A STATISTICAL APPROACH TO JOB MATCHING PROBLEM VIA DIFFERENCE METRICS AND DATA MINING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AHMET FATİH ORTAKAYA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
STATISTICS

DECEMBER 2016

Approval of the thesis:

A STATISTICAL APPROACH TO JOB MATCHING PROBLEM VIA DIFFERENCE METRICS AND DATA MINING

submitted by **AHMET FATİH ORTAKAYA** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Statistics Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** ___________________

Prof. Dr. Ayşen Dener Akkaya
Head of Department, **Statistics** ___________________

Assoc. Prof. Dr. Özlem İlk Dağ
Supervisor, **Statistics Dept., METU** ___________________

Assoc. Prof. Dr. Cem İyigün
Co-Supervisor, **Industrial Engineering Dept., METU** ___________________

Examining Committee Members:

Prof. Dr. Jülide Yıldırım Öcal
Economics Dept., TEDU ___________________

Assoc. Prof. Dr. Özlem İlk Dağ
Statistics Dept., METU ___________________

Assoc. Prof. Dr. Berna Burçak Başbuğ Erkan
Statistics Dept., METU ___________________

Assoc. Prof. Dr. Ceylan Talu Yozgatlıgil
Statistics Dept., METU ___________________

Assoc. Prof. Dr. Osman Abul
Computer Engineering Dept., TOBB ETÜ ___________________

Date: 23.12.2016

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Ahmet Fatih ORTAKAYA

Signature:

ABSTRACT

A STATISTICAL APPROACH TO JOB MATCHING PROBLEM VIA DIFFERENCE METRICS AND DATA MINING

Ortakaya, Ahmet Fatih
Ph.D., Department of Statistics
Supervisor : Assoc. Prof. Dr. Özlem İlk Dağ
Co-Supervisor: Assoc. Prof. Dr. Cem İyigün

December 2016, 139 pages

Labor market vision has changed from a stock perspective to a flow perspective in the recent years. Majority of labor markets in many high income countries are specified by these gross flows. Due to these large flows, a major issue arises in matching workers and jobs in labour market which results in coexistence of high number of unfilled vacancies and unemployed people.

Different approaches are applicable in the literature to match the right candidate with the right job post. Yet, as far as we know a sophisticated statistical analysis or a procedure for employing a Job Matching Scheme (JMS) for Turkish Employment Agency (TEA) does not exist. The main aim of this thesis study is to develop a statistical approach for designing a JMS by proposing a new Classification Algorithm for Categorical Data with Incremental Feature Selection

(CACDIFES) and a Matching Algorithm which consists of a combination of scoring and sorting algorithms by using Independently Weighted Overlap Metric (IWOM). Apart from the studies in the literature, this thesis proposes a new Incremental Feature Selection (IFS) algorithm, an Independently Weighted Value Difference Metric (IWVDM) and a modified version of Overlap Metric (OM) which can be applied to any type of categorical data sets.

Algorithms proposed in this thesis are applied to TEA data set and data sets obtained from UCI Machine Learning Repository. Experimental results reveal that our proposed metric is superior to previously introduced ones, and our JMS is able to match all vacant jobs with suitable job seekers.

Keywords: Job Matching Scheme (JMS), Classification Algorithm for Categorical Data with Incremental Feature Selection (CACDIFES), Independently Weighted Value Difference Metric (IWVDM), Independently Weighted Overlap Metric (IWOM), Matching

ÖZ

FARK METRİĞİ VE VERİ MADENCİLİĞİ KULLANILARAK İŞ EŞLEŞTİRME PROBLEMİNE İSTATİSTİKSEL BİR YAKLAŞIM

Ortakaya, Ahmet Fatih
Doktora, İstatistik Bölümü
Tez Yöneticisi : Doç. Dr. Özlem İlk Dağ
Ortak Tez Yöneticisi : Doç. Dr. Cem İyigün

Aralık 2016, 139 sayfa

Son yıllarda işgücü piyasası vizyonu stok perspektifinden akış perspektifine yönelmiştir. Yüksek gelir grubundaki birçok ülkedeki işgücü piyasası brüt iş akışı çokluğu ile ifade edilmektedir. Büyük iş akışlarından ötürü, iş gücü piyasasındaki işçiler ile açık bulunan iş pozisyonlarını eşleştirme sorunu ortaya çıkmakta ve bu durum, yüksek sayıda açık iş pozisyonu ve işe yerleşemeyen kişilerin var olmasına neden olmaktadır.

Doğru işi doğru kişi ile eşleştirmek için literatürde farklı uygulamalar mevcuttur. Bilindiği kadarıyla Türkiye İş Kurumu'nun istatistiksel analiz yöntemleri aracılığı ile geliştirmiş olduğu bir iş eşleştirme programı mevcut değildir. Bu tez çalışmasının temel amacı; Artırımlı Değişken Seçimini İçeren Yeni Bir Kategorik

Sınıflama Algoritması (ADSKSA) ve Bağımsız Ağırlıklandırılmış Örtüşme Metriğinin (BAÖM) kullanıldığı puanlama ve sıralama algoritmalarının kombinasyonundan oluşan bir eşleştirme algoritması önererek istatistiksel bir yaklaşım ile iş eşleştirme programı geliştirilmesidir. Literatürdeki çalışmalardan farklı olarak bu tez çalışması tüm kategorik veri setlerine uygulanabilir yeni bir değişken seçimi algoritması, Bağımsız Ağırlıklandırılmış Değer Fark Metriği (BADFM) ve Örtüşme Metriğinin (ÖM) modifiye edilmiş bir versiyonunu önermektedir.

Bu tez çalışmasında önerilen algoritmalar Türkiye İş Kurumu veri seti ve Kaliforniya Üniversitesi-Irvine Makine Öğrenmesi Deposu veri setlerine uygulanmıştır. Uygulama sonuçları; bu tezde önerilen metriğin literatürdeki metriklere göre daha üstün olduğunu göstermekte ve geliştirilen iş eşleştirme programı, bütün müsait işlerle uygun iş arayanları eşleştirebilmektedir.

Anahtar Kelimeler: İş Eşleştirme Programı (İEP), Artırımlı Değişken Seçimini İçeren Kategorik Sınıflama Algoritması (ADSKSA), Bağımsız Ağırlıklandırılmış Değer Fark Metriği (BADFM), Bağımsız Ağırlıklandırılmış Örtüşme Metriği (BAÖM), Eşleştirme

To My Wife

ACKNOWLEDGEMENTS

First and foremost, I would like to express profound gratitude to my thesis supervisor Assoc. Prof. Dr. Özlem İlk Dağ for her invaluable knowledge, academic experience, support, encouragement, supervision and useful suggestions throughout this research work. Second, I would like to express my sincere gratitude to my co-advisor Assoc. Prof. Dr. Cem İyigün for his continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. Without their moral support, continuous guidance, suggestions and encouragements it would be impossible to complete this study successfully and as early as possible. Their guidance throughout the study, support, and feedbacks has turned this study to an invaluable learning experience for me.

I would also like to acknowledge my appreciation to my examining committee, Prof. Dr. Jülide Yıldırım Öcal, Prof. Dr. İnci Batmaz, Assoc. Prof. Dr. Berna Burçak Başbuğ Erkan, Assoc. Prof. Dr. Ceylan Talu Yozgatlıgil, Assoc. Prof. Dr. Osman Abul and Asst. Prof. Dr. Fehmi Tanrısever for spending their precious time in reviewing my thesis. Their valuable comments and suggestions have improved the quality of my thesis.

My sincere thanks go to employees of TEA who have spent their precious time, even days, for obtaining data set that is used in my thesis. I also wish to thank Prof. Dr. John W. Emerson from Yale University for his valuable comments on revising my R-code.

Last but not least, I would like to thank my wife for supporting me in any means throughout writing this thesis and my life in general.

TABLE OF CONTENTS

CHAPTERS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AHP	:	Analytic Hierarchy Process
ALMP	:	Active Labor Market Policy
ANP	:	Analytic Network Process
AVDM	:	Augmented Value Difference Metric
AWVDM	:	Attribute Weighted Value Difference Metric
CACDIFES	:	Classification Algorithm for Categorical Data with Incremental Feature Selection
CJO	:	Consultant of Job and Occupation
DC	:	Data Compression
EC	:	European Commission
ETL	:	Extraction, Transformation and Loading
FDM	:	Frequency Difference Metric
FEO	:	Fair Employment Opportunities
FS	:	Feature Selection
HRLTU	:	High Risk of Long Term Unemployment
IFS	:	Incremental Feature Selection
IWOM	:	Independently Weighted Overlap Metric
IWVDM	:	Independently Weighted Value Difference Metric
JMS	:	Job Matching Scheme
JSS	:	Job Seeker’s Specifications
KL	:	Kullback-Leibler
K-L	:	Karhunen-Loève
KMCMM	:	*k*-Means Cluster-Based Mean-and-Mode
k-NN	:	*k*-Nearest Neighbor
LFP	:	Labor Force Participation

MBR	Memory-Based Reasoning
MCDM	Multi-Criteria Decision-Making
MDS	Multidimensional Scaling
NACE	Statistical Classification of Economic Activities in the European Community
NCBMM	Natural Cluster Based Mean-and-Mode
NLM	Nonlinear Mapping
ODVDM	One Dependence Value Difference Metric
OECD	Organization for Economic Co-operation and Development
OM	Overlap Metric
PCA	Principle Component Analysis
RCBMM	Rank Cluster Based Mean-and-Mode
RO	Representative Object
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TEA	Turkish Employment Agency
TURKSTAT	Turkish Statistical Institute
UA	Unemployment Assistance
UCI	University of California Irvine
UI	Unemployment Insurance
UK	United Kingdom
VDM	Value Difference Metric
VGEO	Very Good Employment Opportunities
VR	Vacancy Requirements
WEO	Weak Employment Opportunities

CHAPTER 1

INTRODUCTION

Developing a dynamic job matching model by taking the heterogeneity between vacancies and job seekers into account requires an interdisciplinary perspective. This study stands in the cross-section of labor economy, data mining and statistics, and such a model can be generated by the combination of these three disciplines. According to an extensive literature review, early job matching models in labor economy consist of aggregated job matching functions or models obtained by empirical studies which are based on econometric analysis. Studies concerning this issue in Turkey mainly focus on determinants of employment or unemployment without taking industrial economic activity of the jobs in labor market into account. In addition, preliminary meetings with Turkish Employment Agency (TEA) reveal that Agency does not implement a job matching model or a procedure considering the heterogeneity of the job seekers and vacancies. Thus, a nationwide statistical job matching model for Turkey does not exist and it is developed through this thesis study.

In this thesis, we will focus on generating a dynamic Job Matching Scheme (JMS) by proposing a new Classification Algorithm for Categorical Data with Incremental Feature Selection (CACDIFES) and a matching algorithm which consists of scoring and sorting. As job matching problem is a comprehensive issue, we partition our problem into two main phases. In the first phase we will classify job seekers depending on their employment opportunities, and in the second phase we will employ scoring and sorting to match job seekers who are classified as having Very Good Employment Opportunities (VGEO) for the available job posts. Steps for the generation of these phases are defined in the following chapters.

This thesis consists of six main chapters. These chapters are organized in this way: In the first chapter, problem definition, objective of this thesis, motivation and contributions to literature will be given. In the second chapter, a literature review which covers job matching functions, studies concerning the determinants of employment and unemployment in Turkey and job matching studies of TEA will be defined. Moreover, in this chapter, overview of classification methods and major tools used in classification will be given. In the third chapter, our proposed method for classification, CACDIFES, will be explained. In addition, this chapter includes experimental results of CACDIFES Algorithm using UCI Machine Learning Repository data sets. In the fourth chapter, our proposed matching algorithm will be described. In the fifth chapter, implementation of JMS will be given. Furthermore, in this chapter a detailed data cleaning process for TEA data set is presented. In the final chapter, discussion and conclusion of this thesis will be given.

1.1. Background

Before introducing the background of our research question, definitions of unemployment, unemployment ratio and unemployed are given. Unemployment can be defined as the state of being unemployed. Unemployed people includes people above a certain age during the reference period who are without work, (i.e. neither in paid employment nor in self-employment), available for work at present (i.e. available for paid employment or self-employment) and seeking work (i.e. had taken certain measures to find paid employment or self-employment such as registration to public or private employment agencies, checking employment worksites, etc.) (ILO, 1982).

Unemployment rate is defined as the ratio of unemployed people over the labour force where the labour force includes the unemployed people plus people in paid-employment or self-employment. According to this definition, unemployment rate in 2014 was reported to be 7.3% (on the average) for OECD countries and 9.9% for Turkey. Although, this rate decreased to 6.8% for OECD countries in 2015, it increased to 10.3% for Turkey (OECD, 2016a).

Employment / unemployment issue is one of the top priority agenda for almost all countries in the world. Most of the countries aim for decreasing unemployment rate and increasing the overall employment rate by applying macroeconomic or microeconomic policies (TEA, 2014a). Due to global economic crises during 2008, unemployment rate in the world has increased rapidly and this made countries to take serious precautions to fight against unemployment. In our country the main responsible nationwide institution for mediation of the services of job and employees and development and protection of employment is TEA.

It is known that one of the most powerful tools for fighting unemployment is the creation of the workforce that the labor market requires. Furthermore, directing this created amount of workforce to most suitable vacancies in a short time is as important as the creation of it to keep the long term unemployment rate down and to sustain economic development.

According to information concerning the job and employee services provided by the TEA, total number of people registered as unemployed in Agency’s database in 2014 is 2,747,978. Among these unemployed, the number of people who applied for a job in 2014 is 2,375,583. Within this time period total number of vacancies in the system is 1.735,892. Out of these vacancies only 701,435 of them are filled by the applicants in 2014 (40.4% of the total vacancies) by the studies of TEA. Besides, total number of orientations (directions) through employee seekers for registered vacancies in 2014 is 8,782,612 (The term orientation is used for direction of job applicants to certain open vacancies by TEA who are deemed to be suitable candidates for those jobs). Out of total number of directions, 701,435 of the applicants (7.9%) are able to fill vacancies (TEA, 2014b). Thus, we still have 1,034,457 unfilled vacancies and 2,046,543 unemployed people left in the system for 2014 which clearly shows the urgent need for a sophisticated JMS for TEA.

1.2. Objective of This Thesis

According to the preliminary meetings with TEA, although there are different techniques for job matching programs in the literature, TEA does not employ a nationwide statistical job matching model or a procedure which uses the job seekers' specifications, labor market conditions and heterogeneity among vacancies. The main objective of this thesis study is to develop a statistical approach for designing a JMS by proposing a new classification and matching algorithms. Since we have a complicated problem to solve, we break down our problem of interest into two main phases: Classification and Matching.

In the first phase, job seekers will be classified depending on their employment opportunities, and in the second phase those who have a high rate of transition through labour market (who are classified as job seekers with Very Good Employment Opportunities) will be matched by the vacancies using scoring and sorting algorithms (Figure 1).

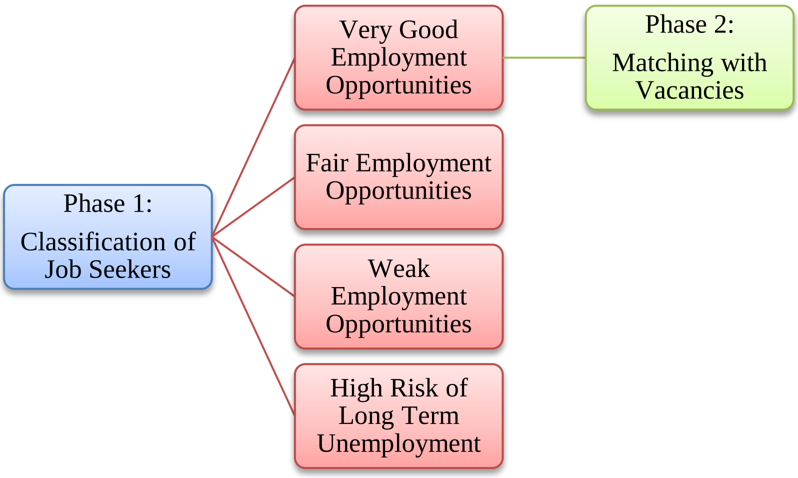
Phase 1:
Classification of Job Seekers
Very Good Employment Opportunities
Fair Employment Opportunities
Weak Employment Opportunities
High Risk of Long Term Unemployment
Phase 2:
Matching with Vacancies

Figure 1. Phases of Job Matching Scheme

1.3. Motivation

In the previous sections, urgent need for developing a job matching model for TEA is explained in details. Specifically, having a high rate of unemployment is one of the major structural issues for Turkish economy. In order to lower the unemployment rate, quality and quantity of jobs in labour market should be increased and they should be accessible by suitable job seekers. However, due to large number of job creation and destruction and a high rate of heterogeneity in labour market conditions, job seekers have difficulties in matching with right posts which undoubtedly results in high number of unfilled vacancies and unemployed workers. Main motivation in this thesis is to address this issue and seek a solution by developing a JMS for TEA by using a combination of classification and matching algorithms.

After a comprehensive literature review on job matching functions, we discovered that majority of job matching functions stem from studies in labor economy (aggregate job matching functions) and others obtained by empirical studies which mainly focus on the determinants of employment/unemployment. As far as we know none of these studies concentrates on matching high number of job seekers with high number of vacancies with respect to Job Seeker’s Specifications (JSS) and Vacancy Requirements (VR). Hence, our second motivation in this thesis is to propose a statistical approach to one major structural issue in labour economy by using a large-scaled micro level data set with the help of data mining tools.

Difference metrics used in traditional categorical classification approaches do not take the dependence structure among attributes into account and they make strong assumptions on attribute independence which are not realistic in many real-world data sets. Another major issue is that these metrics do not consider the attribute importance (i.e. they do not differentiate attributes depending on their relevance to class variable). Thus, our third motivation is to generate new difference metrics which take the dependence structure among attributes and their relevance to class variable into account.

Finally, many-to-many matching of two different data sets (for instance, job seekers and employee seekers) is one of the major issues in any decision making problem since each set may contain different attributes in different scales. Our fourth motivation is to generate a scoring algorithm which calculates the distance of all observations in data set one with all observations in data set two (many with many) and propose a sorting algorithm which matches the observations in both sets depending on their magnitudes of distances.

1.4. Contributions to Literature

Major literature contributions of this thesis are twofold due to its interdisciplinary behavior. These contributions can be grouped under the domains of data mining and statistics and labour economy.

From data mining and statistical perspective, major literature contribution of this thesis is proposing a new categorical classification algorithm including a feature selection and a new difference metric and presenting a matching algorithm by making use of scoring and sorting which possess a modified version of Overlap Metric (OM).

From labor economy perspective, major literature contribution of this thesis is generating a national-scale JMS for TEA by using a large-scaled micro-level administrative data. While doing so this thesis fills a significant gap in fighting against unemployment by serving as a decision support system for TEA. This thesis does not only help improving institutional capacity of TEA but also it acts as an instrument for increasing the efficiency of services concerning job and employees provided by TEA.

Implementation of JMS is done by using a real data set obtained from TEA. This data set satisfies the two major features of big data which are high number of attributes (90 attributes with 54 of them being categorical) and high number of instances (it has more than 2.3 million records). Since our data is not clean and well-structured a comprehensive data cleaning process is employed (see sections

2.4.1 and 5.2). Hence, this thesis provides a good example of working with large-scale and noisy micro-level administrative data.

Majority of the classification algorithms in literature are not data oriented. In this thesis study, we develop a data oriented difference metric including an embedded feature selection and data compression algorithms. Next, by using our generated difference metric, we propose a classification algorithm for categorical data.

Matching algorithm generated in this thesis proposes an alternative approach to Multi-Criteria Decision-Making (MCDM) problems. Previously introduced MCDM methods are highly subjective and they are mainly appropriate for solving univariate MCDM problems. However, our matching algorithm provides a solution to multivariate MCDM problems, it has a well theoretic base (not subjective) and it takes the dependence structure among attributes into account.

Contributions of classification and matching algorithms are described in details in the following sections.

1.4.1. Contributions of CACDIFES algorithm

Majority of the clustering and classification studies in literature mainly focus on numerical attributes which have a natural ordering. However, there is also a growing attention on categorical domains in which attributes do not have a natural ordering. Generally, clustering and classification algorithms discovered so far requires different types of input parameters which are hard to guess in advance, such as stopping criteria in different optimization process or radius of clusters, etc. Most of these input parameters are not known before the analysis or can only be guessed with domain knowledge (Carbonera and Abel, 2014a; 2014b; Ganti et al., 1999).

Majority of the clustering algorithms (conventional clustering algorithms) discovered so far are capable of clustering small or medium sized data sets having smaller size of dimension. They directly focus on clustering data points in domain taking the value difference among data objects into account. Hence, these conventional clustering algorithms do not scale well with high dimensional large-scaled data sets due to memory space issues (Memory-Based Reasoning – MBR) (Kasif et al., 1998). Besides, depending on the type of the data set being clustered, these algorithms try to evaluate the distances among data objects and cluster centers (mean for numerical data and mode for categorical data) during the iteration process which requires data set to be stored in main memory and to be scanned more than once. In order to cope with increasing number of dimensions and data points different techniques are proposed such as feature selection, dimensionality reduction or attribute weighting (Guyon and Elisseeff, 2003).

In some algorithms, representative data points (or vectors) are defined as cluster centers and points are assigned to those clusters depending on distance measures being used (Zhang, 2006). During the clustering process cluster centers might be adjusted (if more appropriate points are found, i.e. if within cluster similarity increases). Also, different methods are proposed for finding seeds to discover cluster centers (or representative points) such as employing a random sampling procedure by using Chernoff bounds (Motwani and Raghavan, 1995; Guha et al., 1998). Although these methods help reducing the number of data points, minor changes with the seeds result in totally different clustering outcomes which clearly indicate a high rate of dependence to the order of the input data.

Majority of the algorithms discovered so far make use of the idea of attribute independence. However, most of the real life data sets include possible correlation structures among attributes as defined in the previous section and attribute independence assumption is not valid for these data sets. Different studies in the literature investigate this issue by using attribute weighting. Some of the existent studies introduce data driven approaches and some require domain knowledge for assigning weights for attributes. Yet, none of these methods are superior to other (Li and Li, 2011; Li et al., 2013, Jiang and Li, 2013; Jiang et al., 2014).

In this study, we focus on generating a new Classification Algorithm for Categorical Data with Incremental Feature Selection. For instance, our proposed algorithm will be used for grouping job seekers with respect to their degree of employability by taking into account their personal specifications, qualifications and skills. Contributions of this algorithm are given as follows:

- This algorithm can be applied to high dimensional categorical data sets,
- It only requires a single input parameter and it can be guessed by the use of a graphical representation,
- It uses an IFS method to take the correlation structure between attributes into account so that it can be applied to any type of real life data sets,
- Apart from many algorithms in literature, it has a well-designed theoretical basis (i.e., it is generated by using information theory),
- It presents a pre-processing step (Data Compression, DC, step) for identifying dense points (Representative Objects) in data set by using OM which can be applied in any type of categorical clustering/classification algorithm,
- It introduces a new difference metric named as Independently Weighted Value Difference Metric (IWVDM) based on symmetric uncertainty,
- It can be used as a tool in missing value imputation methods,
- It does not require any distributional assumption for data set,
- Input order of data points is not important and classification outcome is unique.

In fact, this algorithm is easy to apply and quite useful for real world data sets. By further modification, this algorithm can be applied to any type of data sets (numerical or mixture type of high dimensional data sets).

1.4.2. Contributions of matching algorithm

Matching two different data sets is one of the most difficult tasks in decision making problems. Some of the previously introduced methods are mainly concerned with MCDM problems. These problems include choosing finite number of alternatives from a set of observations by using scoring and ranking with respect to different criteria (attributes). Commonly used methods are Analytic Hierarchy Process (AHP), Analytic Network Process (ANP), utility models like ELECTRE, PROMETHEE and TOPSIS (Saaty, 1990; Saaty 2008; Bufardi et al., 2004; Mergais et al., 2007; Bogdanovic, 2012).

Although these MCDM methods are quite useful and widely used, the hierarchies of criteria (i.e., the importance of attributes) are mainly based on previous work, user experience or empirical studies. First, only a small percentage of studies are based on empirical studies. Others employ a pair-wise comparison procedure for criteria (for attributes) in a scale from 1 to 9 (1 equally preferred and 9 extremely preferred) depending on decision goal by user. Major problem with this procedure is that it is highly subjective (i.e. result changes from one user to another with respect to decision (selection) purpose).

Second, MCDM methods can be used for a single decision making problem. For example, if we would like to fill in a single vacancy by an employee who has a specific degree then we should look for job seekers who only hold that required degree. Unfortunately, our problem of interest is to fill in different types of vacancies which have different types of requirements by job seekers who have different types of skills and job relevant specifications (many with many). We might consider our problem as a multivariate case of MCDM.

In this thesis, we propose a matching algorithm based on scoring and sorting which includes IFS and DC by taking into account the dependence structure among attributes. Apart from the studies in literature which are mainly based on subjective decision making procedure, our algorithm is based on information theory, and it does not require user-domain knowledge. Our proposed algorithm provides a different approach to multivariate MCDM problems. By further modification our matching algorithm can be employed to any type of matching problems which include numerical, categorical or mixture type of high dimensional data sets.

Matching algorithm proposed in this thesis includes a modified version of OM. OM does not make use of attribute importance and it does not take into account dependence structure among attributes (see section 3.1.1 for a detailed discussion). Thus, we propose a new difference metric named as Independently Weighted Overlap Metric (IWOM) by using IFS and DC phases of CACDIFES algorithm for achieving attribute independence and importance. Overall contributions of matching algorithm are as follows:

- This algorithm can be applied to multivariate MCDM problems,
- It takes into account the correlation structure between attributes so that it can be applied to many real life data sets,
- Compared to many MCDM methods, which are based on subjective decision making procedures, this algorithm is designed by using information theory,
- It uses an IFS method controlling the dependence structures of attributes,
- It introduces a new difference metric named as Independently Weighted Overlap Metric (IWOM) based on symmetric uncertainty,
- It does not require any distributional assumption for data set,
- By making suitable modification, it can be applied to any type of matching problems which include numerical or mixture type of high dimensional data sets.

CHAPTER 2

LITERATURE REVIEW

In this part of the thesis, studies concerning job matching models and methods used in classification will be reviewed. The early job matching schemes found in the literature enlightens the importance of the subject from different perspectives. Majority of these studies consist of job matching functions mainly used in the labor economy. In sections 2.1 and 2.2, job matching functions which stem from the studies in labor economy, and the job matching functions that are obtained by focusing on empirical studies will be introduced. In section 2.3, information about TEA and job matching studies of the Agency will be given.

In section 2.4, overview of classification algorithms, data cleaning process, scale conversion and data transformation techniques, dealing with missing value issues will be given. In section 2.5, brief information on feature selection methods will be presented. In section 2.6, difference metrics commonly used in categorical classification algorithms will be described. These three sections will constitute a basis for CACDIFES and Matching Algorithms which are proposed in this thesis. Finally in section 2.7, certain performance measures for classification algorithms will be given.

2.1. Concept of Job Matching Functions

Majority of the labour markets in many high income countries have changed substantially in the recent years. These significant changes are highly affected by the transformation in labor market vision. It is known that the labor market vision has changed from a stock perspective to a flow perspective. On account of this structural change and large flows, labor markets trouble matching workers and jobs suitably, so vacancies and unemployed workers still exist (Blanchard and Diamond, 1989; Burda and Wyploz, 1994; Steven Davis et al., 1996; Smith and Zenou, 2003).

Job matching functions are used to denote the mathematical relationship of flow of job matches on account of stock of job searchers and stock of available jobs. The continuous time job matching function per unit is given by;

$$mL = m(uL, vL), \tag{1}$$

where, L is the number of identical workers in the labor force, u is the unemployment rate (i.e. the fraction of unmatched workers) and v is defined as the number of identical vacancies as a fraction of labor force (i.e. vacancy rate). Major assumption in job matching function given in (1) is that only uL unemployed workers and vL vacancies engage in matching. Other assumptions are; this function is assumed to be increasing with respect to uL and vL, it is homogenous of degree one and it is concave (Pissarides, 2000).

Beginning from the 1970s, the job search theory became a corner stone used in the labor market analysis. Mortensen (1986) stated that experts view on the employment and the unemployment has changed since 1970s and the characteristic of the employees, their work experiences and the needs of the labor market have serious influences on the stock flow of jobs. In addition, he mentioned that the duration of Labor Force Participation (LFP) and the duration of unemployment of different demographical groups differ when unemployment and LFP data is considered jointly.

Mortensen (1986), Mortensen and Pissarides (1999) were the first to apply the job search and matching models on the labor market studies. These models later are used for evaluating the probability of employment in the labor market, determining the duration of unemployment and investigating the efficiency of the employment in policy studies. Kiefer and Neumann (1979) was the first to employ an empirical study of standard matching function by using reduced models under the assumption of constant reservation wage. In their study they obtained two different models; one with constant and the other with variable reservation-wage. Soon after, Flinn and Heckman (1982) estimated the structural job search models. Empirical studies have shown that labor force in less efficient firms / sectors is destined to become even least efficient firms / sectors (Foster, Haltiwanger, and Krizan, 2001; Lentz and Mortensen, 2008). On the other hand, the earnings of those who change their jobs are reported to become higher than they used to be (Bartel and Borjas, 1981; Mincer, 1986; 1994).

Flinn (2006) evaluated the impact of minimum wage on the welfare and employment by using a discrete time Nash equilibrium model. In another study, Flinn (2002) evaluated the impact of minimum wage on the distribution of wage by using a longitudinal data analysis in Italy and the USA. The outputs of the estimated structural search models are found to be compatible with the wage differences. Employees with similar qualifications in different firms are found to be paid different wages. Hence, employees respond to that situation by moving from less-paid jobs to high-paid jobs. Jolivet et al. (2006) found evidence for OECD countries and Christensen et al. (2005) found evidence for Denmark.

Frederiksen (2008) investigated the impact of gender on the permanent employment and the probability of leaving job. According to his findings, women has a higher probability of leaving job, thus, employment permanence for a given job is less stable for them compared to men.

Lindeboom et al. (1994) estimated a general model which makes job seekers and employee seekers to interact by using different channels. Four sources of interaction channels (job advertisements, informal ways, employment offices and other sources of interaction) are proposed in their study in which the vacancy announcing and job seeking take place. Four different matching functions are estimated for each channel.

Alba et al. (2012) analyzed the transitions out of unemployment for benefit recipients in Spain by using the data from Integrated Benefits System (Spanish administrative data which provides information about the unemployment benefits received by each worker). Data include information on unemployment insurance (UI) and unemployment assistance (UA) being received, benefits duration, demographical information about the receiver, etc. In their analysis, they study unemployment exit rates around benefit exhaustion.

Andrews et al. (2001) estimated the probability of a match for contacts between job seekers and vacancies by using microeconomic data for a particular labour market in UK, and they observed that the determinants of matching is related to the characteristics of job seeker, labour market conditions and vacancies. They employed a probit regression model to estimate the probability of a match.

Other empirical studies investigating the probability of employment and the duration of unemployment indicates that the decision of labor force participation is affected by the factors such as gender, education level, decision on fertility, inequality of wages, marital status, labor market conditions, ease of finding temporary jobs and unemployment insurance (Caucutt et al., 2002; Gray and Hunter, 2002; Veracierto, 2008; Kyyrä and Ollikainen, 2008; Tatsiramos, 2009; Boone et al., 2009; Arcidiacono et al., 2010; Kahn, 2012; Erceg and Levin, 2014).

2.2. Studies for the Determinants of Employment and Unemployment in Turkey

Majority of the empirical studies in Turkey concerning employment-unemployment issue focus on determinants of the unemployment. Taşçı and Tansel (2006) investigated the determinants for the duration of youth unemployment in Turkey using the data from 2000-2001 household labor force survey by Turkish Statistical Institute (TURKSTAT). According to their analysis, the probability of the re-employment of young women is smaller than those of men. Besides, they point out the importance of geographical differences for the determination of unemployment duration. There is not a significant difference in the probability of employment among high school and vocational high school graduates. Having a university degree for the young men increases the probability of finding a job but it is not the case for the young women.

Taşçı (2008) explored the determinants of the density of job search of unemployment by using the 2001 household labor force survey data by TURKSTAT. In his study, he identifies how the density of job-search is affected by the gender, location of residence (urban-rural areas), family types and the labor market. Intensity of job search is estimated by using an ordered-probit model taking into account the sample selection problem. According to results of the study, the density of job search of unemployed staying in urban areas is higher than those who stay in rural areas. Higher levels of education increase the density of job search. The density of job search for women is smaller than men. There exists a reverse-U type of relation between age and density search of job.

Taşçı and Darıcı (2009) studied the determinants of the unemployment in Turkey taking into account the gender difference by using different kinds of unemployment definitions. Household labor force survey data held in 2006 by TURKSTAT was used in their analysis. According to the results of their analysis, probability of women's becoming unemployed is smaller than men. Living in urban areas decreases the probability of being unemployed. Compared to those who do not have any levels of education, as the education level of men increases the probability of being unemployed decreases. Converse is true for women. Both for men and women they discovered a reverse-U type of relation between age and the probability of being unemployed.

One of the most powerful tools for fighting against unemployment is the creation of the workforce that the labor market needs. In addition, directing the created amount of workforce to most suitable vacancies (matching) in a short period is as important as the creation of it so as to keep the long term unemployment rate down and to sustain economic development (World Bank, 2014). In Turkey the main responsible nationwide institution for mediation of the services for vacancies and employees is TEA. Agency is defined as a complementary governmental body for developing national employment policies, protecting employment and preventing the unemployment. One of the fundamental activities of TEA is to match job seekers with employee seekers.

For the implication of TEA's fundamental goals, Kumaş (2010) emphasized the importance of developing a user-friendly web-based system in which job seekers and employee seekers can access and directly do their matches. Öz (2010) stated that the periods of deterioration of the negative relation between the number of unemployed and the vacancies shows the weak performance of match in labor markets. Mismatch between the requirements of open job posts and the qualifications of the job seekers can be regarded as a statement of matching shocks which results in a long term unemployment and equilibrium in higher rate of unemployment even the aggregate demand tend to recover after the economic crises. He underlined the importance of this issue by requiring further studies.

Koçak and Akman (2013) emphasized the importance of the services given by Consultants of Job and Occupation (CJOs) of TEA and implied the necessity of increasing the awareness of the job seekers about these services so as to overcome the problem of unemployment in their survey study in Yalova. They found out that due to not knowing the way of seeking job and limited or no information about the employment market, the unemployment rate and its duration increase.

According to the results of their study, majority of the job seekers asked for an assistance during their job search process. Owing to these findings they conclude that all job seekers should be given professional assistance, and they should be monitored during the overall process of the job seeking and employment by the expert personnel of TEA.

2.3. Job Matching Studies of TEA

TEA is the main nationwide responsible institution for providing services of job and employees in Turkey. It does not only implement policies for protecting, improving and maintaining employment services that labour market requires but also it uses Active Labor Market Policy (ALMP) instruments such as temporal income support for boosting the employability of the workforce (TEA, 2013b).

The regulation of public employment services dates back to 1936 in Turkey. There are several improvements both in the services of job and employment and in the regulations concerning these services due to transitional behavior of labor market conditions. After all, TEA obtained its establishment law through Turkish Employment Agency Law with the number 4904 in 2003. The new law aims for increasing the institutional capacity of TEA by enhancing the efficiency of its activities. Moreover, this law is designed to improve the democratic governance of TEA by including social parties and representatives of workers and employers.

Other important advancements by the new law are the establishment of private employment agencies, constitution of regional and provincial directorates, employment of experts in TEA and finally employment of CJOs in provincial directorates and in employment offices. These crucial steps have significant effect on the development of institutional capacity of TEA. Furthermore, TEA carries on the market analysis in different job sectors on account of the needs of labor market so as to increase the performance of its services. These studies constituted by regular visits of workplaces, demand analysis of the workforce, identification of the open job posts and assistance to job seekers in job search process. By the help of these studies both in local and central level, job seekers are aimed to be directed to suitable jobs (i.e. job matching). The overall study of job matching in TEA has five fundamental steps (TEA, 2013b):

1. **Application to TEA's units**: Job seekers apply for a job through electronic channels (web-based application) or they can apply through TEA's provincial directorates or job and employment offices in person. All this information is registered to TEA's database.
2. **Investigation of application:** Information concerning the application is investigated and a pre-evaluation process is carried on whether the eligibility criteria for the job being applied is satisfied or not.
3. **Invitation to job interview:** Applicants who are found eligible for a given job are invited to a job interview and further documents related to this job are acquired (if necessary).
4. **Job evaluation:** Performance of applicants in job interview, information being collected, number of open job posts registered in the system, and personal assessments of CJOs for the applicants are evaluated to check whether the eligibility criteria for the post are satisfied or not.
5. **Direction to vacancy or training:** According to overall evaluation period those who satisfy the eligibility criteria for the given job posts are directed to these jobs and those who might face the long term unemployment risk can be directed to different types of vocational trainings (Figure 2).

These steps are implemented by the personnel who work on behalf of the TEA. The final decision on directing applicant to a given vacancy or a vocational training program is given by the expert view of these personnel. This evaluation process might be a difficult task considering so many indicators or employment specific information about the applicants with respect to different job sectors in different provinces of Turkey. Hence, for designing an optimal procedure based on objective criteria, a well-designed JMS is to be created.

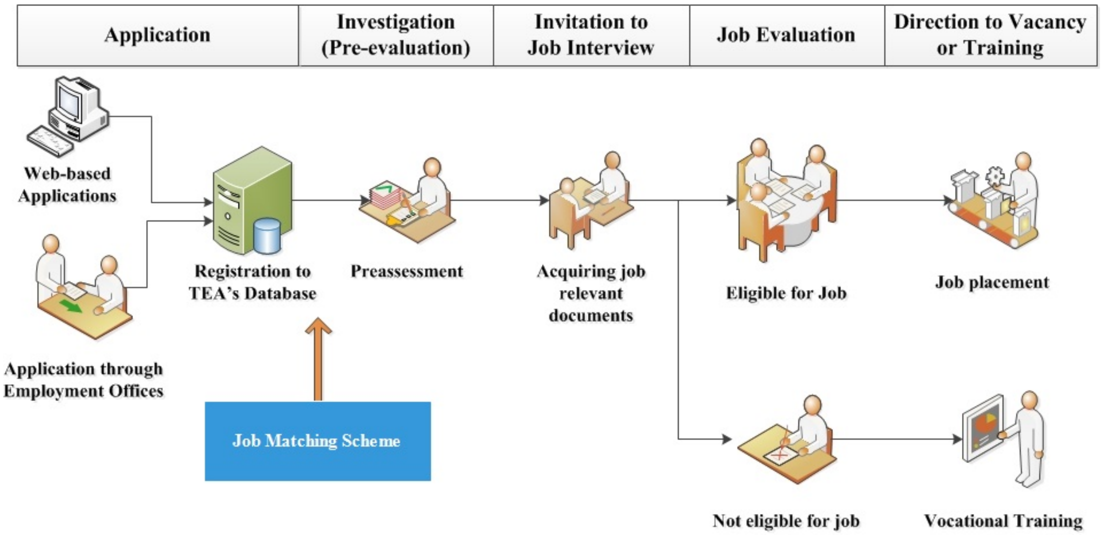
Application
Investigation (Pre-evaluation)
Invitation to Job Interview
Job Evaluation
Direction to Vacancy or Training
Web-based Applications
Application through Employment Offices
Registration to TEA's Database
Preassessment
Acquiring job relevant documents
Eligible for Job
Job placement
Job Matching Scheme
Not eligible for job
Vocational Training

Figure 2. Job Matching Studies of TEA

JMS generated in this thesis study is aimed to be put in use after the application process (Figure 2). Thus it is aimed to act as a decision support system for the employees of TEA in the overall process.

2.4. Main Concepts in Classification Methods

Briefly, classification is one of the major data mining methods which are used to predict group membership for data points (Phyu, 2009). Classification algorithms are used to categorize objects into one of several predefined sets. Generally, classification and clustering methods are confused. Main difference between these two algorithms is that in classification, group of objects are put into predefined sets whereas in clustering, these sets are unknown before the analysis. Clustering algorithms (also named as taxonomy analysis, segmentation analysis or unsupervised classification) are used to group objects into sets in which the objects are similar within the sets and objects in different sets are dissimilar (Gan et al., 2007; Aggarwal, 2014).

Before exploring classification algorithms, terminology frequently encountered will be introduced. The terms such as *data point, observation, individual, object, item* or *tuple* are used to define a single record. A data point in a high-dimensional space will be defined as *variable*. An *attribute* or *feature* is used to denote a dimension in a *d*-dimensional space. *Target variable* (or *class variable*) is the prediction class of objects which is the main aim of research in classification algorithms. Major difference between regression and classification is that the target variable in classification is a discrete variable whereas in regression it can be discrete or continuous. *Training set* is referred as set of objects which are used to build predictive model (whose class labels are known in advance) and *testing set* is referred to objects which are used to test (validate) the predicted model (Jain et al., 1999; Gan et al., 2007; Aggarwal, 2014).

Generally, classification algorithms consist of two major phases: training and testing. In training phase, a classifier or a model is constructed using training data. In testing phase, this built model is used to assign labels to unlabeled testing objects. Training phase is the learning step of classification algorithm. In this phase, algorithm learns the interrelations among features and target variable by using training set and in testing phase testing objects are labeled by using these learned structure among features and target variable (Aggarwal, 2014).

Classification algorithms can be used in a variety of fields. Major domains that classification algorithms employed are customer market analysis, diagnosis of diseases, biology (biological data analysis), social network analysis, categorization of documents, multimedia analysis, etc (Aggarwal, 2014).

There are different classification algorithms presented in literature. Most common categorical classification algorithms are; decision tree induction, naive Bayes algorithm, rule-based classification, support vector machines (SVMs), *k*-nearest neighbor (*k*-NN), etc (Saranya et al., 2014).

Decision tree is an analytical model used in data mining and statistics. It uses decision tree as a predictive model for mapping an item's observations to drive conclusions about an item's end value. In other words, they are used to classify observations by sorting them depending on their feature values. As a general representation in decision tree, leaves show the class labels and branches show the associations of features which are related to class labels (Soundarya and Balakrishnan 2014). It is a directed tree which has a node named as root. Root has no incoming edges whereas all other nodes have only one incoming edge. A node that has an outgoing edge is named as internal node while all others are named as leaves or decision nodes. Algorithm works in a recursive manner. First, an attribute is selected as a root node. Second, observation space is divided into two or more sub-spaces by internal node till all instances are classified. These algorithms are useful as a decision support tool however they do not work efficiently with large-scale data sets which have high number of attributes (Sakshi, and Khare, 2015).

Naive Bayes Algorithm uses a simple probabilistic classifier which stems from the Bayes' Theorem. It makes strong assumption on attribute independence. It is robust to noise; it can be trained in an efficient manner in supervised learning problems, and it is easy to construct and interpret. Yet, it requires a high number of observations to obtain good results, it is a lazy learner, and it needs all training set to be stored (Archana and Elangovan, 2014).

Rule-based classification algorithms are generated by using "if-then" conditions. If part (rule antecedent) simply covers the observations which satisfy the "if condition" and else part (consequent) covers the observations which does not satisfy the "if condition". Quality of the classification is measured by accuracy and coverage. Accuracy is defined as the ratio of observations which are correctly classified over observations matched by the rule. Coverage is defined as the ratio of observations matched by the rule over all observations in training set (Saranya et al., 2014; Sujatha et al., 2013). These algorithms are easy to apply and interpret. However, there are certain issues with these algorithms. First, for complicated problems which cannot be structured by straightforward “if conditions”, these algorithms cannot be employed. Second, they do not take into account the attribute importance or relevance with respect to class variable by using statistical tools.

The algorithm of SVM is first introduced by the studies of Vladimir N. Vapnik and Alexey Ya Chervonenks in 1963 and a classifier is generated by Boser et al. (1992). SVMs mainly concern with pattern recognition (linear or non-linear). The main goal in SVM is to construct an optimal hyper plane which linearly separates patterns in data set. In other words, optimal hyper plane is defined to be the one selected from set of hyper planes which maximizes the margin on hyper plane while classifying patterns (i.e. as margin size gets larger, correctness of classification of patterns gets higher). These algorithms are accurate in classification, robust to noise, efficient in text categorization and they are memory-intensive. However, they are binary classifiers (for problems which have more than two classes, these algorithms can be used in a pair-wise manner), and they have a high computational cost (Archana and Elangovan, 2014). Although they are mainly designed for numerical data sets, they can also be used in categorical data sets, too.

One of the major distance-based classification algorithms is k-nearest neighbor (k-NN) and its variants (Cover and Hart, 1967; Aha et al., 1991; Aha, 1992). Basically, it is used to find the class of an observation based on the nearest neighbors of that observation whose class labels are known in advance. k-NN algorithm can successfully estimate the probabilities for class membership but it is not always a good class probability estimator (Li et al., 2013). k-NN is a probability-based classifier which looks for k-nearest neighbor of a test object and then it estimates its class membership probability. For an arbitrary object x, k-NN simply assigns x to the most common class using the k-nearest neighbor of it as shown by;

$$c(x) = \text{argmax}_{c \in C} \sum_{i=1}^{k} \delta\big(c, c(y_i)\big), \tag{2}$$

where, k represents the number of neighbors, y_i's are the k-NNs of x and $\delta\big(c, c(y_i)\big)$ is 1 for $c = c(y_i)$ and 0 elsewhere.

k-NN algorithm is easy to apply and interpret. Also, training part of the algorithm runs very fast. Moreover, it is robust to noisy data and it suits well for multi-class problems (Archana and Elangovan, 2014). Yet, there exist certain drawbacks with k-NN algorithm. First, k-NN employs a simple voting process in which each neighbor is treated in the same manner during the decision process. For example, if there exist correlation structure among attributes than they will probably be seen more frequent with class label which will result in overestimating their effect in classification. Second, standard Euclidean distance and its variants are commonly used as a distance measure which results in a phenomenon named as curse of dimensionality (Beyer et al., 1999; Hinneburg and Keim, 1999) due to possible high numbers of irrelevant attributes. Finally, the size of the neighborhood is given artificially as an input parameter. These drawbacks, lowers the performance of k-NN algorithm. In order to overcome these issues of k-NN we developed CACDIFES Algorithm (see section 3.1).

One of the major issues in classification algorithms is preparing the data for classification. Depending on the classification purpose, this pre-processing stage might include several steps such as data cleaning, scale conversion and data transformation, missing value imputation and feature selection. These steps are briefly described in sub sections 2.4.1 through 2.4.3. Although feature selection is defined under pre-processing stage, we would like to extend the discussion in a separate section since it is a major phase of our proposed classification algorithm.

2.4.1. Data cleaning

Basically, data cleaning (data cleansing or scrubbing) process includes removing noise from data by applying certain type of techniques. These techniques include eliminating errors and inconsistencies from data so as to improve its quality. Generally, data cleaning issues might arise in single database or files due to spelling errors, invalid or missing information during data entrance. However, when data set is to be observed (or merged) from multiple data sources then situation becomes more complicated. In that case, certain features might become redundant, there might be duplicated records, etc (Rahm and Do, 2000).

Typically, building a clean data warehouse includes three main stages: Extraction, Transformation and Loading (ETL). These three stages cover, extraction and transformation of features, matching and integration of extracted features, and finally filtering and aggregating data which will be stored in data warehouse. Bear in mind that each stage of ETL process is done on separate data stages. Although there are different tools for completing these three major stages, commonly majority of tasks are handled manually (Rahm and Do, 2000).

Data cleaning process highly depends on the classification purpose, type of data being used, data collection method, availability of memory and storage, etc. In some cases (such as our case), certain rules (business analytics rules, control and validation mechanisms) should be set in order to increase the quality of data. Yet, this requires certain amount of domain knowledge. A detailed data cleaning process for TEA data set is presented in section 5.2.

2.4.2. Scale conversion and data transformation

Scale conversion is one of the most important steps in majority of the data mining studies. It is known that in different types of the classification analysis, variables which define the objects may not be measured or observed in the same scales (e.g. some can be interval type while others are categorical). To apply most of the classification methods, different variables with different scales should be converted to the same scale. However, considering the fact that measures of dissimilarity (such as Euclidean distance) are highly sensitive to huge differences in magnitudes of input data, converting variables into the same scale may not be a good idea for the analysis. Thus, data transformation techniques may need to be applied (Jain and Dubes 1988; Gan et al., 2007).

Different types of scale conversion techniques are described in detail by Anderberg (1973). Some major techniques are conversion from nominal to ordinal, nominal to interval, ordinal to nominal, ordinal to interval, interval to nominal, interval to ordinal and binarization. In this thesis, we mainly used scale conversion from interval to ordinal.

Conversion from interval to ordinal generates equally ranked objects within a category while possessing the ordinal relation among objects in different categories. However by using this conversion, information on the size of the differences among objects in different categories and information on the differences among objects in the same category are lost. Eleven different methods are suggested for scale conversion from interval to ordinal such as linear discriminant function, one-dimensional hierarchical linkage method, Ward's hierarchical clustering method and Cochran and Hopkins method (Anderberg, 1973; Gan et al., 2007).

Data standardization techniques should be applied with caution since they may cause information loss in scale and location of data. However, these techniques should be applied in case of having high amounts of difference in the magnitudes of input variables due to its effect on dissimilarity measures such as Euclidean distance (Milligan and Cooper, 1988). Before applying classification methods for a given data set, it is worth to know the data collection method and types of variables in input data for avoiding standardization errors which results in high amounts of information loss (Gan et al., 2007).

Depending on the type of data, some well-known data standardization methods are mean, median, standard deviation, range, *z*-score, Huber's estimate (for further details see Milligan and Cooper, 1988, and Jain and Dubes, 1988).

Data transformation techniques mainly focus on the whole data set rather than each variable one by one. Some well-known data transformation techniques are Principle Component Analysis (PCA), Singular Value Decomposition (SVD) and The Karhunen-Loève (K-L) Transformation. However, majority of these methods are used for numerical data sets (Gan et al., 2007).

2.4.3. Dealing with missing values

Dealing with missing values in data mining studies is one major area of interest. As it is known, in most of the real-world data sets we face two main problems: data containing useful information can be missing or there can be errors in the data sets. If most of the measurements of a record are missing then this record should be removed from the data set (Kaufman and Rousseeuw, 1990). Fujikawa and Ho (2002) emphasized three major cases in which missing values can occur. They might occur in many variables (several attributes or columns), in a number of records (instances or rows) or they might occur randomly both in many variables and in a number of rows. In order to deal with missing values, two major groups of methods are used. These are prereplacing methods and embedded methods. In prereplacing methods, missing values are replaced before the classification process take place whereas in embedded methods missing values are handled during the classification process.

In literature, different methods are used to overcome the missing value problems. Fujikawa and Ho (2002) investigated the methods employed in dealing with missing values (Table 1) and proposed three different cluster-based algorithms which rely on the mean-and-mode method. These methods are Natural Cluster Based Mean-and-Mode algorithm (NCBMM), Rank Cluster Based Mean-and-Mode algorithm (RCBMM) and *k*-Means Cluster-Based Mean-and-Mode algorithm (KMCMM).

Table 1. List of Methods Used for Dealing with Missing Values

Method	**Group**	**Attributes**	**Case**	**Cost**
Mean-and-mode method	Pre-replacing	Numerical & Categorical	Case 2	Low
Linear regression	Pre-replacing	Numerical	Case 2	Low
Standard deviation method	Pre-replacing	Numerical	Case 2	Low
Nearest neighbor estimator	Pre-replacing	Numerical & Categorical	Case 1	High
Decision tree imputation	Pre-replacing	Categorical	Case 1	Middle
Autoassociative neural net.	Pre-replacing	Numerical & Categorical	Case 1	High
Casewise deletion	Embedded	Numerical & Categorical	Case 2	Low
Lazy decision tree	Embedded	Numerical & Categorical	Case 1	High
Dynamic path generation	Embedded	Numerical & Categorical	Case 1	High
C4.5	Embedded	Numerical & Categorical	Case 1	Middle
Surrogate split	Embedded	Numerical & Categorical	Case 1	Middle

NCBMM method cannot be directly applied in unsupervised clustering; on the other hand RCBMM and KMCMM can be used in both supervised and unsupervised data clustering. Depending on the type of the attribute, NCBMM divides the observations into clusters by the help of class attribute and missing values of observations are filled by using the mode or mean of each cluster. RCBMM is independent of class attribute and it is used to fill in missing values for categorical attributes. This method includes three major steps. In the first step, a ranking algorithm is employed in which the distance of all categorical attributes from a missing one is measured. The one having the shortest distance is used for clustering. In the second step, all observations are divided into clusters and finally missing values are filled by each cluster mode.

KMCMM is also independent of class attributes and it is used to fill in missing values for numerical attributes. This algorithm uses the absolute correlation coefficient between all numerical attributes and the missing value *a*, then it ranks them in an increasing order. Next, based on the missing value *a*, *k*-means algorithm (Macqueen, 1967; Hartigan, 1975) is used to divide data into *k* clusters. Finally mean of each cluster is used to fill in missing attribute on *a* (Fujikawa and

Ho, 2002; Gan et al., 2007).

In our thesis we mainly used RCBMM and KMCMM for filling missing values of categorical and numerical attributes, respectively. There are other methods applicable for dealing with missing values in data mining studies. These studies are listed in Table 1, where missing values occurring in several variables are defined as case 1 and the ones occurring in a number of records are defined as case 2 (Fujikawa and Ho, 2002).

2.5. Feature Selection Methods

Majority of the classification algorithms use feature selection methods to increase classification accuracy and efficiency by removing irrelevant or redundant features. Main concern in these methods is to select subset of informative features to optimize solution (i.e. features which are most relevant to classification purpose) instead of using whole feature space which results in a high computational cost. Other benefits of feature selection are decreasing storage requirements, reducing training and optimization times and overcoming the curse of dimensionality (see section 2.4) (Guyon and Elisseeff, 2003; Al Aghbari, 2010).

There is an increasing trend in variable and feature selection methods in recent years (Blum and Langley, 1997; Kohavi and John, 1997; Guyon and Elisseeff, 2003). Majority of previous works include selection of features for smaller domains (i.e. feature up to 40). However, todays data mining studies might include domains which have tens of thousands of feature. For example classification studies concerning gene selection from a DNA microarray data requires subsetting variables approximately 50,000. Thus, todays feature selection methods should work with larger domains and overcome possible redundancy and irrelevance among features depending on classification goal.

There are two major types of feature selection methods used in literature: filter approach and wrapper approach. In filter approach major concern is to find set of informative features and remove (filter out) less relevant ones by using certain statistical analysis (such as ranking by correlation coefficients). Then, classification algorithm is run on the set of selected features. Filter approach is independent of the algorithm used for classification and it serves as a preprocessing step. In wrapper approach, set of features are searched by using cross-validation and classification performance is compared with respect to each selected sets. Then, the set that maximizes the classification performance is selected. Wrapper approach comes with a better classification performance but it results in higher computation cost than filter approach. Basically, filter approach tries to obtain a set of features which maximizes classification accuracy while wrapper approach selects set of features which minimizes classification error (Guyon and Elisseeff, 2003; Liu and Kender, 2003; Al Aghbari, 2010).

In order to decrease computation and storage cost with large-scaled micro level categorical data sets a suitable filter approach can be used. In this thesis, a new IFS method is proposed which is based on information theory and a well-structured graphical solution. Our method is independent of the classification algorithm used and it is categorized under filter approach (see section 3.1.2.1).

2.6. Difference Metrics

Difference metrics are the key elements of classification algorithms. Efficiency of these algorithms (scalability, running time, computation cost, performance, etc.) highly depends on the difference metric used. Some well-known difference metrics for categorical data sets are described in this section.

Overlap Metric (OM) is one of the most widely used distance metrics with nominal attributes and it simply evaluates the sum of mismatches among attributes of two objects x & y (Wilson and Martinez, 1997). It is given as:

$$OM(x,y) = \sum_{i=1}^{d} \delta\big(a_i(x), a_i(y)\big) \qquad (3)$$

with

$$\delta\big(a_i(x), a_i(y)\big) = \begin{cases} 0 & \text{if } a_i(x) = a_i(y) \\ 1 & \text{if } a_i(x) \neq a_i(y) \end{cases}$$’

where d is the number of attributes, $a_i(x)$ and $a_i(y)$ are the attribute values of x and y, respectively. OM is a metric function and clearly it satisfies the following three conditions:

- $OM(x, y) = 0$, iff $x = y$,
- $OM(x, y) = OM(y, x)$ (symmetry)
- $OM(x, z) \leq OM(x, y) + OM(y, z)$ for any $x, y, z \in \Omega_d$ (categorical sample space).

Another famous difference metric for categorical data is Value Difference Metric (VDM) (Stanfill and Waltz, 1986). It is defined as;

$$VDM(x, y) = \sum_{i=1}^{d} \sum_{j=1}^{l} \left| P\left(c_j | a_i(x)\right) - P\left(c_j | a_i(y)\right) \right|, \qquad (4)$$

where d is the number of attributes, l is the number of classes, c_j is the output class, $a_i(x)$ and $a_i(y)$ are the attribute values of x and y, respectively. $P\left(c_j | a_i(x)\right)$ is the conditional probability of output class c_j of x given attribute A_i has the value of $a_i(x)$. VDM makes use of the idea that difference between two values of an attribute is smaller provided that they are correlated with the class variable in a similar manner. In a more explicit form VDM uses the correlation between the value of A_i and class c_j rather than the values of attributes. By using this idea another metric named as Frequency Difference Metric (FDM) is proposed (Jiang et al., 2014):

$$FDM(x, y) = \sum_{i=1}^{d} \sum_{j=1}^{l} \left| F\left(c_j, a_i(x)\right) - F\left(c_j, a_i(y)\right) \right|, \qquad (5)$$

where $F\left(c_j, a_i(x)\right)$ and $F\left(c_j, a_i(y)\right)$ represent the joint frequencies of class label c_j and attribute values $a_i(x)$ and $a_i(y)$ respectively. There is a strong relation between FDM and VDM. In a more explicit form,

$$\begin{aligned} VDM(x,y) &= \sum_{i=1}^{d} \sum_{j=1}^{l} \left| \frac{P\left(c_j, a_i(x)\right)}{P\left(a_i(x)\right)} - \frac{P\left(c_j, a_i(y)\right)}{P\left(a_i(y)\right)} \right| \\ &= \sum_{i=1}^{d} \sum_{j=1}^{l} \left| \frac{F\left(c_j, a_i(x)\right)/n}{F\left(a_i(x)\right)/n} - \frac{F\left(c_j, a_i(y)\right)/n}{F\left(a_i(y)\right)/n} \right| \\ &= \sum_{i=1}^{d} \sum_{j=1}^{l} \left| \frac{F\left(c_j, a_i(x)\right)}{F\left(a_i(x)\right)} - \frac{F\left(c_j, a_i(y)\right)}{F\left(a_i(y)\right)} \right|. \end{aligned} \quad (6)$$

It can be inferred from (5) and (6) that VDM is a weighted version of FDM where weights are $F\left(a_i(x)\right)$ and $F\left(a_i(y)\right)$. It is mentioned that in some cases $F\left(a_i(x)\right)$ and $F\left(a_i(y)\right)$ can be very small and this results in an undefined or very large fractions in VDM. Thus, FDM is argued to be more efficient than VDM (Jiang et al., 2014). Although experimental results might show FDM’s superiority to VDM in some data sets, removing weights from VDM is theoretically unjustifiable. Moreover, both FDM and VDM use the assumption of attribute independence.

Another common distance metric for nominal attributes which takes dependence structure among attributes into account is One Dependence Value Difference Metric (ODVDM) (Li and Li, 2011). It uses Bayesian Network Classifiers for identifying the dependence relation between attributes. It is given as;

$$ODVDM(x,y) = \sum_{i=1}^{d} \sum_{j=1}^{l} \left| P\left(c_j | a_i(x), a_{ip}(x)\right) - P\left(c_j | a_i(y), a_{ip}(y)\right) \right|, \quad (7)$$

where $a_{ip}(x)$ and $a_{ip}(y)$ show the value of parent nodes for A_{ip} and other parameters are the same as defined in (4).

A further study is carried on by Jiang and Li (2013) and they proposed Augmented Value Difference Metric (AVDM) which takes into account the dependence structure of attributes. It is defined as follows:

$$AVDM(x,y) = \sum_{i=1}^{d-1} \sum_{l=i+1}^{d} \sum_{j=1}^{l} \left| P\left(c_j|a_i(x), a_l(x)\right) - P\left(c_j|a_i(y), a_l(y)\right) \right|, (8)$$

where $a_i(x)$ and $a_l(y)$ are the attribute values of A_i and A_l. Due to estimating pairwise dependence structure of each attributes computational complexity of the AVDM is higher than VDM and ODVDM. Besides, another critical issue is that pairwise independence does not necessarily imply the independence of all attributes.

Difference metrics stated above (VDM, FDM, ODVDM and AVDM) do not take into account the attribute importance (or relevance) for target variable. VDM (with modified versions) or FDM uses the correlation structure of attributes with the output class individually. ODVDM and AVDM use pairwise dependence relation of attributes with the output class. For solving this issue another difference metric is proposed by Li et al. (2013). This metric uses mutual information between attributes and target variable C and named as Attribute Weighted Value Difference Metric (AWVDM). It is given as;

$$AWVDM(x,y) = \sum_{i=1}^{d} \sum_{j=1}^{l} I(A_i, C) \left| P\left(c_j|a_i(x)\right) - P\left(c_j|a_i(y)\right) \right|, (9)$$

where $I(A_i, C) = w_i$ are used as weights. However, mutual information can be greater than one (see section 3.1.1) and it favors attributes with more values. In addition, weighting scheme used in AWVDM does not take the dependence structure among attributes into account. It only uses the amount of information that each attribute has on target variable C. In other words, AWVDM favors attributes which are highly correlated with the output class and this will possibly result in overestimating the effect of attributes which has a high rate of dependence.

As far as we know, a difference metric which takes into account attribute importance (relevance) with class variable and correlation structure among all attributes at the same time does not exist, and it is proposed in this thesis study (see section 3.1.2.3).

2.7. Evaluating Classification Performance

Major concern in classification algorithms is finding a model which truly classifies objects in data set. In order to evaluate the quality of different classification algorithms different methods are used in literature depending on the classification problem (i.e., the characteristic of problem and data set). Generally, evaluation measures are defined by a 2×2 contingency table which is commonly known as confusion matrix (see Table 41 in Appendix D). A confusion matrix represents number of objects which are correctly or incorrectly classified for each class. Some well-known performance measures obtained by confusion matrix are accuracy, sensitivity, specificity, prevalence and detection rate (Costa et al., 2007; Powers, 2011).

Sensitivity (recall or true positive rate) shows the proportion of true positive cases which are correctly predicted as positive. Sensitivity is used to define how many of the relevant cases are defined by true positive rate. Information that it provides is interpreted differently depending on the case it is used. In information retrieval it is not highly valued. In computational linguistic it can be neglected. However, in medical research it is noted as one of the primary performance measures.

Specificity (inverse recall or true negative rate) represents the proportion of true negative cases which are correctly predicted as negative. Specificity simply shows the test performance on detecting objects without an event (e. g. patients identified as healthy who are known not to have disease in medical research) (Powers, 2011).

Positive predictive value is defined as the number of true positive cases over summations of true positive and false positive cases. In other words, it is the ratio of true positives divided by all cases predicted as positive. Similarly, negative predictive value represents the number of false positive cases over number of all cases predicted as negative (Fletcher et al., 2012).

Prevalence is used to show the proportion of true positive and false negative cases over all cases. In a more explicit form, it represents the frequency of events over all cases.

Detection rate represents the proportion of true positive cases which are correctly predicted as positive over all cases. Similar to sensitivity, in medical experiments higher values of detection rate is primary objective since idea is to obtain all real positive cases. On the other hand, detection prevalence shows the cases which are predicted as positive over all cases.

Accuracy (or Rand Accuracy) is used to define sum of true positive and true negative cases over all cases. It is one of the major performance measures used to check the quality of the classification. Balanced accuracy is simply the arithmetic mean of sensitivity and specificity.

Although these performance measures are interpreted differently depending on the classification purpose, higher values in accuracy commonly show higher quality in classification (see Appendix D for mathematical definitions of these performance measures).

CHAPTER 3

PROPOSED METHOD FOR CLASSIFICATION

In this part of the thesis, our proposed CACDIFES algorithm and experimental results by using this algorithm are presented. In the first section, CACDIFES Algorithm is introduced and in the second section series of experiments for testing the efficiency of CACDIFES algorithm are given.

CACDIFES Algorithm is based on information theory and it consists of three main phases: Incremental Feature Selection (IFS) phase, Data Compression (DC) phase and Classification phase (Figure 3).

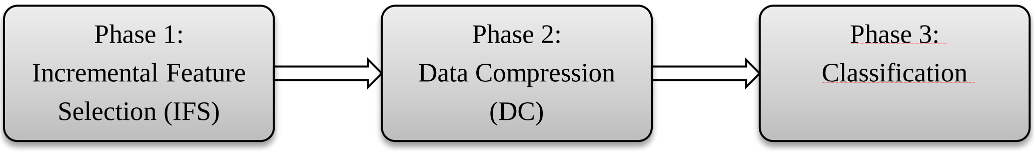

Figure 3. Phases of CACDIFES Algorithm

In section 3.1.1, tools such as entropy, mutual information and symmetric uncertainty which are used to develop CACDIFES algorithm (specifically first phase of the algorithm) are briefly described. In section 3.1.2, three phases of

CACDIFES algorithm, given in Figure 3, are introduced. Finally, our proposed difference metric, IWVDM which is used in phase 3, is presented in the same section.

In section 3.2, efficiency of our proposed difference metric IWVDM is compared with other three well-known difference metrics, VDM, FDM and OM, via series of experiments by using 9 different data sets obtained from UCI Machine Learning Repository.

3.1. CACDIFES Algorithm

In this section, we introduce notation that will be used throughout the thesis. Let $D = \{(x, C)\}$ be a categorical data set with n objects, where C shows the class membership (target variable or class variable C) with number of classes l, and each x is an instance represented by a vector of attribute values $\{a_1(x), a_2(x), \dots, a_d(x)\}$. Here, d shows the number of attributes (number of dimensions), $a_j(x)$ shows the value of jth attribute A_j of x. A is a set of d categorical attributes $A = \{A_1, A_2, \dots, A_d\}$ and domain of each $A_j \in A$ $(1 \le j \le d)$ is finite.

For illustrating the idea, consider a 3-dimensional data set with 10 objects representing the certain specifications of job applicants given in Table 2.

Table 2. Representation of Example Data Set

Data Object	A_1 **(Gender)**	A_2 **(Marital St.)**	A_3 **(Disability St.)**	C **(Employability St.)**
x_1	Male	Married	Yes	Weak Emp. Opp.
x_2	Male	Married	No	Good Emp. Opp.
x_3	Male	Married	Yes	Weak Emp. Opp.
x_4	Male	Married	No	Good Emp. Opp.
x_5	Female	Single	Yes	Fair Emp. Opp.
x_6	Female	Single	Yes	Fair Emp. Opp.

Table 2. (Continued)

Data Object	A_1 **(Gender)**	A_2 **(Marital St.)**	A_3 **(Disability St.)**	C **(Employability St.)**
x_7	Female	Single	Yes	Fair Emp. Opp.
x_8	Female	Single	Yes	Fair Emp. Opp.
x_9	Male	Married	No	Good Emp. Opp.
x_{10}	Male	Single	Yes	Fair Emp. Opp.

Here A is a set of 3 categorical attributes, i.e., $A = \{A_1, A_2, A_3\}$ with, $A_1 = \{Male, Female\}$, $A_2 = \{Married, Single\}$ and $A_3 = \{Yes, No\}$. Now let us assume that we have three classes (target variable) with c_1, c_2 and c_3 representing weak employment opportunities, good employment opportunities and fair employment opportunities of the applicants, respectively.

As it can be inferred, full dimensional data space (categorical sample space Ω_d) is constituted by the Cartesian product of the domain of each attribute $\Omega_d = A_1 \times A_2 \times \ldots \times A_d$. In the given example Ω_d has 8 elements. Searching for full-dimensional space and discovering dense points might be quite difficult as the number of dimensions increase. Technically, for a binary coded 20 dimensional categorical data set, total number of dense points is $2^{20} = 1{,}048{,}576$ which might be greater than the total number of objects (n). Although this number is theoretically quite high (in the case where each object is located in distinct points), majority of the time objects are located nearby or at the same point and an important subset of Ω_d covering all points will be sufficient.

In addition, traditional approaches do not take into account the dependence structure among attributes which is not realistic with many real-world data sets. For example, a candidate job applicant can only possess a specific certification if s/he holds a certain degree of education (or above) which clearly indicates a correlation structure between two attributes (education level vs. existence of certain certification). Hence, attribute independence assumption is usually not valid while working with many real life data sets (Li and Li, 2011; Li et al., 2013, Jiang and Li, 2013).

In order to cope with these issues we present a Classification Algorithm for Categorical Data with Incremental Feature Selection (CACDIFES). We base our algorithm on information theory. Therefore, some main concepts on information theory which are used to develop CACDIFES are given in section 3.1.1.

3.1.1. Entropy, mutual information and symmetric uncertainty

Entropy (or Shannon's Entropy (Shannon, 1948)) is a measure of uncertainty of a random variable X. In a more explicit form entropy of a discrete variable X which has a probability distribution $p_X(x)$ is given by (Cover and Thomas, 2006);

$$H(X) = -\sum_{x \in X} p_X(x) \log(p_X(x)). \tag{10}$$

Bear in mind that if log base is taken as 2, represented entropy is in bits and if natural log base is used then entropy is in nats. Equation (10) can be expressed as $H(X) = E[-\log(p_X(x))]$. Joint entropy of two discrete random variables X and Y can be obtained as follows;

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p_{XY}(x,y) \log(p_{XY}(x,y)), \tag{11}$$

where $p_{XY}(x,y)$ is the joint probability mass function of X and Y. Joint entropy of X and Y is sub-additive, i.e. $H(X,Y) \leq H(X) + H(Y)$ and equality holds if X and Y are independent. From statistical perspective, we might infer that if X and Y are dependent, then they will result in a reduction in their joint entropy. By introducing the conditional entropy $H(X|Y=y)$, we can identify the uncertainty about X when the outcome y is obtained.

$$\begin{aligned} H(X|y) &= -\sum_{x \in X} p_{X|Y}(x|y) \log(p_{X|Y}(x|y)) \\ &= -E\big[\log(p_{X|Y}(x|y))\big]. \end{aligned} \tag{12}$$

By averaging equation (12) over $y \in Y$ we get,

$$\begin{aligned} H(X|Y) &= -\sum_{x\in X}\sum_{y\in Y} p_Y(y)p_{X|Y}(x|y)\log(\,p_{X|Y}(x|y)) \\ &= -\sum_{x\in X}\sum_{y\in Y} p_{XY}(x,y)\log(\,p_{X|Y}(x|y)) \\ &= H(X,Y) - H(Y). \end{aligned} \tag{13}$$

In a similar manner, $H(Y|X) = H(X,Y) - H(X)$. There is a significant connection between Kullback-Leibler (KL) divergence and entropy (Akaike, 1973, 1974). KL divergence (or relative entropy) is a measure for identifying the difference of two probability distributions over the same space. For two probability distributions P and Q, KL divergence over the same set is defined as follows:

$$D(P||Q) = \sum_{x\in X} P(x)\log\left(\frac{P(x)}{Q(x)}\right). \tag{14}$$

As the KL divergence gets smaller, probability distributions P and Q becomes closer and vice versa. However, KL divergence is not a symmetric distance measure. A symmetric distance metric is generated from KL divergence named as KL distance metric:

$$D(P||Q) = \sum_{x\in X}\left[(P(x) - Q(x))\log\left(\frac{P(x)}{Q(x)}\right)\right]. \tag{15}$$

KL distance can be defined as the natural distance function from a true probability distribution to a target probability distribution (Kullback and Leibler, 1951; Kullback, 1959). It is one of the most widely used distance metric in speech processing, statistical language modeling, information retrieval, etc (Bigi, 2003).

Following the definition of entropy and conditional entropy, now we will define the mutual information (information gain) between X and Y. Simply, mutual information is used as a measure of dependence between two random variables.

$$\begin{aligned} I(X,Y) &= \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\left[\frac{P(x,y)}{P(x)P(y)}\right] \\ &= H(X)+H(Y)-H(X,Y). \end{aligned} \tag{16}$$

Mutual information has advantages compared to correlation coefficient since it does not only measure the linear dependence, and variables used in the estimation do not need to be Euclidean (Brillinger, 2004).

Higher order mutual information can be defined as;

$$I(X,Y,Z) = H(X)+H(Y)+H(Z)-H(X,Y,Z). \tag{17}$$

Mutual information has the following properties;

1- $I(X,Y) = I(Y,X)$ (Symmetry)
2- $I(X,Y) \geq 0$ (Equality holds if X and Y are independent)
3- $I(X,Y) = I(U,V)$ if $u = u(x)$ and $v = v(x)$ are individually 1-to-1 measurable transformations (Invariance).

Mutual information can be used as a correlation measure between two random variables, yet it can take values greater than one. Besides, this measure favors variables with more values (Yu and Liu, 2003). In order to overcome these issues, symmetric uncertainty is introduced. Symmetric uncertainty of two random variables is given as;

$$U(X,Y) = 2\left[\frac{H(X)+H(Y)-H(X,Y)}{H(X)+H(Y)}\right] = 2\left[\frac{I(X,Y)}{H(X)+H(Y)}\right]. \tag{18}$$

It can be inferred that if X and Y are independent, then $H(X)+H(Y)=H(X,Y)$, and $U(X,Y)=0$. If X completely defines Y, or vice a versa, then $H(X)=H(Y)=H(X,Y)$ which results in $U(X,Y)=1$. This guarantees that $0 \leq U(X,Y) \leq 1$ (Press et al., 1988).

3.1.2. Phases of CACDIFES

Algorithm of CACDIFES consists of three main phases: Incremental Feature Selection (IFS) phase, Data Compression (DC) phase and classification phase. These phases will be described in the following sections.

3.1.2.1. Incremental feature selection phase

In this part of the algorithm an IFS method is proposed by using information theory. Intuitively, if an attribute A_j does not contain information about the class variable C, then they are independent (i.e. $I(C, A_j) = 0$). We will make use of this fact and try to investigate if our attributes contain information (pairwise mutual info) about our target variable C. For discovering if our entire set of attributes $A_1, A_2, \dots, A_d$ contain information about target variable C, we need to find

$$\begin{aligned} I(C; A_1, A_2, \dots, A_d) &= H(C) + H(A_1, A_2, \dots, A_d) - H(C, A_1, A_2, \dots, A_d) \\ &= H(C) - H(C|A_1, A_2, \dots, A_d). \end{aligned} \qquad (19)$$

Mutual information for C and entire set of attributes can be defined by chain rule:

$$I(C; A_1, A_2, \dots, A_d) = I(C, A_1) + \sum_{j=2}^{d} I\big(C; A_j | A_1, \dots, A_{j-1}\big). \qquad (20)$$

It is obvious that adding more attributes increases the joint entropy ($I(C; A_1, A_2) \geq I(C, A_1)$). Consider the case where $d = 2$.

$$\begin{aligned} I(C; A_1, A_2) &= I(C, A_1) + I(C; A_2|A_1) \\ I(C, A_1) &= H(C) - H(C|A_1) \text{ (from 19)} \\ I(C; A_2|A_1) &= I(C; A_2, A_1) - H(C) + H(C|A_1) \\ &= [H(C) - H(C|A_1, A_2)] - H(C) + H(C|A_1) \\ &= H(C|A_1) - H(C|A_1, A_2) \\ I(C; A_1, A_2) &= I(C, A_1) + I(C; A_2|A_1) \text{ (from first line)} \\ &= H(C) - H(C|A_1) + H(C|A_1) - H(C|A_1, A_2) \\ &= H(C) - H(C|A_1, A_2). \end{aligned} \qquad (21)$$

If $I(C;A_1,A_2) > I(C,A_1) + I(C,A_2)$, then there exists a positive relation between A_1 and A_2 which provides higher information on target variable C when both A_1 and A_2 are used together. Converse is also true for negative relation between A_1 and A_2. Since, our main aim is to include attributes which contain high information on class C with low rate of interdependence, we need to find attributes which maximizes the joint mutual information while controlling attribute dependence.

We introduce an IFS method which satisfies these conditions: First, we estimate the mutual information $I(C,A_j)$ between each attribute A_j for $(j = 1, \ldots, d)$ and the target variable C. Second, a hypothesis test is employed to identify if $I(C,A_j)$ is statistically different from 0 or not.

$$\begin{aligned} &H_0\colon I(C,A_j) = 0 \\ &H_1\colon I(C,A_j) \neq 0. \end{aligned} \tag{22}$$

Under the null hypothesis (i.e. if A_j and C are independent), $2n\hat{I}(C,A_j)$ is asymptotically distributed as chi-square with $\chi^2_{(t_j-1)(l-1)}$ (Christensen, 1997) where t_j shows the number of categorical values that A_j can take, l is the number of classes in C and n is the number of observations. Depending on the results of hypothesis tests, we identify attributes which contain information on C. Suppose that a subset of attributes $SA = \{A_1, \ldots, A_s\}$ $(s \leq d)$ is found to be informative for class C (where they are sorted depending on their information level on C in descending order).

In the next step, we introduce a forward selection method and a stopping criteria by using the importance of contribution of additional attribute to the total mutual information for target variable C. As specified in the above discussion, adding more attributes increases the joint mutual information. Thus, we need a proper method to see if the additional information from new attribute to the mutual information is statistically significant or not. Set of attributes which maximizes the joint mutual information of attributes and class variable, $\text{argmax}_{A_j \in A}\, I(C; A_1, \dots, A_j)$, is the same as those that minimizes the conditional mutual information of target variable and additional feature given the selected features, $\text{argmin}_{A_j \in A}\, I(C; A_j | A_1, \dots, A_{j-1})$. In a more explicit form;

$$\text{argmax}_{A_j \in A}\, I(C; A_1, \dots, A_j) = \text{argmin}_{A_j \in A}\, I(C; A_j | A_1, \dots, A_{j-1}), \quad (23)$$

where $j = 1, \dots, s$. Intuitively, adding new attributes will surely increase (decrease) the magnitude of joint mutual information (conditional mutual information) but after certain number of attribute addition, this increase (decrease) will tend to be very small and eventually, an additional attribute will not provide statistically significant information on C. We employ a graph-based approach to satisfy the condition stated in (23). We plot the joint mutual information of informative attributes vs. informative attributes and conditional mutual information vs. informative attributes (For instance, see Figures 9 and 10 in section 5.3). These two graphs provide an intuitive idea for determining the attributes which contain statistically significant information on C.

However, best approach is to define a hypothesis testing procedure to see if $I(C; A_{j+1}, A_1, \dots, A_j) - I(C; A_j, A_1, \dots, A_{j-1}) = 0$ or not. From Sun Han (1980) and Srinivasa (2003), we know that; $2n\hat{I}(C; A_{j+1}, A_1, \dots, A_j) \sim \chi^2_{r_{t_{j+1}}(l-1)}$ for large n and under the semi-independence of C and $A_1, \dots, A_{j+1}$ where $r_{t_j} = \prod_{i=1}^{j} \left(r_{t_j} - 1\right)$ shows the number of categories that the joint distribution of $A_1, \dots, A_j$ can take. Yet, the distribution of $I(C; A_{j+1}, A_1, \dots, A_j) - I(C; A_j, A_1, \dots, A_{j-1})$ is mathematically intractable. As an alternative to hypothesis testing procedure, we use another graph for the differences of estimated joint mutual information (For instance, see Figure 11 in section 5.3).

By using graphical inspection, we employ a user-defined threshold β for controlling the contribution of additional attribute to the joint mutual information of target variable and previously added attributes. Algorithm will stop if $\frac{\hat{I}(C;A_1,\ldots,A_j)}{\hat{I}(C;A_1,\ldots,A_s)} \geq \beta$ where $j = 1, \ldots, s^* \leq s$. If β is chosen as too small, then the overall information provided by the informative attributes would not be acquired. Conversely, if it is chosen as too large, then we might face the risk of selecting irrelevant attributes. Experimental results of CACDIFES Algorithm reveal that choosing β between 0.90 and 0.95 provides satisfactory results.

Next step is to identify the dependence structure of informative attributes and remove the redundant ones by using symmetric uncertainty. Due to symmetry, we have $\binom{s^*}{2} = s^*(s^* - 1)/2$ number of symmetric uncertainty estimates. Symmetric Uncertainty (**U**) matrix of the informative attributes is given by;

$$\mathbf{U} = \begin{bmatrix} 1 & \cdots & U(A_1, A_{s^*}) \\ \vdots & \ddots & \vdots \\ U(A_{s^*}, A_1) & \cdots & 1 \end{bmatrix}_{s^* \times s^*}. \qquad (24)$$

Using the similar idea as in multicollinearity in regression analysis, attributes with high rate of symmetric uncertainty will be removed. Since exact distribution of symmetric uncertainty is unknown and there is not a statistical test for testing whether it is different from 0 or above some certain threshold value, we present a heuristic procedure for this purpose. Bear in mind that eliminating attributes will result in a decrease in information ratio. Hence, we consider the percentage of decrease in information ratio while eliminating the attributes with dependence structure with the rest. In our analysis we suspect dependence structure if at least two symmetric uncertainty estimate of an attribute is greater than 0.5.

3.1.2.2. Data compression phase

In this phase, a Data Compression (DC) algorithm is employed by using the final features obtained in IFS phase. The term “Data Compression” in this thesis is referred to “removal of duplicated records” and should not be confused with the definition of data compression in computer science which involves encoding information by using fewer bits than the original version. Having the total number of attributes decreased from d to s' in IFS phase, we are expecting to discover more similar objects in data space. Now we introduce the concept of Representative Objects (ROs). ROs are defined as data points in D for which $d(x,y)=0$ for $\forall\, x,y \in D$ where $x \,\&\, y$ show the objects in D. Distance function, $d(.,.)$, is the OM given in (3) in section 2.6. OM clearly helps discovering ROs in a dataset.

Due to the nature of categorical data set (i.e. no ordering is present), the levels of the attributes are exchangeable (Gordon, 1999). Thus, any two or more points having OM as 0 are considered as $RO_t \in \Omega_r$ with $\Omega_r \subset \Omega_{s'}$, where $t = 1, \dots, m$ with $1 \leq m \leq n$. We are expecting m to be much smaller than n. For our categorical data space $\Omega_{s'}$, we can clearly define a metric space by $(\Omega_{s'}, s')$.

Majority of the algorithms require overall data to be stored in main memory at a time which results in memory space issues. Although these non-incremental algorithms are efficient in many aspects they do not scale well with large-scaled data sets due to high amount of memory requirements (Aranganayagi and Thangavel, 2010). Obtaining ROs by using a non-incremental algorithm seems impractical since dataset should be scanned repeatedly.

Algorithm: Data Compression (DC)

```
Step 1: Initialize $x_1 \rightarrow RO_1 \,\&\, j = 1, m = 1$
Step 2: Assign $x_i \rightarrow RO_j \,\&\, m = j$, if $d(RO_j, x_i) = 0$, for $i = 2, \dots, n \,\&\, j = 1, \dots, m$
Step 3: Else assign $x_i \rightarrow RO_{j+1} \,\&\, m = j + 1$
```

Figure 4. Data Compression (DC) Algorithm

In order to overcome the issues stated in the previous page, we present an incremental DC algorithm in Figure 4. Bear in mind that first point is chosen as seed. For assigning each x_i to its representative objects, data set is scanned for once and either x_i is assigned to current RO_j or it is assigned to RO_{j+1}. For decreasing the computational complexity, proper initialization points can be chosen as seed. For example, modes of each attribute can be found and data objects covering majority of the modes can be chosen as seed.

3.1.2.3. Classification phase

In this phase, we adopt a difference metric similar to VDM. As it is discussed in section 2.6, difference metrics such as VDM, FDM, ODVDM and AVDM do not take into consideration the attribute importance for target variable. VDM type of metrics only uses correlation structure of attributes with the output class individually. Some make strong assumptions on attribute independence. Another metric, AWVDM, allow for identifying dependence relation of attributes with the output class, but it does not account for finding dependence structure among attributes.

In order to overcome these issues we introduce Independently Weighted Value Difference Metric (IWVDM). We assign weights by using symmetric uncertainty between attributes and class variable C, but the major difference between AWVDM and IWVDM is that, IWVDM takes the dependence structure among attributes into account. IWVDM is defined as follows:

$$IWVDM(x, y) = \sum_{i=1}^{s\prime} \sum_{j=1}^{l} w_i \left| P\left(c_j | a_i(x)\right) - P\left(c_j | a_i(y)\right) \right|. \qquad (25)$$

In feature selection phase, we obtained a smaller number of s' informative attributes which do not have a dependence structure $SA' = \{A_1, \dots, A_{s\prime}\}$ where they are sorted in descending order. We know that VDM types of different metrics are sensitive to irrelevant (or redundant) attributes and removing them increases the efficiency of these metrics. Now, we will estimate the weights of attributes by using symmetric uncertainty as:

$$\begin{aligned}
&w_1 = U_1(C, A_1),\\
&w_2 = U_2(C; A_2, A_1) - w_1,\\
&\qquad \vdots\\
&w_{s'} = U_{s'}(C; A_{s'}, A_1, \dots, A_{s'-1}) - w_{s'-1}.
\end{aligned} \tag{26}$$

We will adjust the weights so that they will sum up to one. Our weights favor the attributes which contain high rate of information on class variable C. Now we will use IWVDM for class assignment of our objects. In order to use IWVDM for class assignment we need to transform our input space by using a probabilistic model. Kasif et al. (1998) proposed a method for transformation of nominal attributes into probability distribution (probability MBR transform) by assuming the conditional independence of the joint probability distribution as given below:

$$P(a_1(x), a_2(x), \dots, a_d(x), c) = P(c) \prod_{i=1}^{d} P(a_i(x)|c). \tag{27}$$

It can be inferred that equation (27) requires attribute independence. Considering we have d-dimensional input space and l number of classes, our transformed input space turns into ld-dimensional space. In a more explicit form, ith attribute value $a_i(x)$ is transformed into discrete probability distribution as

$$f_{MBR}\left(c_j|a_i(x)\right) = \begin{cases} P\left(c_j|a_i(x)\right), & \text{for } j = 1, \dots, l \\ 0, & \text{elewhere} \end{cases}. \tag{28}$$

An MBR transformation of an object x is defined as $\{P(c_1|a_1(x)), \dots, P(c_l|a_1(x)), P(c_1|a_2(x)), \dots, P(c_l|a_2(x)), P(c_1|a_d(x)), \dots, P(c_l|a_d(x))\}$. For better understanding, let's employ MBR to our example data set given in Table 2. Let the first 8 objects be our training set and the final 2 objects be our testing set. First object, $x_1 = \{Male, Married, Yes\}$ is transferred into $\left\{\langle\frac{2}{4},\frac{2}{4},\frac{0}{4}\rangle\,\langle\frac{2}{4},\frac{2}{4},\frac{0}{4}\rangle\,\langle\frac{2}{6},\frac{0}{6},\frac{4}{6}\rangle\right\}$ where the first number in the first vector is given as;

$$P(\text{Weak Emp. Opp.}\,|\text{Male}) = \frac{N_{W\cap M}}{N_M} = \frac{2}{4}, \qquad (29)$$

where $N_{W\cap M}$ represents the frequency of joint occurrence of weak employment opportunities and male, and N_M represents the frequency of male in training set.

First three numbers in “{ }” given above are related with weak employment opportunities. Out of these three numbers given in “< >”, the first one is related with gender being “male”, the second one is related with marital status being “married” and the third one is related with disability statues being disabled (“yes”). Similarly, second three numbers in “{ }” are related with good employment opportunities and third three are related with fair employment opportunities. All probability estimates for each object are given in Table 3.

Table 3. Joint Probabilities of Each Attribute and Class Labels

Data Object	**A_1: Gender**			**A_2: Marital Status.**			**A_3: Disability Status.**		
	c_1	c_2	c_3	c_1	c_2	c_3	c_1	c_2	c_3
x_1	0.5	0.5	0	0.5	0.5	0	1/3	0	2/3
x_2	0.5	0.5	0	0.5	0.5	0	0	1	0
x_3	0.5	0.5	0	0.5	0.5	0	1/3	0	2/3
x_4	0.5	0.5	0	0.5	0.5	0	0	1	0
x_5	0	0	1	0	0	1	1/3	0	2/3
x_6	0	0	1	0	0	1	1/3	0	2/3
x_7	0	0	1	0	0	1	1/3	0	2/3
x_8	0	0	1	0	0	1	1/3	0	2/3
x_9	0.5	0.5	0	0.5	0.5	0	0	1	0
x_{10}	0.5	0.5	0	0	0	1	1/3	0	2/3

We estimate the $IWVDM(x_i, x_j)$ for $i = 1,..,8$ and $j = 9, 10$ with $w_1 = 1/5$, $w_2 = 1/5$ and $w_3 = 3/5$ where w_1, w_2 and w_3 are the attribute weights of A_1, A_2 and A_3, respectively (w_1, w_2 and w_3 are chosen as an example). We have 3-dimensional data and 3 classes. $IWVDM(x_1, x_9)$ is given by;

$$IWVDM(x_1, x_9) = \sum_{i=1}^{3} \sum_{j=1}^{3} w_i \left| P\left(c_j | a_i(x)\right) - P\left(c_j | a_i(y)\right) \right|$$
$$= \frac{1}{5}\left(\left|\frac{2}{4} - \frac{2}{4}\right| + \left|\frac{2}{4} - \frac{2}{4}\right| + |0 - 0| \right) + \frac{1}{5}\left(\left|\frac{2}{4} - \frac{2}{4}\right| + \left|\frac{2}{4} - \frac{2}{4}\right| + |0 - 0| \right) + \frac{3}{5}\left(\left|\frac{1}{3} - 0\right| + |0 - 1| + \left|\frac{2}{3} - 0\right| \right) = 1.2. \qquad (30)$$

Table 4. IWVDM of Test Objects ($i = 1, \dots, 8$)

ith Object	**$IWVDM(x_i, x_9)$**	**$IWVDM(x_i, x_{10})$**	**Class Label of ith Object**
1	1.2	0.4	Weak Emp. Opp.
2	0	1.6	Good Emp. Opp.
3	1.2	0.4	Weak Emp. Opp.
4	0	1.6	Good Emp. Opp.
5	2	0.4	Fair Emp. Opp.
6	2	0.4	Fair Emp. Opp.
7	2	0.4	Fair Emp. Opp.
8	2	0.4	Fair Emp. Opp.

Table 4 shows the estimated IWVDM of test objects from (30). According to the estimated values, we can see that 3-nearest neighbor (3-NN) of x_9 are x_2, x_4 (with 0 distance) and x_1 or x_3 with an equal distance of 1.2. Furthermore, class labels of x_2 and x_4 are good employment opportunities and this label is also assigned to object x_9 (i.e., class label is assigned by using the mode of 3-NN objects' class labels). By using the same algorithm class label of x_{10} is assigned as fair employment opportunities (Since there exist 6 nearest neighbor for object 10 with equal distance of 0.4, and 4 objects out of 6 have class labels as fair employment opportunities).

As it is defined in section 2.4, *k*-NN algorithm has certain drawbacks. In order to address these issues, we initiate the algorithm by employing IFS and DC to increase its efficiency. Next, performance of *k*-NN highly depends on the distance metric used due to its distance-based learning nature (Mitchell, 1997). We use IWVDM as a distance function in *k*-NN and define CACDIFES in Figure 5.

Algorithm: CACDIFES

Incremental Feature Selection (IFS) Phase

Step 1: Estimate $I(C, A_j)$ from (19) for $j = 1, \dots, d$.

Step 2: Test if $I(C, A_j)$ is different from 0. Then, record $SA = \{A_1, \dots, A_s\}$ with $s \leq d$, depending on attributes' information level.

Step 3: Plot the graph of $\hat{I}(C; A_1, \dots, A_j)$ v.s. $A_1, \dots, A_j$ and $\hat{I}(C; A_j | A_1, \dots, A_{j-1})$ v.s. $A_1, \dots, A_j$ to discover A_j's which satisfy $\text{argmax}_{A_j \in A}\, I(C; A_1, \dots, A_j)$ and $\text{argmin}_{A_j \in A}\, I(C; A_j | A_1, \dots, A_{j-1})$ for $j = 1, \dots, s$.

Step 4: Subset from attributes till $\frac{\hat{I}(C; A_1, \dots, A_j)}{\hat{I}(C; A_1, \dots, A_s)} \geq \beta$, where $j = 1, \dots, s^* \leq s$.

Step 5: Identify dependence structure of informative attributes by using **U** and remove those whose two or more symmetric uncertainty estimate is greater than 0.5. Then, record $SA' = \{A_1, \dots, A_{s'}\}$ with $s' \leq s^*$. (Final set of informative attr.)

Data Compression (DC) Phase

Step 6: Find unique records in data set by using DC Algorithm

Classification Phase

Step 7: Estimate IWVDM by using weights defined in (26)

Step 8: Employ an attribute weighted *k*-NN algorithm by using IWVDM.

Figure 5. CACDIFES Algorithm

3.2 Experimental Results of CACDIFES Algorithm

In this part of the thesis, results of a series of experiments for testing the efficiency of CACDIFES algorithm are presented. In order to see if our proposed difference metric IWVDM works efficiently or not, we compare its performance in classification by using three other difference metrics, which are OM, FDM and VDM in each experiment. For performance check of the experiments, we use performance measures given in section 2.7.

All the experiments presented in this chapter are performed on Intel(R) Core(TM) i7-4720HQ CPU @ 2.60GHz processor laptop with 16GB of memory and running on Windows 10 Home, x64. Experiments are run on R package (x64) version 3.3.1 (R core team, 2016). The following R-packages are used in the experiments:

- biganalytics (Emerson and Kane, 2016)
- bigmemory (Kane et al., 2013)
- caTools (Tuszynski, 2014)
- cvTools (Alfons, 2012)
- FSelector (Romanski and Kotthoff, 2014)
- Infotheo (Patrick, 2014)
- matrixcalc (Novomestky, 2012)
- mgcv (Wood, 2011)

In experiments, we use 10 different data sets obtained from UCI Machine Learning Repository which is one of the main data source used by machine learning community for empirical analysis of machine learning algorithms (Lichman, 2013). For description and the first two phases of CACDIFES algorithm see Table 5.

Table 5. Description, IFS and DC Results of UCI data sets

No	Dataset	IFS Results				DC Results		
		# of Attr.	# of Attr. after IFS	Attr. Red. Ratio (%)	Preserved Inf. Ratio (%)	# of Obj	# of Obj After DC	Compr. Rate (%)
1	Breast Cancer	9	5	44.44	100.00	286	286	0.00
2	Credit-g	20	9	55.00	92.30	1,000	878	12.20
3	vote	16	11	31.25	94.80	435	435	0.00
4	SPECT	22	17	22.73	93.24	267	209	21.72
5	Mushroom	22	1	95.45	90.69	8,123	10	99.88
6	Segment	19	7	63.16	91.21	1,500	316	78.93
7	Hypothyroid	29	2	93.10	94.97	3,772	3,772	0.00
8	Ionosphere	34	6	82.35	94.22	351	351	0.00
9	US Census - Income	60	6	90.00	89.79	2,458	174	92.92
10	Diabetes	8	5	37.50	87.40	768	768	0.00

These data sets are used in the majority of classification and clustering experiments. However as we employ IFS to these data sets we see significant reduction on the number of attributes since many attributes do not contain statistically significant information on class variables (i.e. $I(C, A_j) \cong 0$). For example, favorite "Mushroom" data set has 22 attributes but only one of them contains statistically significant information on target variable. By preserving 90.69% of the information, total number of attributes is reduced to one. It is also interesting that as we employ DC to this reduced mushroom data set, total number of objects counts down to 10 where half of the objects fall in the first class and the other half fall in the second class. It is meaningless to run a classification algorithm using this data set since it is nothing but a duplication of 8,113 objects. Thus, mushroom data set is only used in IFS and DC phases but not in classification.

Note that there is also certain amount of reduction in the duplicated records after DC is employed in some data sets. Some studies in literature (Jiang et al., 2014; Jiang and Li, 2013; 2011; Li et al., 2013; Li and Li, 2011) do not take into account issues stated above. We only see that useless attributes (i.e. attributes that have the same number of values with number of objects in data sets) are removed in those studies. However, classification accuracy is highly affected by duplicated records in data sets since it overestimates its performance. We will use two different setups (with IFS and without IFS) for experiments with UCI data sets.

Table 6. Classification Accuracy of CACDIFES in UCI data sets

No	**Dataset**	***k***	**With IFS**				**Without IFS**			
			IWVDM	**VDM**	**FDM**	**OM**	**IWVDM**	**VDM**	**FDM**	**OM**
1	Breast Cancer	9	**75.58**	62.79	72.09	74.42	67.44	**70.93**	69.77	68.6
2	Credit-g	7	**71.48**	67.68	69.58	68.82	74	73.67	72	**74.33**
3	vote	7	**96.95**	96.95	96.95	88.55	**95.42**	93.13	**95.42**	86.29
4	SPECT	5	**88.89**	82.54	84.13	79.92	79.41	75	**86.76**	83.23
5	Segment	7	**69.47**	64.21	68.42	61.05	63.16	68.42	**69.47**	62.11
	Segment (without DC)		**83.11**	82.89	82.89	81.78	91.56	**94.22**	93.78	92.22
6	Ionosphere	5	**93.33**	90.48	88.57	81.63	**76.96**	71.3	73.48	76.09
7	Hypothyroid	7	**95.15**	94.35	94.67	94.49	**85.94**	85.11	85.73	85.1
8	US Census – Income (*)	13	**67.31**	40.38	59.61	50	75.92	**95.49**	87.55	91.11
	US Census - Income (*) (without DC)		**94.44**	93.89	93.22	94.03	74.9	89.79	88.6	**90.77**
9	Diabetes (**)	3	80	80.87	80	**81.3**	**74.35**	72.61	72.17	71.13

(*) 0.1% of data is sampled from original data which has a size 2,458,285 (sample size 2,485).
(**) Performance increase obtained with FS in VDM, FDM and OM compared to Jiang et al. (2014).

In Table 6, *k* shows the number of nearest neighbors in *k*-NN algorithm. In each experiment we use 70% of data set as training and 30% as testing. Training data sets are used for constructing classifiers and these classifiers are used in assigning labels to unlabeled testing objects. (For detailed estimated performance measures in each experiment, see Appendix C).

Conditional on the use of IFS, highest values within each row are highlighted in bold. As we can infer from Table 6, our proposed metric IWVDM with IFS works efficiently in many of the data sets (8 out of 9 data sets) compared to other difference metrics. Note that classification accuracy results of IWVDM with IFS for all data sets are higher compared to those of IWVDM without IFS. Highest accuracy increase in IWVDM with IFS compared to those without IFS is obtained with US Census data without DC, where the accuracy increase is 19.54%.

It is interesting to note that for some data sets such as Credit-g and SPECT, specificity is obtained significantly higher than sensitivity (see Table 34 and 37 in Appendix C). Main reason for this situation is that prevalence for these two data sets are not too high (33.84% for Credit-g and 14.29% for SPECT). In a more explicit form, when the events are rarely seen in data sets, classification algorithm struggles to detects the true classes of these events (specificity > sensitivity). However, when prevalence is not very low (i.e. when number of events in the data set is not very low) we obtain sensitivity greater than specificity.

We can also infer from Table 6 that some data sets (such as vote dataset) are highly “separable” which might indicate high classification accuracy no matter which method is used.

In each experiment, DC is used. However, for Segment and US Census data sets we also employ classification with and without DC. It is also interesting to note that classification accuracy differs with the use of DC (as in the case of segment data set, it is 69.47% with DC and 83.11% without DC). As mentioned before, duplicated records affect the classification, and this result in overestimation in the performance measures. Normally, without using IFS and DC, classification accuracy by employing IWVDM in segment dataset is 91.56%. However, with the use of IFS and DC it reduces to 69.47% which seems more reliable.

CHAPTER 4

PROPOSED METHOD FOR MATCHING

In this part of the thesis, our proposed matching algorithm is given. Matching algorithm proposed in this thesis consists of two main phases: scoring and sorting. Scoring phase of the matching algorithm, includes the calculation of distances between two different data sets by using a modified version of OM. Sorting phase of the algorithm, contains a heuristic procedure for determining suitable matches by using a data driven threshold parameter. These two phases will be defined in section 4.1 and 4.2, respectively. After defining these two main phases, overall JMS framework will be given as a final section of this chapter.

4.1. Matching Algorithm

Matching two different data sets is one of the major issues in decision making problems. Especially, if data sets to be matched are large-scaled and include high number of attributes in different scales the problem becomes even more complicated with possible memory-space issues. In order to address these points, we structured down our matching problem into two main phases: scoring and sorting.

4.1.1. Scoring phase

Before defining scoring phase of the matching algorithm, we will formulate the problem as follows: Let $D_1 = \{(x, C_1)\}$ and $D_2 = \{(y, C_2)\}$ be two data sets with n_1 and n_2 objects, respectively. C_1 and C_2 show class variable with number of classes l_1 and l_2, and each x and y are instances represented by vector of attribute values $\{a_1(x), a_2(x), \ldots, a_{d_1}(x)\}$ and $\{a_1(y), a_2(y), \ldots, a_{d_2}(y)\}$. Here, d_1 and d_2 show the number of attributes in D_1 and D_2, respectively. $a_j(x)$ shows the value of jth attribute A_j^1 of x and $a_t(y)$ represents the value of tth attribute A_t^2 of y. A^1 is a set of d_1 attributes $A^1 = \{A_1^1, A_2^1, \ldots, A_{d_1}^1\}$ and domain of each $A_j^1 \in A^1$ $(1 \leq j \leq d_1$) is finite. Similarly, A^2 is a set of d_2 attributes $A^2 = \{A_1^2, A_2^2, \ldots, A_{d_2}^2\}$ and domain of each $A_t^2 \in A^2$ $(1 \leq t \leq d_2$) is also finite.

Number of dimensions in each data set (d_1 and d_2) is not necessarily the same. However, distance between x and y can only be measured by the number of attributes which coincides. This process of matching suitable attributes in different data sets can be done by using a heuristic procedure. Unfortunately, there is no one-fits for all method to decide on coinciding attributes. For some attributes it might be quite obvious (such as age, gender, etc.) but for others user domain knowledge might provide significant help. Attributes which have the same data type and scale can be compared by using OM. Attributes in different scales can be converted to the same scale by using methods described in section 2.6 and then OM can be used. Nevertheless, critical issue is that OM does not take into account the attribute relevance (i.e. it equally treats all attributes), and it does not consider dependence structure among them.

In order to solve these issues, we propose a modification in OM by employing two major adjustments:

- IFS and DC phases of CACDIFES Algorithm are used to guarantee attribute independence, and
- A weighting scheme defined in (26) is employed to identify attribute importance.

We name our difference metric as Independently Weighted Overlap Metric (IWOM). IWOM for the objects x & y is given as:

$$IWOM(x,y) = \sum_{i=1}^{d} w_i \delta\big(a_i(x), a_i(y)\big) \tag{31}$$

with

$$\delta\big(a_i(x), a_i(y)\big) = \begin{cases} 0 & \text{if } a_i(x) = a_i(y) \\ 1 & \text{if } a_i(x) \neq a_i(y) \end{cases},$$

where d is the number of informative attributes that coincides, w_i is the weight defined in (26), $a_i(x)$ and $a_i(y)$ are the attribute values of x and y, respectively.

Another major issue is to decide whether to match data set 1 with data set 2 or data set 2 with data set 1 (since n_1 and n_2 are not necessarily the same). This is highly related to research question. For example, in our case we intend to match job seekers with vacancies. It is quite obvious that matching vacancies with job seekers is another valid research question and can be done, too.

Third issue is to decide whether to calculate IWOM for all objects in data set 1 with all objects in data set 2 or not. Grouping down objects in data set 1 to match certain data objects in data set 2 will surely decrease running time, but it will also prevent some objects to be matched even if they have smaller IWOM values. For example, grouping job seekers with respect to their location (province of residence) and matching them by the vacancies with coinciding location might increase the chances of matching (i.e. prioritizing with respect to location of job seekers and vacancies). However, this might also result in eliminating the matching of job seekers by the vacancies from different locations despite the possible higher rate of suitability among them.

Instead of such grouping procedure, we calculate IWOM for all observations in data set 1 with all observations in data set 2 (many with many) and propose a certain data driven threshold (τ) for deciding candidate matches.

4.1.2. Sorting phase

Sorting phase of the algorithm includes ordering estimated IWOM of objects in data sets with respect to their magnitudes. In a more explicit form, we will have an $n_1 \times n_2$ matrix of estimated IWOMs given in (32):

$$\mathrm{IWOM}(x,y) = \begin{bmatrix} d(x_1,y_1) & d(x_1,y_2) & \dots & d(x_1,y_{n_2}) \\ d(x_2,y_1) & d(x_2,y_2) & \cdots & d(x_2,y_{n_2}) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_{n_1},y_1) & d(x_{n_1},y_2) & \cdots & d(x_{n_1},y_{n_2}) \end{bmatrix}_{n_1 \times n_2}, \qquad (32)$$

where $d(x_i,y_j)$ represents the $IWOM(x_i,y_j)$ for $i = 1,\dots,n_1$ and $j = 1,\dots,n_2$. We will organize rows of $\mathrm{IWOM}(x,y)$ in ascending order and record the indices of ordered objects. Objects in data set 1 which have smaller IWOM values than τ are nominated as candidate matches (in the same order with IWOM values) for the corresponding object in data set 2.

Threshold parameter is obtained by using the prior information which includes the previous matches (such as previous job fillings) in the data set. On account of using IWOM, distances between previously matched objects in data set 1 with data set 2 are calculated and histogram of these distances is plotted. Experimental results reveal that histogram of calculated distances of the previous matches follow a bimodal distribution in which end point of the first part can be used as threshold parameter, τ (see Figure 12 in section 5.4.3).

Using a threshold parameter τ is not obligatory. If the research question is to find group of object in data set 1 which matches with the objects in data set 2, then a threshold parameter is useful. However, if the research question is only to find the best matches then one might pick the objects in data set 1 with the smallest IWOM values that correspond to objects in data set 2. Matching algorithm is defined in Figure 6.

Algorithm: Matching

Step 1: Apply scale conversion and data transformation methods in case it is necessary (Pre-processing step).

Step 2: Apply IFS and DC phases in CACDIFES Algorithm.

Step 3: Arrange attributes in each data set which coincide.

Step 4: Calculate IWOM of data sets given in (31) by using weights defined in (26).

Step 5: Choose threshold parameter τ for IWOM values.

Step 6: Sort IWOM values in descending order which fall below τ and nominate those as candidate matches (in the order of IWOM values).

Figure 6. Matching Algorithm

4.2. JMS Framework

JMS Framework presented in this thesis includes three major stages: Pre-processing, classification and matching. For details of JMS Framework see Figure 7.

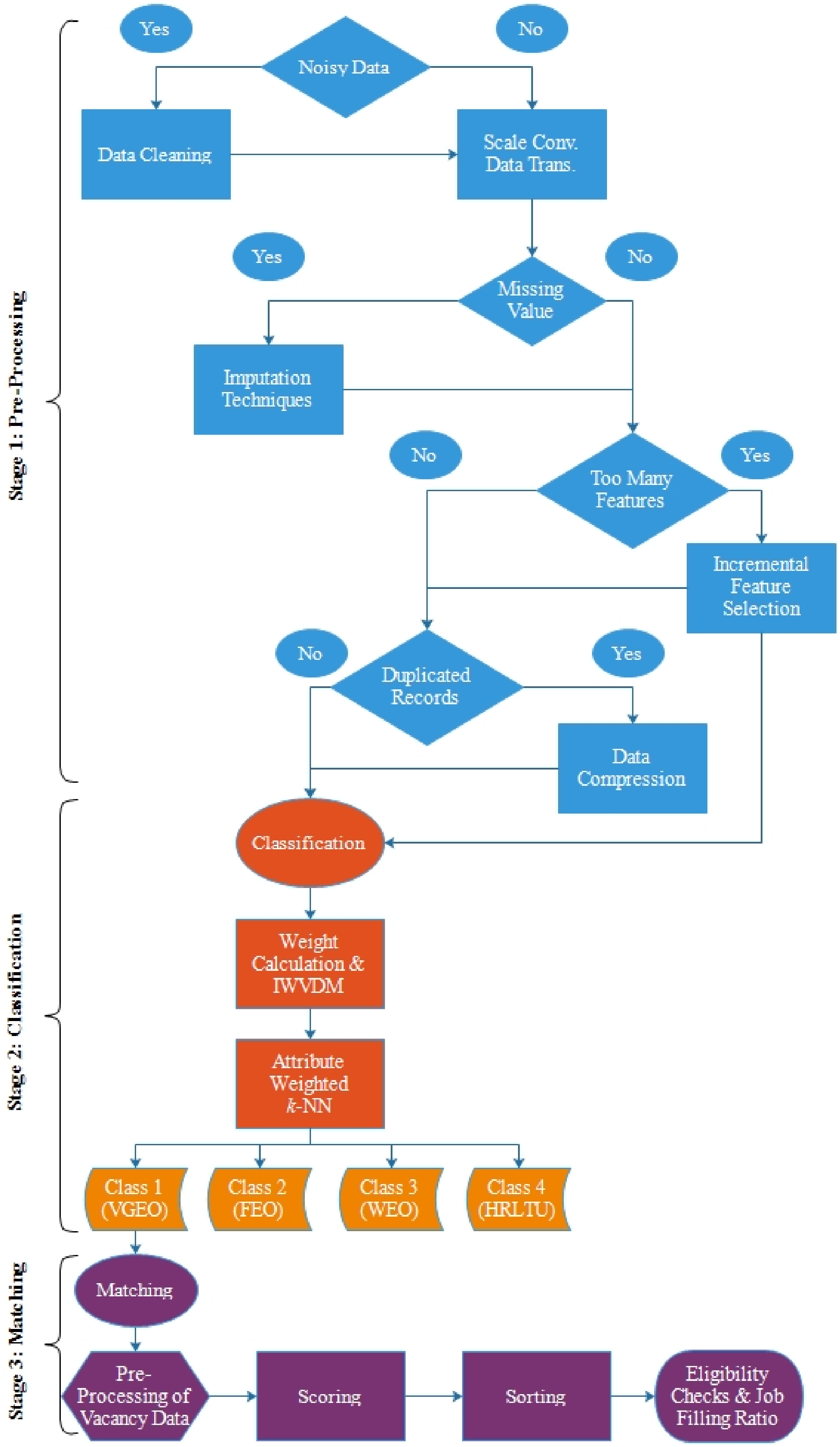

Figure 7. Details of JMS Framework

Fundamentals of pre-processing stage are data cleaning, scale conversion and data transformation, missing value imputation, feature selection and eliminating duplicated records (i.e. data compression). Classification stage includes employing an attribute weighted *k*-NN algorithm for observations (job seekers) by using IWVDM as difference metric. Matching stage covers pre-processing of second data (vacancy data) to be matched, scoring by using IWOM, sorting with respect to calculated scores and checking eligibility of job seekers for any vacancies and calculating the ratio of filled vacancies. Here eligibility refers to job seekers who have a calculated IWOM which falls below τ for any vacancy, and filling ratio represents the percentage of available vacancies filled by any eligible job seekers.

CHAPTER 5

APPLICATION OF JMS ON TEA DATA SET

As given in previous chapters, implementation of JMS includes two main phases: classification and matching. In this chapter we present the implementation of JMS to TEA data set. First and second sections of this chapter include information concerning TEA data set and a detailed preprocessing step for preparing data for the implementation of algorithms. Third and fourth sections cover implementation of CACDIFES and matching algorithms, respectively.

Before the implementation of JMS to full data, we employ CACDIFES Algorithm to TEA data set by using two different sampling designs. In the first study, we obtain a balanced sample from data set and test our algorithm's efficiency. In the second study, performance of CACDIFES algorithm is tested by using an imbalanced sample. Majority of the classification algorithms are designed to be used in binary classification problems and their efficiency might change for imbalanced data sets with multiple classes. Thus, main concern in these two different sampling designs is to identify CACDIFES Algorithm's efficiency in reflecting the true behavior of our data set. Finally, our algorithm is applied to full TEA data set.

This chapter consists of 4 main sections. Information concerning TEA data set is explained in section 5.1. Unfortunately, data set observed from TEA is neither clean nor well-structured. It requires a complicated process of data cleaning. In section 5.2, data cleaning steps are given briefly. In section 5.3, implementation of

CACDIFES Algorithm is presented. In section 5.4, implication of matching algorithm is described. In this section, job seekers who are classified as VGEO are matched by proper vacancies. Steps for the application of matching algorithm are defined in sections from 5.4.1 through 5.4.4.

5.1. Information on TEA Data Set

According to Statistical Classification of Economic Activities in the European Community, which is generally referred as NACE, jobs in labor market are classified into 17 main categories with respect to their industrial economic activity (NACE Rev. 1.1). The aim of this classification is to standardize and group industrial activity of jobs in Europe. In 2006, second revision on these codes was established (NACE Rev. 2) by Regulation (EC) No 1893/2006 of the European Parliament, and it was put in use in 2008 (Eurostat, 2008). According to this revision, jobs are grouped on account of 4 hierarchical levels. First level has 21 main sections, second level has 88 divisions, third level has 272 groups and final level includes 615 classes (for detailed information see Eurostat (2008)). Being a member of Council of Europe, Turkey amended this regulation, and TEA designed its system with respect to these activity codes (see Appendix A). According to NACE Rev 1.1, jobs in labor market are classified into 17 main and 33 sub categories in Turkey (TEA, 2014a). Since data sets observed from TEA are based on NACE Rev. 1.1, JMS is employed on account of NACE Rev. 1.1.

Due to macroeconomic policies being employed in recent years, there has been a significant change on our community's social and economic life which has a substantial effect on the labour market conditions. Moreover, there is a significant increase on the number of jobs being created by the help of TEA's recent activities. Especially after employing high number of CJOs, the institutional capacity of TEA increased significantly. Due to these important advancements in recent years, JMS is based on the data from TEA's database covering time period 2014-2015.

According to investigation of TEA's database and a detailed literature review, two main sets of data are observed from TEA. First dataset consists of information on applicants (job seekers) and the second dataset consists of information on vacancies. Both data sets cover the time period 01.01.2014 – 13.10.2015. Applicant data set includes as a total of 1,691,827 records with 63 variables whereas vacancy data set contains 627,031 records with 27 variables (Table 7). (For more information on variable types see Appendix B.)

Table 7. Information on Data set

Dataset	**Number of Numerical Variables**	**Number of Categorical Variables**	**Number of Records**		**Total Number of Records**
			01.01.2014 - 31.12.2014	**01.01.2015 - 12.10.2015**	
Applicant	25	38	1,486,927	204,900	1,691,827
Vacancy	9	18	351,863	275,168	627,031

In order to have a better understanding of TEA data sets, descriptive information concerning certain type of variables from these data sets are provided in the following sub sections (5.1.1 and 5.1.2). Descriptive information given in these sections covers TEA data sets which belong to 2015. After employing data cleaning and data manipulation techniques, our applicant data set covering time period 01.01.2015 – 12.10.2015 is reduced to 175,143 records and our vacancy data set is reduced to 275,118 records (see section 5.2 for details of data cleaning). Therefore, descriptive statistics provided in the following sub sections are obtained from cleaned version of 2015 data sets.

5.1.1. Descriptive information on applicant data set

Histogram of the ages of the job seekers is plotted in Figure 8. As it can be seen, ages of the job seekers seem to follow a right skewed distribution.

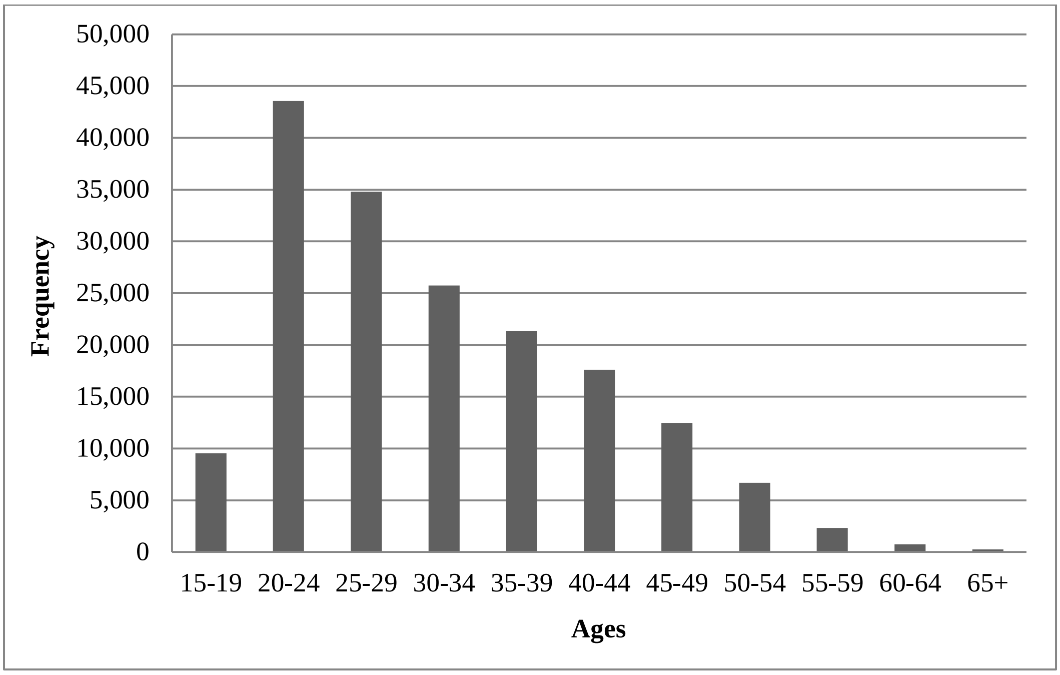

Figure 8. Histogram of Ages of Job Seekers

Figure 8 also indicates that there is high percentage of young people reported as unemployed in TEA's database. Considering the Turkey's youth unemployment rate in 2015 as 18.5%, this number is not surprising (Youth unemployment rate is defined as the number of unemployed aged among 15-24 as a percentage of youth labor force (OECD, 2016b)).

When we consider the gender distribution of job seekers we see that 61.1% of job seekers are males whereas 38.9 % of job seekers are females. Civil status of job seekers indicates that, 48.6% of the job seekers are married, 47.8% of them are single and 3.6% of them are reported as divorced. Less than 1% of job seekers have disabilities.

Percentage of job seekers who received unemployment benefits from TEA is 25.8%. Average period of receiving unemployment benefit within a year is 7 months. Average amount of unemployment benefit provided within a year is

3,750 TL.

Having a driving license is not very common among job seekers. Although it is required in certain type of vacancies, only 11.3% of job seekers own a driving license.

Table 8. Education Level of Job Seekers

Education Level	**Frequency**	**(%)**	**Cumulative (%)**
Illiterate	9,891	5.65	5.65
Literate (but no education)	6,196	3.54	9.19
Primary education	88,690	50.64	59.82
Secondary education	38,891	22.21	82.03
Two year college	13,349	7.62	89.65
Bachelor's degree	17,508	10.00	99.65
Master's degree	599	0.34	99.99
Ph. degree	19	0.01	100.00
Total	**175,143**	**100**	

Table 8 indicates that education level of the job seekers is very low. We can see that majority of the job seekers (59.82%) have primary education or less. Only 10.35% of job seekers hold a bachelor degree or more.

Table 9. Education Type of Job Seekers

School Type / Institution Type / Faculty Type	**Frequency**	**(%)**
No school	16,087	9.19
Primary school	71,881	41.04
Secondary school	16,961	9.68
High school / High school (with distance option)	19,087	10.90
Vocational education high school	16,031	9.15

Table 9. (Continued)

School Type / Institution Type / Faculty Type	Frequency	(%)
Anatolian high school	4,090	2.34
Technical college	781	0.45
Vocational technical college	12,275	7.01
Faculty with distance learning	3,540	2.02
Faculty of agriculture	561	0.32
Faculty of education	2,011	1.15
Faculty of communication	370	0.21
College of arts	302	0.17
Faculty of arts and science	3,115	1.78
Faculty of economics and administrative science	4,483	2.56
Law	220	0.13
Medicine	49	0.03
Faculty of engineering and architecture	3,299	1.88
Total	**175,143**	**100**

Table 9 shows the school type, institution type or faculty type that job seekers graduated from. It can be inferred from Table 9 that only small percentage of job seekers (6.38%) has bachelor degrees from faculty of engineering, law, medicine, arts and science, economics and administrative science in which their employment opportunities might be reasonably higher compared to other job seekers.

When we investigate the occupation of job seekers, we see that 33.42% of them are reported as blue collar worker (general type) (a person who performs work based on physical power). Second frequent occupation of job seekers is obtained as cleaner (5.51%). Majority of the job seekers (81.1%) reported that they learned their professions by hands on experience in work, while 14% of them learned it in educational institutions and only 4.9% of them learned it in vocational education training programs.

Holding a specific license or job related certifications (such as computer literacy, knowledge of different types of software, etc.) is very low among job seekers. Only 3.5% of job seekers hold these types of documents. Moreover, knowing foreign language is also very rare among job seekers (only 5.15% of job seekers know foreign language).

5.1.2. Descriptive information on vacancy data set

Gender requirement for the vacancy data set is investigated and we see that majority of the jobs (47.04%) require only male workers (Table 10).

Table 10. Gender Requirement for Vacancies

Gender Requirement	Frequency	(%)
Female	36,342	13.21
Female OR Male	109,347	39.75
Male	129,429	47.04
Total	**275,118**	**100**

Table 10 also indicates that for certain number of vacancies (39.75% of all vacancies) the required work can be done by both males and females.

Table 11. Minimum Education Level Required for Vacancies

Education Level	Frequency	(%)	Cumulative (%)
Illiterate	4,659	1.69	1.69
Literate (but no education)	31,408	11.42	13.11
Primary education	159,217	57.87	70.98
Secondary education	59,720	21.71	92.69
Two year college	9,153	3.33	96.02
Bachelor's degree	10,940	3.98	99.99
Master's degree	16	0.01	100.00
Ph. degree	5	0.00	100.00
Total	**275,118**	**100**	

Table 11 shows the minimum education level requirement for the vacancies. It can be seen that 57.87% of the vacancies require at least primary education level for applicants. It is also interesting to note that only 4% of vacancies require bachelor’s degree or more. We may infer that majority of the vacancies registered to TEA’s database do not require higher levels of education.

As we investigate the occupation requirement of vacancies, we see that 20.9% of them look for blue collar worker of general type. Second frequent occupation requirement for vacancies is accountant (3.83%). Surprisingly, 72.65% of the vacancies do not require previous job experience at all. However, 24.15% of the vacancies require previous job experience which is gained by hands on experience.

Driving license requirement is not very common among vacancies. 2.23% of vacancies require owning a driving license. Similarly, foreign language requirement is not frequently seen in vacancies. Specifically, 0.24% of the vacancies require the knowledge of a foreign language. Similar to job seekers so few of the vacancies require certain type of license or job related certificates. Only 0.11% of vacancies require owing these types of documents.

5.2. Data Cleaning Process

TEA’s database includes high volume of information. Certain type of information is registered to TEA’s database from different central databases operated by other governmental bodies through the web service technologies. This enables TEA to cross-check and verify the information entered to their database by job seekers and employee seekers. Nevertheless, the amount of such kind of information is quite limited and majority of the information entered to database is not verified.

The following information are entered by the job seeker and they are not verified: state of workforce, state of job search, social state, disability status, condemned status, priority status, type of driving license, education level (school type, department attended and graduation year), occupation (experience type and length of experience), additional information (courses and certificates), location preference for work, preference of period of work, type of the contract, economic activity of the job, opt for vocational training, level of foreign language and information on previous job experience (firm name, economic activity code, vacancy level – position, date of entrance and leaving). Although for certain type of information (such as education level, disability status, priority status, condemned status, type of driving license, previous job experience) there exists central databases operated by the responsible governmental authorities, TEA does not acquire this information from them via web services.

Besides, data for education level, such as school type and department being attended are defined as free text field in the database rather than a standard drop-down list which highly decrease the quality of the data. For example, for departments being graduated, records like "computerrr engineer, cooomputer engineer, comp. eng.," exist which all actually refers to "computer engineer". We face the similar issue with the school type being attended.

In addition to non-verified and non-standard data, we see that information entered to system is not controlled by certain analytical rules on database side with business analytics tools. In a more explicit form, some employee seekers define a minimum age requirement for a given vacancy below 15 or above 64 which is against the labour law no. 4857 where the working age is defined between 15 and 64. By using certain control operations this type of information can be standardized.

Due to above reasoning and weaker conditions on cross-checks for data entrance to TEA's Database, a data cleaning process is required to prepare the data for the analysis. Steps taken for data cleaning is as follows:

- Data types including dates are standardized ("dd.mm.yyyy").
- All information defined in free text fields such as school type, department and name of the institute/university/college/high-school, etc. are grouped,

and records with similar tags are standardized. Besides, departments in vocational high schools, junior technical and non-technical colleges, industrial schools, universities and institutes are standardized with respect to their curriculum and training program.

- New variables are created from the education field, such as education type (distance learning/formal education), education language, scholarship being received and education option (day/evening).
- Age requirement for the vacancy data is modified, and minimum age requirement defined below 15 is converted to 15, and those defined above 64 is also limited to 64.
- Applicants whose job search profiles are labeled as inactive (or not seeking jobs) in TEA's database are removed.

Non-standard, non-verified and non-controlled information registered to TEA's database are considered to have serious adverse effects on the quality of the data and services provided by the TEA. Measures should be taken to overcome these issues will be presented as policy recommendations in discussion and conclusion.

5.3. Phase 1: Application of CACDIFES Algorithm

As a result of data cleaning process given in previous chapter, we end up with an applicant data set covering time period 01.01.2015 – 12.10.2015 with 175,143 records. There exist natural classes in our data set for identifying the employment opportunities of the job seekers such as previous employment history, number of interviews with CJOs and employee seekers, etc. Hence we generated the target variable from 5 different attributes:

- Number of job application
- Number of job interview
- Job interview with CJOs
- Number of job placement
- Last job placement by TEA

Job applicants who have been placed on at least one job (who has one or more job experience) are classified as applicants who have Very Good Employment Opportunities (VGEO). Applicants who have not been placed on a job but interviewed at least once by employee seekers are classified as applicants with fair employment opportunities (FEO). Those neither employed nor applied to any job but interviewed with CJO's are classified as applicants with weak employment opportunities (WEO). Finally, applicants who applied to at least one job but neither employed nor interviewed by employee seekers are classified as applicants who are under the high risk of long term unemployment (HRLTU) (For class sizes and labels see Table 12). Bear in mind that applicants who are labeled as HRLTU have not been employed for 12 months or more (OECD, 2016c).

Table 12. Classification of Job Applicants (01.01.2015 – 12.10.2015)

Class Names	**Number of Records**	**(%)**	**Class Labels**
Fair Employment Opportunities	19,449	11.10	FEO
High Risk of Long Term Unemployment	9,500	5.42	HRLTU
Very Good Employment Opportunities	27,522	15.71	VGEO
Weak Employment Opportunities	55,044	31.43	WEO
To be Classified	63,628	36.33	
Total	175,143		

There is also another group in the data set: they have not been placed on a job, they did not even apply to any job, neither were they interviewed by employee seekers nor by CJOs. This clearly shows that they are registered to TEA's system but they have not been served yet. Hence, we need to figure out which class (or classes) this group of job seekers should be placed in by using their job relevant specifications (attribute values).

Before employing CACDIFES Algorithm, irrelevant attributes (such as applicant ID, training ID, attributes which include descriptive information, etc.) are removed, and date type attributes are transformed into numerical attributes (such as using age instead of birthdate, etc.). Attributes which include duration (such as experience in months and experience in years) are converted to the same scale and combined. After all these operations, our data set is reduced to 29 categorical attributes.

In the first step, joint mutual information of each attribute and class variable, $I(C, A_j)$, is estimated from (19) for $j = 1, \dots, 29$. In the second step, hypothesis test given in (22) is employed for each attribute at $\alpha = 0.01$ significance level, and all attributes are found to contain statistically significant information on target variable C (i.e., all attributes are found to be informative). In the third step, graph of $\hat{I}(C; A_1, \dots, A_j)$ v.s. $A_1, \dots, A_j$ and $\hat{I}(C; A_j | A_1, \dots, A_{j-1})$ v.s. $A_1, \dots, A_j$ are plotted to discover the set satisfying

$$\mathrm{argmax}_{A_j \in A}\, I(C; A_1, \dots, A_j) = \mathrm{argmin}_{A_j \in A}\, I(C; A_j | A_1, \dots, A_{j-1}), \quad (33)$$

for $j = 1, \dots, 29$. Since all attributes are sorted depending on their information level on C in descending order, here A_1 represents the attribute with the highest joint mutual information with class variable C.

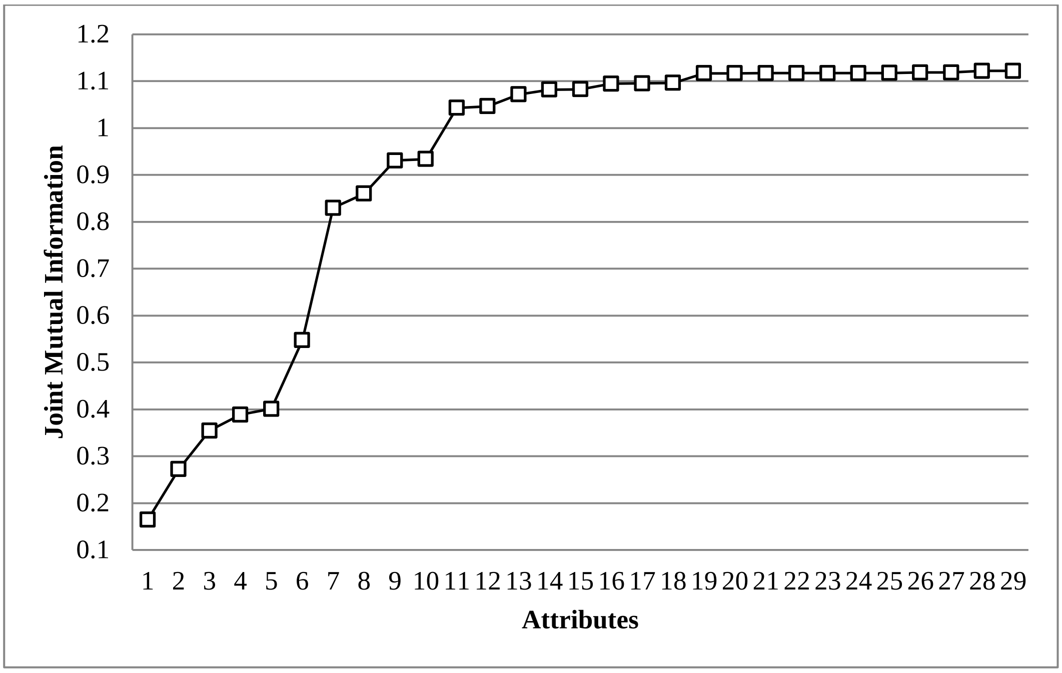

Figure 9. Joint Mutual Information of Attributes

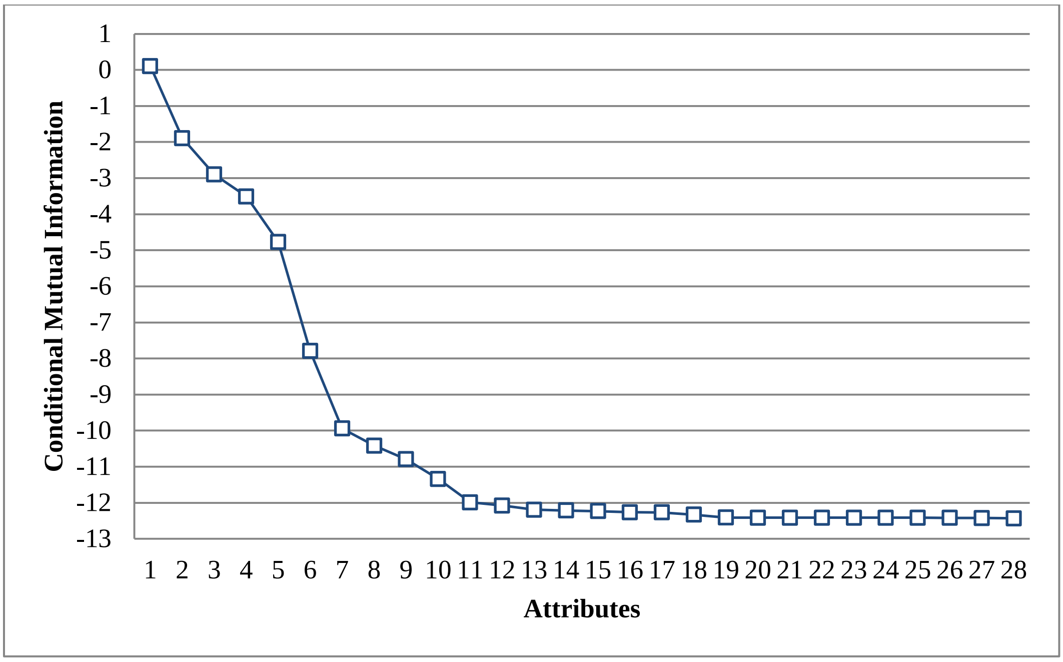

Figure 10. Conditional Mutual Information of Attributes

As we can infer from Figures 9 and 10; first 19 informative attributes seem to have sufficient information on target variable C since they contain 99.5% of the information. In a more explicit form, total information of all attributes and target variable C is estimated as $\hat{I}(C; A_1, \dots, A_{29}) = 1.121708$ and joint mutual information of informative attributes and C is given as $\hat{I}(C; A_1, \dots, A_{19}) = 1.116757$. Thus the information ratio is $\frac{\hat{I}(C;A_1,\dots,A_{19})}{\hat{I}(C;A_1,\dots,A_{29})} = 0.9955$.

In a coarser manner, one may choose to decide on the sufficiency of information ratio as 95%. Then, algorithm will stop after subsetting the first 12 informative attributes due to assigning $\beta = 0.95$ (in step four) as a threshold parameter for the information gain. (Information ratio of first 13 informative attributes and target variable is $\frac{\hat{I}(C;A_1,\dots,A_{13})}{\hat{I}(C;A_1,\dots,A_{29})} = 0.955 \geq 0.95$). Hence, we can neglect the information from the remaining 17 attributes.

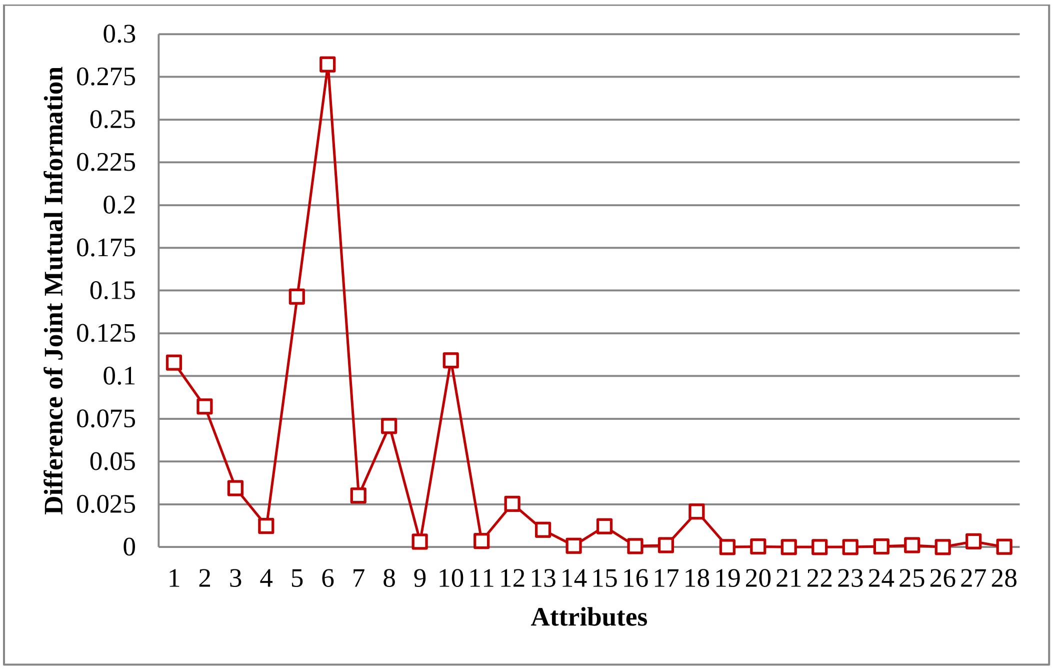

Figure 11. Difference of Joint Mutual Information of Attributes

Figure 11 shows the change in the difference of joint mutual information with the additional attribute. It can be inferred that we do not observe significant changes after the 19th attribute (i.e., $\hat{I}(C;,A_1,\dots,A_{20}) - \hat{I}(C;,A_1,\dots,A_{19}) \approx 0$). Figure 11 also provides visual inspection on the dependence structure of informative attributes with respect to their contribution to total information. If all attributes were strictly independent, then we would possibly expect Figure 11 to be a non-increasing function (in fact it should be decreasing). Since we add attributes to joint mutual information starting from the highest informative one (such as A_1 in SA), the difference between the joint mutual information of informative attributes, $\hat{I}\big(C;A_1,\dots,A_j,A_{j+1}\big) - \hat{I}\big(C;A_1,\dots,A_{j-1},A_j\big)$, should always decrease. Yet, we see significant jumps at certain attributes. Bearing in mind this graphical inspection, we estimate the Symmetric Uncertainty matrix (**U**) of informative attributes, and check if any of the estimated value is greater than 0.5.

Table 13. Symmetric Uncertainty Matrix of Informative Attributes

Attribute Numbers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	1.00	0.06	0.04	0.07	0.06	0.29	0.13	0.18	0.02	0.15	0.06	0.12
2	0.06	1.00	0.02	0.02	0.02	0.03	0.01	0.02	0.00	0.02	0.01	0.00
3	0.04	0.02	1.00	0.41	0.44	0.01	0.01	0.01	0.04	0.00	0.04	0.00
4	0.07	0.02	0.41	1.00	0.73	0.03	0.02	0.01	0.04	0.01	0.04	0.00
5	0.06	0.02	0.44	0.73	1.00	0.02	0.01	0.01	0.05	0.00	0.04	0.00
6	0.29	0.03	0.01	0.03	0.02	1.00	0.05	**0.61**	0.01	**0.57**	0.06	0.10
7	0.13	0.01	0.01	0.02	0.01	0.05	1.00	0.02	0.01	0.01	0.01	0.00
8	0.18	0.02	0.01	0.01	0.01	**0.61**	0.02	1.00	0.01	**0.75**	0.06	0.12
9	0.02	0.00	0.04	0.04	0.05	0.01	0.01	0.01	1.00	0.01	0.00	0.01
10	0.15	0.02	0.00	0.01	0.00	**0.57**	0.01	**0.75**	0.01	1.00	0.05	0.15
11	0.06	0.01	0.04	0.04	0.04	0.06	0.01	0.06	0.00	0.05	1.00	0.02
12	0.12	0.00	0.00	0.00	0.00	0.10	0.00	0.12	0.01	0.15	0.02	1.00

As listed in Table 13, attributes 6 and 10 satisfy our condition for the removal, since $U(A_6, A_8) = 0.61$, $U(A_6, A_{10}) = 0.57$ and $U(A_8, A_{10}) = 0.75$. We may infer that by including attribute 8 we include the information from both attributes 6 and 10. A_6 has 1,041 levels representing the departments being graduated from and removing that attribute will surely increase the computational efficiency in further steps. A_8 shows the school type being graduated (with 18 levels), and A_{10} shows the education level (with 8 levels). By using the same reasoning, we might remove A_8 but school type being graduated from also covers the information on education level. As a result we choose to remove A_{10}. By further discarding the attributes 6 and 10, we obtain the joint mutual information of informative attributes and class variable as 1.017 and information ratio as 91%. Finally, we obtain s' number of informative attributes, $SA' = \{A_1, \dots, A_{s'}\}$ with $s' \leq s^*$. This group contains sufficient information on target variable C (In our case, $s^* = 12$ and $s' = 10$ and independent informative attribute numbers are: 1, 2, 3, 4, 5, 7, 8, 9, 11, 12). These independent informative attributes are occupation, number of different occupations, number of unemployment benefits being received, total amount of unemployment benefits received, period of unemployment benefits received (in months), province of residence, school type being graduated from, state of workforce, age and, experience type of the occupation.

We also use Cramér's V to measure the dependence structure among attributes. Cramér's V (or Cramér's phi) is used as a measure for association between two categorical variables based on Pearson's chi-square test statistics (Cramér, 1946). The outcome of the test statistics is between 0 and 1, where 0 shows no relationship and 1 shows perfect relationship. For two categorical variables A_1 and A_2, Cramér's V is given by;

$$V = \sqrt{\frac{\chi^2/n}{\min(t_1-1, t_2-1)}}, \tag{34}$$

where, n is the sample size, the t_1 and t_2 show the number of categorical values that A_1 and A_2 can take, and χ^2 shows the chi-square test statistics with degrees of freedom $(t_1 - 1) \times (t_2 - 1)$.

Table 14. Cramér's *V* estimates of Informative Attributes

Attribute Numbers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	1.00	0.27	0.31	0.26	0.25	0.39	0.21	0.59	0.22	0.42	0.20	0.69
2	0.27	1.00	0.10	0.10	0.10	0.16	0.11	0.12	0.05	0.11	0.08	0.03
3	0.31	0.10	1.00	0.34	0.37	0.14	0.10	0.05	0.15	0.04	0.13	0.04
4	0.26	0.10	0.34	1.00	0.59	0.11	0.06	0.05	0.19	0.06	0.09	0.05
5	0.25	0.10	0.37	0.59	1.00	0.11	0.07	0.04	0.20	0.04	0.09	0.04
6	0.39	0.16	0.14	0.11	0.11	1.00	0.10	**0.83**	0.15	**0.72**	0.16	0.46
7	0.21	0.11	0.10	0.06	0.07	0.10	1.00	0.08	0.13	0.12	0.05	0.10
8	0.59	0.12	0.05	0.05	0.04	**0.83**	0.08	1.00	0.10	**0.73**	0.14	0.42
9	0.22	0.05	0.15	0.19	0.20	0.15	0.13	0.10	1.00	0.10	0.05	0.06
10	0.42	0.11	0.04	0.06	0.04	**0.72**	0.12	**0.73**	0.10	1.00	0.15	0.39
11	0.20	0.08	0.13	0.09	0.09	0.16	0.05	0.14	0.05	0.15	1.00	0.15
12	0.69	0.03	0.04	0.05	0.04	0.46	0.10	0.42	0.06	0.39	0.15	1.00

Cramér's *V* estimates for our 12 informative attributes are given in Table 14. Bryman and Cramer (1997) interpret the estimated Cramér's *V* which is given in Table 15.

Table 15. Interpretation of Cramér's *V* estimates

Cramér's *V*	**Interpretation**
< 0.19	Very low association
0.20 - 0.39	Low association
0.40 - 0.69	Modes association
0.70 - 0.90	High association
0.91 - 1	Very high association

According to Table 14 and 15, we see that there is a high association between attributes 6, 8 and 10. Thus we need to eliminate some of these attributes by maintaining the joint mutual information level. Apparently, our approach for discarding attributes by using symmetric uncertainty seems logical and it is verified by Cramér’s *V* estimates.

As given in section 3.1.2.1, joint mutual information always increases with the addition of new attributes. If all our attributes are independent from each other and we start adding attributes to joint mutual information depending on their information level on C, then we should expect $\hat{I}(C; A_1, \dots, A_j, A_{j+1}) - \hat{I}(C; A_1, \dots, A_{j-1}, A_j)$ to be decreasing function of A_j for $j = 1, \dots, s$. However, Figure 11 shows the converse. Significant jumps in attributes 6, 8 and 10 in Figure 11, Symmetric Uncertainty estimates given in Table 13 and Cramér’s *V* estimates given in Table 14 imply dependence structure among attributes. Simply, Figure 11 shows the existence of possible dependence structure among attributes, and Table 13 and 14 show which attributes might be the cause of this.

In step 6, DC algorithm defined in section 3.1.2.2 is employed. Since the number of informative attributes is reduced to 10, DC is expected to work more efficiently. After employing DC 111,515 records are reduced to 87,773 (compression rate is 22%).

In step 7, weights of the independent informative attributes are estimated by using symmetric uncertainty as defined in (26). Estimated weights are given in Table 16. Bear in mind that weights are estimated by using full data.

Table 16. Attribute Weights

Attribute Numbers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Attribute Weights	0.290	0.132	0.100	0.021	0.015	0.172	0.102	0.064	0.096	0.007

In order to have an intuitive idea whether our IWVDM works efficiently or not, we employed two different types of simple random sampling procedure, and we sampled 15% of the data. Considering we have an unbalanced data set, i.e., the number of objects is highly different in each class (see Table 17), we use both balanced and imbalanced samples for investigating the efficiency of our algorithm.

Table 17. Class Distribution of Compressed Data

Class Names	**Number of Records**	**(%)**	**Class Label**
Fair Employment Opportunities	15,453	17.61	**FEO**
High Risk of Long Term Unemployment	8,109	9.24	**HRLTU**
Very Good Employment Opportunities	16,413	18.71	**VGEO**
Weak Employment Opportunities	47,758	54.44	**WEO**
Total	**87,733**	**100**	

5.3.1. Application with balanced sample

In each application of CACDIFES algorithm with balanced, imbalanced and full data, we split 70% of data set into training and 30% of data set into testing. Training data sets are used for constructing classifiers and these classifiers are used in assigning labels to unlabeled testing objects.

Table 18. Descriptive Information on Balanced Sample

Sample / Training / Testing	**Class Labels**				**Total**
	FEO	**HRLTU**	**VGEO**	**WEO**	
Size of Sample	3,290	3,290	3,290	3,290	13,160
(%)	25	25	25	25	100
Size of Training Set	2,299	2,333	2,306	2,274	9,212
(%)	24.96	25.33	25.03	24.69	100
Size of Testing Set	991	957	984	1,016	3,948
(%)	25.10	24.24	24.92	25.73	100

Table 18 represents the descriptive information of balanced sample. Next, attribute weighted *k*-NN algorithm is employed, and optimal number of *k* which corresponds to highest overall accuracy of the algorithm is found as 11. Since we estimated weights by using full data which is unbalanced, our weights will reflect the true behavior of unbalanced data. Thus our difference metric should work more efficiently by using imbalanced data. Not surprisingly, Table 19 implies that after using IFS and DC, classification accuracy by using VDM is slightly superior to other 3 metrics including IWVDM.

Table 19. Performance Estimates (in %'s) of Classification on Balanced Sample

For *k*=11	**IWVDM**				**Overall Accuracy**
	FEO	**HRLTU**	**VGEO**	**WEO**	
Sensitivity	49.34	48.90	56.61	47.54	50.58
Specificity	81.40	81.78	80.94	90.08	
Pos Pred Value	47.06	46.20	49.64	62.40	
Neg Pred Value	82.74	83.34	84.89	83.21	
Prevalence	25.10	24.24	24.92	25.73	
Detection Rate	12.39	11.85	14.11	12.23	
Detection Prevalence	26.32	25.66	28.42	19.60	
Balanced Accuracy	65.37	65.34	68.77	68.81	
For *k*=11	**VDM**				**Overall Accuracy**
	FEO	**HRLTU**	**VGEO**	**WEO**	
Sensitivity	47.12	45.25	59.45	53.05	**51.27**
Specificity	80.18	82.45	81.65	90.83	
Pos Pred Value	44.35	45.20	51.82	66.71	
Neg Pred Value	81.90	82.47	85.85	84.81	
Prevalence	25.10	24.24	24.92	25.73	
Detection Rate	11.83	10.97	14.82	13.65	
Detection Prevalence	26.67	24.27	28.60	20.47	
Balanced Accuracy	63.65	63.85	70.55	71.94	
For *k*=11	**FDM**				**Overall Accuracy**
	FEO	**HRLTU**	**VGEO**	**WEO**	
Sensitivity	43.90	42.63	57.93	51.57	49.06
Specificity	81.03	80.71	81.21	89.22	
Pos Pred Value	43.67	41.42	50.58	62.38	
Neg Pred Value	81.17	81.47	85.32	84.17	
Prevalence	25.10	24.24	24.92	25.73	
Detection Rate	11.02	10.33	14.44	13.27	
Detection Prevalence	25.23	24.95	28.55	21.28	
Balanced Accuracy	62.46	61.67	69.57	70.40	

Table 19. (Continued)

For *k*=11	OM				Overall Accuracy
	FEO	HRLTU	VGEO	WEO	
Sensitivity	41.57	40.65	62.09	51.08	48.91
Specificity	81.64	82.72	78.41	89.19	
Pos Pred Value	43.14	42.94	48.84	62.08	
Neg Pred Value	80.65	81.33	86.17	84.03	
Prevalence	25.10	24.24	24.92	25.73	
Detection Rate	10.44	9.85	15.48	13.15	
Detection Prevalence	24.19	22.95	31.69	21.18	
Balanced Accuracy	61.61	61.68	70.25	70.14	

Although overall accuracy results are not very high with balanced sample, we see from Table 19 that balanced accuracy for job seekers who are classified as VGEO is around 70%. Since our JMS design uses the outcome of the classification as an input to the matching, high percentage of correctly classified job seekers as VGEO is of our interest.

In section 2.7, prevalence is defined as the frequency of events over all cases. We would expect prevalence to be the same as the percentage of testing set with respect to classes and it is clear from Tables 18 and 19, these two aforementioned are the same. We also see that detection rates are very low. We know that detection rate is the ratio of true positive cases that are correctly predicted over all cases and summation of detection rates for each class gives the overall accuracy which is also low.

5.3.2. Application with imbalanced sample

Since our IWVDM is generated by using weights which are estimated from the imbalanced data, we employ a simple random sampling procedure and obtain an imbalanced sample to reflect the real behavior of the data set (Table 20).

Table 20. Descriptive Information on Imbalanced Sample

Sample / Training / Testing	**Class Labels**				**Total**
	FEO	**HRLTU**	**VGEO**	**WEO**	
Size of Sample	2,402	1,229	2,462	7,067	13,160
(%)	18.25	9.34	18.71	53.70	100
Size of Training Set	1,657	855	1,730	4,970	9,212
(%)	17.99	9.28	18.78	53.95	100
Size of Testing Set	745	374	732	2,097	3,948
(%)	18.87	9.47	18.54	53.12	100

Population	15,443	8,109	16,413	47,758	87,723
(%)	17.60	9.24	18.71	54.44	100

By following the same steps as in section 5.3.1 and taking $k = 11$, we employ attribute weighted k-NN algorithm. Performance results of the CACDIFES algorithm with imbalanced sample is given in Table 21.

Table 21. Performance Estimates of Classification on Imbalanced Sample

For *k*=11	**IWVDM**				**Overall Accuracy**
	FEO	**HRLTU**	**VGEO**	**WEO**	
Sensitivity	57.05	13.90	37.43	77.73	**60.31**
Specificity	84.67	98.27	90.05	62.51	
Pos Pred Value	46.40	45.61	46.13	70.14	
Neg Pred Value	89.45	91.60	86.34	71.24	
Prevalence	18.87	9.47	18.54	53.12	
Detection Rate	10.76	1.32	6.94	41.29	
Detection Prevalence	23.20	2.89	15.05	58.87	
Balanced Accuracy	70.86	56.08	63.74	70.12	

Table 21. (Continued)

For *k*=11	**VDM**				**Overall Accuracy**
	FEO	**HRLTU**	**VGEO**	**WEO**	
Sensitivity	49.93	17.11	44.95	75.87	59.68
Specificity	85.48	96.84	88.03	66.02	
Pos Pred Value	44.44	36.16	46.08	71.67	
Neg Pred Value	88.01	91.78	87.54	70.72	
Prevalence	18.87	9.47	18.54	53.12	
Detection Rate	9.42	1.62	8.33	40.30	
Detection Prevalence	21.20	4.48	18.09	56.23	
Balanced Accuracy	67.71	56.98	66.49	70.94	
For *k*=11	**FDM**				**Overall Accuracy**
	FEO	**HRLTU**	**VGEO**	**WEO**	
Sensitivity	46.98	1.87	33.33	82.31	58.94
Specificity	86.76	99.80	92.04	49.54	
Pos Pred Value	45.22	50.00	48.80	64.89	
Neg Pred Value	87.56	90.67	85.85	71.20	
Prevalence	18.87	9.47	18.54	53.12	
Detection Rate	8.87	0.18	6.18	43.72	
Detection Prevalence	19.61	0.35	12.66	67.38	
Balanced Accuracy	66.87	50.84	62.69	65.92	
For *k*=11	**OM**				**Overall Accuracy**
	FEO	**HRLTU**	**VGEO**	**WEO**	
Sensitivity	40.81	4.55	42.62	80.97	59.04
Specificity	88.85	99.47	89.02	52.03	
Pos Pred Value	45.99	47.22	46.92	65.66	
Neg Pred Value	86.58	90.87	87.21	70.70	
Prevalence	18.87	9.47	18.54	53.12	
Detection Rate	7.70	0.43	7.90	43.01	
Detection Prevalence	16.74	0.91	16.84	65.50	
Balanced Accuracy	64.83	52.01	65.82	66.50	

As it is shown in Table 21, our proposed algorithm by using IWVDM has slightly better classification accuracy than the previously introduced difference metrics with imbalanced sample. This suggests that weight assignment procedure in IWVDM reflects the true behavior of the data set, and it is more proper for the imbalanced data sets with more than one class which can be encountered in many real-life data sets. This is one of the major issues in classification algorithms; classification accuracy is highly affected by the classes with predominant nature since they prevent true classification of objects that belong to classes with fewer objects. However, our algorithm prevents this situation, and it can efficiently classify objects with rare existence in domain.

Similar to discussion in previous sub section, prevalences for each class are found to be the same as their sampling percentages in testing set (Table 21). Note that attribute independence assumption for FDM, VDM and OM restricts their usage in many real-life data sets. However, using IFS enables these three metrics to be used in data sets which have dependence structure among attributes. By using IFS and DC we seem to increase the classification accuracy of these three metrics in our data set.

5.3.3. Application with full data

In the previous two sections CACDIFES Algorithm is employed to balanced and imbalanced samples. Now we will use our proposed algorithm for the complete applicant data set.

Table 22. Full Data Classification Results

Before Classification			**Classified Objects**		**After Classification**	
Class Names	**Number of Records**	**(%)**	**Number of Records**	**(%)**	**Number of Records**	**(%)**
FEO	19,449	11.10	21,134	33.21	40,583	23.17
HRLTU	9,500	5.42	1,098	1.73	10,598	6.05
VGEO	27,522	15.71	6,745	10.60	34,267	19.57
WEO	55,044	31.43	34,651	54.46	89,695	51.21
To be Classified	63,628	36.33				
Total	**175,143**		**63,628**		**175,143**	

By using 111,515 objects as training set, we classified 63,628 objects with CACDIFES Algorithm. According to resulting classification outcome given in Table 22, majority of the objects (54.46% of them) are classified as WEO. After we sum the classified objects with the training set, we may see that only 19.57% of job seekers have high probability of transition from unemployment to labour market (For a detailed discussion see chapter 6).

5.4. Phase 2: Application of Matching Algorithm

In this section, we employ our matching algorithm to data set observed from TEA. As it is defined in section 1.2, our aim is to match job seekers (data set 1) with vacancies (data set 2). In the first phase of JMS, job seekers are classified depending on their employment opportunities by using CACDIFES Algorithm. In matching phase, job seekers who are classified as VGEO are matched by proper vacancies.

Matching algorithm is employed in six steps (see Figure 6), which are preprocessing, application of IFS and DC, attribute arrangement, calculation of IWOM, calculation of τ and sorting. These steps are defined in the following sections.

5.4.1. Preprocessing: IFS and DC steps

In the first step, we employ data cleaning process defined in section 5.2 and our vacancy data set covering the time period 01.01.2015 – 12.10.2015 is reduced to 275,118 records.

In the second step, IFS and DC phases of CACDIFES Algorithm is employed to vacancy data set, and independent and informative attributes are gathered. In order to employ IFS and DC, class variable for vacancy data set is generated depending on the vacancy status. If a given vacancy was previously filled then it is classified as “filled”, and if it was not, it is identified as “unfilled”. According to this classification rule, out of 275,118 records, 6,330 vacancies (2.3% of vacancies in 2015) are previously filled and 268,788 vacancies (97.7%) are unfilled.

Main concern in class label assignment is to investigate which factors (attributes) possess significant information on a given vacancy’s being filled. In other words, the idea is to explore decision making process of employee seekers by identifying the importance of these attributes from their perspective. Next, steps from 1 to 4 in CACDIFES Algorithm (see Figure 5) are employed and number of informative attributes is reduced to 9. In order to check whether there exists dependence structure among informative attributes, we examine Symmetric Uncertainty matrix (**U**). According to **U** matrix given in Table 23, none of the symmetric uncertainty estimates is greater than 0.5, which indicates attribute independence.

Table 23. Symmetric Uncertainty Matrix of Informative Attributes for Vacancy Data

Attribute Numbers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	1.00	0.08	0.06	0.13	0.04	0.15	0.10	0.07	0.03
2	0.08	1.00	0.32	0.43	0.33	0.03	0.04	0.30	0.00
3	0.06	0.32	1.00	0.19	0.26	0.03	0.04	0.25	0.00
4	0.13	0.43	0.19	1.00	0.11	0.02	0.01	0.23	0.00
5	0.04	0.33	0.26	0.11	1.00	0.02	0.04	0.01	0.00
6	0.15	0.03	0.03	0.02	0.02	1.00	0.02	0.04	0.04
7	0.10	0.04	0.04	0.01	0.04	0.02	1.00	0.01	0.04
8	0.07	0.30	0.25	0.23	0.01	0.04	0.01	1.00	0.00
9	0.03	0.00	0.00	0.00	0.00	0.04	0.04	0.00	1.00

Obtained informative attributes in vacancy data are occupation, minimum and maximum age requirements for males and females, gender, minimum education level, location (province) of vacancy and experience type. Next, DC is employed and total number of records is reduced from 275,118 to 143,924 (compression rate is 47.6%).

5.4.2. Attribute arrangement and IWOM calculation steps

In the third step, attributes in job seekers data which correspond to informative and independent attributes in vacancy data are gathered. As a result a new applicant data set is generated from job seekers using following attributes of applicants: occupation, age, gender, education level, province of residence and experience type. Although there are more attributes in both data sets which coincide, only these set of attributes contain statistically significant information.

In the fourth step, IWOM values of job seekers and vacancies are calculated. In order to find IWOM, weight assignment procedure defined in (26) is used for informative attributes in vacancy data.

Table 24. Attribute Weights for Vacancy Data

Attribute Numbers	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Attribute Weights	0.282	0.133	0.140	0.018	0.036	0.066	0.263	0.017	0.045

We may infer from Table 24 that most informative attributes are the first and seventh attributes, which are occupation and province of vacancy. Their total weight is 0.545. Intuitively, we may think that in order to have an efficient match between job seekers and vacancies these two attributes should be the same. Next, IWOM matrix defined in (32) is calculated. As it is discussed in section 4.1.1, for the same nominal attributes which are in the same scale in vacancy and applicant data sets IWOM works successfully. However, we have interval type of attributes for age in vacancy data set and numerical type of an attribute for age in applicant data set. In fact, there are 4 attributes for age requirement in vacancy data set which are minimum and maximum age limits for males and females, respectively.

We make use of the following heuristic for calculating the IWOM for age:

- If the applicant's age is smaller than the minimum age limit or greater than the maximum age limit, then the distance is taken 1.

- If the applicant's age is greater than the minimum age limit and smaller than the maximum age limit, then distance is taken as $1/|a_i(x) - a_i(y)|$,

where, $a_i(x)$ represents the applicant's age and $a_i(y)$ shows minimum or maximum age limit requirements. In a similar manner minimum distance in education level is taken as 0 if education level of the applicant is the same as the vacancy requirement, and it is taken as $1/|a_i(x) - a_i(y)|$ if education level of applicant is higher than the vacancy ($a_i(x)$ shows the applicant's education level and $a_i(y)$ represents the minimum education requirement for vacancy). Therefore, each attribute has distance values between $[0,1]$. Multiplying with its corresponding weight, each IWOM value lies in the [0,1] interval.

5.4.3. Calculation of τ and sorting steps

After calculating IWOM for applicants and vacancies, next step is to find applicants who are suitable for vacancies by using IWOM values. Bear in mind that, smallest value of IWOM for a given vacancy does not necessarily mean a suitable match between that vacancy and applicant. Hence, we define an upper bound so that IWOM values which falls below that bound are identified as candidate matches.

In order to estimate an upper bound τ for a given vacancy, a data driven approach is used. Dataset observed from TEA includes previous job fillings. By using data set covering time period 01.01.2014-12.10.2015, we obtain information on 11,667 vacancies being filled by applicants. Table 25 shows frequency of attributes that coincide in vacancy and applicant data sets. As it can be seen from Table 25 attributes listed in bold letters do not contain sufficient observations in vacancy or applicant data set (or both in some attributes). Not surprisingly, after we employ IFS and DC to vacancy data set, these attributes (listed in bold in Table 25) are eliminated.

Table 25. Attributes Coincide in Applicant and Vacancy Data sets

Applicant Data Set	**Freq.**	**Vacancy Data Set**	**Freq.**
Age	11,667	Minimum Age Requirement for Females	11,667
		Maximum Age Limit for Females	11,667
		Minimum Age Requirement for Males	11,667
		Maximum Age Limit for Males	11,667
Gender	11,667	Gender	11,667
Province of Residence	11,667	Province of Vacancy	11,667
Driving License	**1,041**	**Driving License**	**59**
Education Level	11,667	Minimum Education Level	11,667
		Maximum Education Level	11,667
Preference of Economic Activity	**188**	Economic Activity of Vacancy	11,667
Previous Job's Economic Activity	**254**		
Applicant Data Set	**Freq.**	**Vacancy Data Set**	**Freq.**
Occupation	11,667	Occupation	11,667
Duration of Work Experience (in months)	**259**	**Minimum Duration of Work Experience (in months)**	**5**
Experience Type	11,667	Experience Type	3,366
Additional Info-1	**207**	**Additional Info-1**	**2**
Additional Info-2	**100**	**Additional Info-2**	**1**
Additional Info-3	**50**	**Additional Info-3**	**0**
Foreign language	**276**	**Foreign Language Requirement**	**5**
Preference of Period of Work	11,667	Preference of Period of Work	11,667

Final set of attributes which are used in the calculation of IWOM are as follows:

- Minimum Age Requirement for Females
- Maximum Age Limit for Females
- Minimum Age Requirement for Males
- Maximum Age Limit for Males
- Gender
- Province of Vacancy

- Minimum Education Level
- Experience Type
- Occupation

By using IWOM, distances between previously filled applicants and vacancies are calculated. Histogram of calculated IWOM is given in Figure 12. As we may infer from Figure 12, distribution of distances seems to follow a bimodal distribution. We investigate the histogram and experimentally obtain a global upper bound τ for detecting matches and mismatches with respect to occupation. Specifically, IWOM values which are smaller than $\tau = 0.27$ are defined to be matches, and those that are higher than τ are denoted as mismatches between vacancies and applicants with respect to occupation. We can estimate the mismatch ratio with respect to occupation as 27.8%.

In order to explain mismatch on account of occupation, we illustrate a simple case in Table 26. Consider two different applicants whose occupation is electrician. First applicant filled a waiter position while second applicant filled an electrician position.

Table 26. Example for Match and Mismatch

No	Applicant's Occupation	Occupation in Vacancy being filled	Match/Mismatch
1	Electrician	Waiter	Mismatch
2	Electrician	Electrician	Match

As it can be inferred from Table 26, second applicant is a match with respect to his/her occupation whereas first applicant is a mismatch.

Skill mismatch is one of the top priority policy issues across the countries in the world. Simply, skill mismatch takes place as the employees have more or fewer skills that the vacancies require (Klosters, 2014). According to 2013 Survey of Adult Skills by OECD, 13% of workers are reported to be underqualified than those required by the vacancies and 21% are overqualified (average of OECD counties) (OECD, 2013). Unfortunately, recent available data for Turkey belongs to 2005, and it implies that 40% of workers are overqualified and 3% are underqualified for the requirements of jobs (Quintini, 2011).

Although skill mismatch defined above and mismatch ratio with respect to occupation calculated in this thesis are different, we believe our results might provide insights to policy makers concerning this issue.

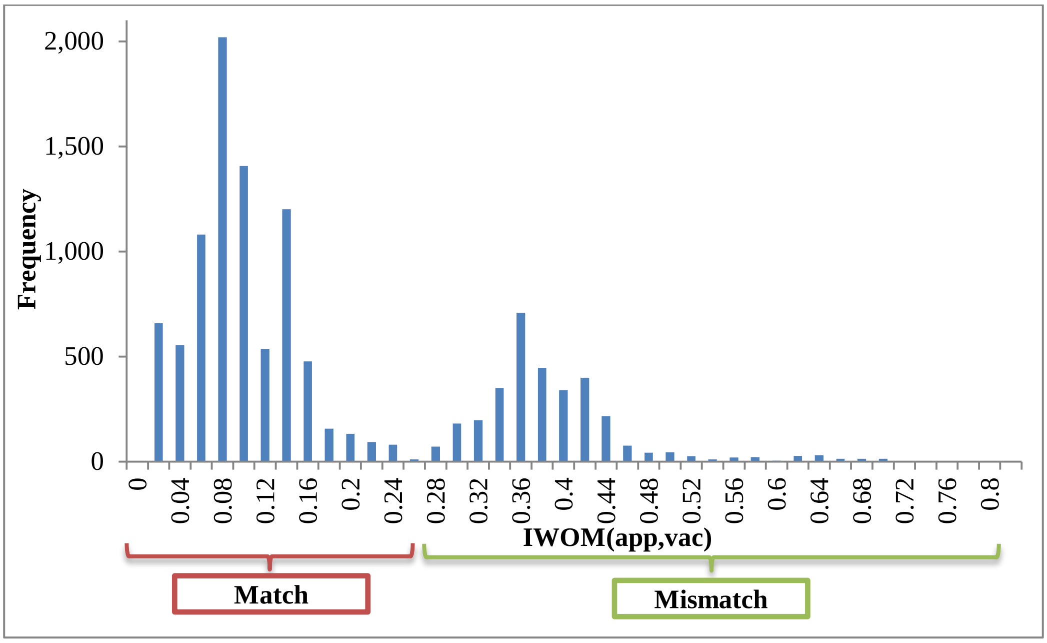

Figure 12. Histogram of IWOM Values for Previous Job Fillings

Figure 12 implies that our IWOM works and we might classify IWOM values with respect to matches in occupation by using $\tau = 0.27$ as global upper bound. In order to clarify this, we refer to Table 25 which shows 11,667 vacancies being filled by applicants covering time period 01.01.2014-12.10.2015. As we investigate these previous job fillings on account of the occupation requirements of the vacancies, we see that 9,212 applicants have the same occupation with the vacancy (i.e., match with respect to occupation). However, in 2,455 job fillings, applicants have different occupations than the vacancy requirements (mismatches). In other words, despite the efforts of TEA, only 78.9% of job fillings occur by satisfying the occupation requirement of the vacancies.

In addition to low level of vacancy filling ratio (only 40.4% of all vacancies registered to TEA were able to be filled in 2014. See section 1.1), matching job applicants with respect to occupation requirement in these fillings is not as high as it is desired. Our proposed matching algorithm can efficiently match job seekers with vacancies and it also prevents mismatching with respect to occupation (Table 27).

Table 27. Classification Summary of IWOM using $\tau = 0.27$ vs. Observed Data with respect to Matches in Occupation

		Predicted by Method		**Total**
		Match	**Mismatch**	
Observed	**Match**	8,415	797	9,212
	Mismatch	0	2,455	2,455
	Total	8,415	3,252	11,667

Table 27 indicates that if we use $\tau = 0.27$ for IWOM values, candidate matches among vacancies and applicants can be classified correctly with respect to occupation. Classification accuracy is 93.17% (see Table 28). Also, we may see that our classification approach correctly classifies all mismatches in the original data which shows that our proposed upper bound seem to work efficiently. In simple words, if our matching algorithm had been used in the first place, no mismatch would have occurred.

Table 28. Performance Estimates of Classification of IWOM vs. Observed Data with respect to Occupation (in %'s)

Sensitivity	100.00
Specificity	75.49
Pos Pred Value	91.35
Neg Pred Value	100.00
Prevalence	72.13
Detection Rate	72.13
Detection Prevalence	78.96
Balanced Accuracy	87.75
Overall Accuracy	**93.17**

In sorting step, rows of **IWOM** are organized in ascending order and indices of applicants are recorded. Applicants who have smaller IWOM values than $\tau = 0.27$ are nominated as candidate matches (in the same order with IWOM values) with the corresponding vacancies.

5.4.4. Matching using sub samples and full data for TEA dataset

In this section, first we split job seekers into 15 sub samples (sampling without replacement) who are classified as VGEO with CACDIFES Algorithm. Major concern here is to decrease the computational cost. Next, we run matching algorithm for each sample which has a size ~2,285 and calculate IWOM for each objects in each sample. Then, we combine all samples to obtain IWOM of all applications who are classified as VGEO. Thus, in full data matching we use

34,267 applicants who are labeled as VGEO (Table 22) and 275,118 vacancies covering time period 01.01.2015 – 12.10.2015 (See section 5.4.1). Finally, eligibility and ineligibility percent of job seekers for vacancies, ratio of filled and unfilled vacancies and average number of matches for a given applicant for any vacancy are calculated (Table 29).

A job seeker is called eligible if he/she is found eligible for at least one vacancy. In other words, if the calculated IWOM for a given job seeker for any vacancy is smaller than a pre-defined threshold τ, then he/she is nominated as eligible. Vacancy filling ratio is defined as the percentage of vacancies which can be filled by any eligible job seekers. Note that a given job seeker can be eligible for more than one vacancy. In order to address this issue we calculate the average number of matches for a given applicant (i.e., average number of vacancies which can be filled by a given job seeker).

Table 29. Matching Results with Sub Samples and Full Data

	Sub Sample 1	**Sub Sample 2**	**Sub Sample 3**	**Sub Sample 4**	**Sub Sample 5**	**Sub Sample 6**	**Sub Sample 7**	**Sub Sample 8**
Sub sample size	2,286	2,284	2,285	2,284	2,285	2,284	2,285	2,284
Eligibility (%) of applicants	91.07	90.54	100	91.59	91.02	91.02	91.16	91.15
Ineligibility (%) of applicants	8.93	9.46	0	8.41	8.98	8.98	8.84	8.85
Vacancy filling ratio (%)	40.95	40.7	54.25	40.98	40.45	40.29	40.65	40.1
Unfilled vacancy ratio (%)	59.05	59.3	45.75	59.02	59.25	59.71	59.35	59.9
Average number of matches for an eligible applicant	54	54	65	53	53	53	53	52
Average number of eligible applicants for a vacancy	11	11	11	11	11	11	11	12

Table 29. (Continued)

	Sub Sample 9	**Sub Sample 10**	**Sub Sample 11**	**Sub Sample 12**	**Sub Sample 13**	**Sub Sample 14**	**Sub Sample 15**	**Full Data**
Sub sample size	2,285	2,284	2,285	2,284	2,284	2,284	2,284	**34,267**
Eligibility (%) of applicants	90.42	91.73	91.68	90.9	91.95	91.37	90.58	**91.14**
Ineligibility (%) of applicants	9.58	8.27	8.32	9.1	8.05	8.63	9.42	**8.86**
Vacancy filling ratio (%)	41.24	53.05	40.97	39.85	41.25	40.89	40.51	**100**
Unfilled vacancy ratio (%)	58.76	46.95	59.03	60.15	58.78	59.11	59.49	**0**
Average number of matches for an eligible applicant	54	69	53	52	54	53	53	**55**
Average number of eligible applicants for a vacancy	11	13	12	12	12	11	11	**165**

Table 29 indicates that classification of job seekers with respect to their employment opportunities seems reasonable, since those who are classified as VGEO in the first phase has a high rate of eligibility in matching. We can see that eligibility percentage of job applicants in all sub samples range from 90.42% to 100%. Similarly, only a small percentage of applicants are found ineligible for any vacancies. Specifically, ineligibility percentage of job applicants in all sub samples are between 0% and 9.58%.

Vacancy filling ratio of the job applicants in each sample is greater than 39.85%. The highest vacancy filling ratio is obtained in sub sample 3 which is 54.25%. Vacancy filling ratios might seem small for each sub sample. Main reason of having smaller percentage of vacancy filling ratios can be explained by smaller sizes of sub samples. Simply, vacancy filing ratio is estimated as the number vacancies which can be filled by applicants over number of all vacancies. Thus, smaller the number of applicants (sub samples), smaller the percentage of vacancy filling ratios. From a different perspective of view, if all job applicants were found to eligible for a single vacancy, than vacancy filling ratio would be estimated as

(2,285/275,118)×100 = 0.83% for each sample. However, even for a sub sample of size ~2,285, we see that vacancy filling ratio is very high which clearly shows that many job applicants are found eligible for many vacancies. Specifically, a job seeker can find job opportunities in at least 52 vacancies on the average for all sub samples. When we investigate the situation from employers' perspective, we see that average number of eligible applicants for a vacancy in each sub sample ranges from 11 to 13. This number shows the number of suitable applicants on the average for an announced vacancy.

Bear in mind that 34,267 applicants are labeled as VGEO out of 175,143 records (19.57%) in 2015 data. Final column of Table 29 indicates that when we use all job applicants which are labeled as VGEO in matching, 91.14% of them are found eligible. Moreover, we see that all 275,118 available vacancies can be filled by these applicants (vacancy filling ratio is 100%). When we investigate the average number of matches for an eligible applicant we see that a job seeker can find job opportunities in 55 vacancies on the average. From employee seekers' side, it is clear that our JMS framework can provide suitable job seekers for given vacancies to employee seekers. In other words, an employer can find 165 suitable applicants on the average for his/her announced vacancy.

Table 29 and comments given above clearly indicates that our statistical approach for JMS first classifies job seekers depending on their eligibility, and then, it efficiently matches them with suitable vacancies.

CHAPTER 6

DISCUSSION AND CONCLUSION

The process of developing JMS requires sophisticated statistical approaches which TEA does not employ as far as we know. In this thesis study, we develop a statistical approach for designing a JMS by proposing a new categorical classification and matching algorithms. Our JMS takes into different specifications of job seekers and vacancies account. In other words, it solves a multivariate MCDM problem by using JSS and VR in two phases. First phase of the approach includes classification of job seekers with respect to their employment opportunities into four classes, and second phase includes matching those who have high chance of transition through labour market (i.e., high chance of filling proper vacancies) with proper vacancies.

This thesis acts as a decision support system for TEA by generating a national-scale JMS which uses a large-scaled micro-level administrative data. As it is described in section 2.3, job matching in TEA are implemented manually by the personnel who work on the behalf of TEA in Employment Offices or Provincial Directorates of TEA. The overall process of directing a job seeker to any given vacancy or any type of vocational training program is done by the expert view of these personnel without using a statistical model which takes employment specific information about high number of different applicants and vacancies in different provinces of Turkey into account. Thus, JMS proposed in this thesis provides a powerful tool for TEA in matching all job seekers with vacancies registered to its database by operating centrally and distributing ideal matches to local offices and Provincial Directorates of TEA.

JMS can be converted to a computer application and embedded to TEA's database by making required adjustments. Then, the system might automatically identify candidate matches between job seekers and vacancies and distribute this information to local units of TEA. Depending on the system performance, a batch job can be run in TEA's system for JMS weekly or monthly which will enable new comers (new applicants and new vacancies registered to TEA's database) to be involved in matching. According to these new comers, status of previous candidate matches can be updated.

JMS Algorithm designed in this thesis is a dynamic model and it can be updated from time to time to reflect the changing behavior of labor market. Algorithm itself does not change but the significance level of attributes, their contributions to total mutual information and weights used in the estimation of IWOM and IWVDM might change. Hence, by only making minor adjustments, overall JMS can be updated easily.

We believe that JMS designed in this thesis will contribute to Turkish economy in a great deal by decreasing the efforts and resources (time, money, physical strength, etc.) of unemployed which are spent on job hunting (from job seeker's perspective) and by lowering the costly process of recruitment (from employee seeker's perspective). Moreover, JMS will help TEA to match more suitable job applicants with vacancies in its database which will eventually help decreasing the unemployment rate in Turkey.

Major technical contributions of this thesis are introducing a new categorical classification algorithm (CACDIFES) which includes an Incremental Feature Selection (IFS) and a new difference metric (IWVDM) and presenting a matching algorithm by making use of scoring and sorting which possess a modified version of overlap metric (IWOM).

We present discussion and conclusion under three main topics, which include technical discussion on the proposed algorithms and methods, policy recommendations concerning job and employment services of TEA and future studies which we plan to conduct.

Our proposed CACDIFES Algorithm includes three main phases, IFS, DC and classification, which can be applied (together or one by one as a pre-processing step) in any type of categorical data classification problem. It can also be used in modeling studies which include categorical data. Since differential entropy (i.e., entropy for continuous variables) exists, our algorithm can be applied to numerical and mixture type of data sets by further modification. However, in some cases, there exists mathematical intractability problem in the estimation of differential entropy for some continuous distributions. Yet, numerical methods can be used to overcome these issues.

Compared to many classification algorithms our proposed algorithm requires only a single input parameter (β, threshold for detecting upper information level) which can be guessed by using the support of the algorithm. Due to practical IFS phase of the algorithm, it can be applied to data sets with high number of attributes having possible correlation structures (such as DNA microarray data sets). Moreover, it can be used in missing value imputation such that class variable can be taken as the attribute which has missing values, and rest of the attributes can be used in classification. Then, missing values will be imputed by using class assignment of classification phase.

As it is discussed in previous chapters, VDM type of metrics make strong assumptions on attribute independence, and they do not take into account the attribute importance (or relevance) for target variable, which highly affect their efficiency in most real-life data sets. In fact, there exist certain issues in the implication of these metrics with different data sets (which have dependence structure among attributes) due to above reasoning. However, our proposed difference metric IWVDM considers correlation structure among attributes and proposes a way of obtaining independent informative attributes. Besides, it weights attributes depending on the information that they possess on class variable and includes them in classification with a user-friendly graphical approach without requiring domain knowledge.

For testing the efficiency of our proposed CACDIFES Algorithm and IWVDM, we use 10 data sets obtained from UCI Machine Learning Repository. According to experimental results of CACDIFES, our proposed metric IWVDM with IFS is superior to other 3 difference metrics FDM, OM and VDM in 8 out of 9 data sets. Although these data sets are commonly used by machine learning algorithm community, some of them (like Mushroom, Vote and segment data sets) should be reconsidered to be used in experiments. That is because, our experimental results show that we end up with significant data reduction when IFS and DC is employed in these data sets. In other words, majority of the attributes do not contain statistically significant information on class variable for those data sets. For instance, mushroom data seems a replication of 8,113 records. Moreover, for some data sets (like Vote), no matter which classification algorithm is used we end up with high classification accuracy.

Experiments with UCI Machine Learning Repository data sets also reveal that classification performance highly changes with respect to using DC. As it is discussed in section 3.2, duplicated records affect the classification outcome; they result in overestimation of the performance measures if data objects are classified correctly, and they result in underestimation if data objects are classified wrongly. Hence, in order to have a reliable classification result, duplicated records should be removed from data in any classification studies.

In the implication of JMS, our proposed CACDIFES algorithm and IWVDM are employed first to our dataset obtained from TEA with two different setups (balanced and imbalanced samples). Initial results imply that our design is slightly better than the other 3 difference metrics FDM, OM and VDM in the case of multi-class imbalanced data, which can be encountered in most real life data sets. Besides, our weight assignment process efficiently represents the true behavior of our real data set.

Matching algorithm proposed in this thesis is based on scoring and sorting. This algorithm is designed to provide a different perspective to multivariate MCDM problems. Majority of the methods for MCDM problems are designed for a single decision making problem where hierarchy of criteria are based on user experience, previous work and empirical studies (i.e., it is highly subjective). However, our proposed method is designed to solve multivariate MCDM problems, and it has a well theoretical basis. It includes a modified version of OM named as IWOM where weights are assigned by using the same procedures as in CACDIFES Algorithm (see 3.1.2.3). With the help of weight assignment, IWOM makes use of attribute independence and importance, while ordinary OM does not.

In order to employ Matching Algorithm to TEA data set, we use VR of previously filled vacancies and JSS of applicants who fill those vacancies. By using this prior information we employ IWOM to these records and obtain skill mismatch of applicants (with respect to occupation) as 27.8% (See Figure 12). Note that concept of skill mismatch (data source and estimation method) given in OECD (2013) and Quintini (2011) are different than the mismatch ratio with respect to occupation found in this thesis. Yet, we provide a different perspective to estimation of skill mismatch problem by using a large-scale micro-level administrative data set.

We calculate a global upper bound which works as an eligibility criterion for identifying applicants who are suitable for given vacancies. Experimental results reveal that out of 34,267 applicants who are classified as VGEO (i.e. 19.57% of all applicants from 2015 data) 91.14% of them are found to be eligible for vacancies of 2015 data. This clearly shows that classification outcome from CACDIFES algorithm is verified by matching algorithm due to high rate of resulting eligibility. In different words, as we employ matching algorithm to all applicants who are classified as VGEO and all available vacancies for 2015 data, we obtain vacancy filling ratio as 100% (all vacant jobs in 2015 data can be filled by applicants with VGEO labels). This clearly indicates that our statistical approach for JMS works successfully and leaves no vacancy without being filled. Although our JMS algorithm works successfully, it is worth to mention that classification accuracy of CACDIFES algorithm (first phase) is not very high with TEA data set (it is slightly above 60%). We believe there exist two main reasons for that. First one is the data quality issues of TEA described in details in sections 5.1 and 5.2. By comparing the efficiency of CACDIFES algorithm with data set from TEA and data sets from UCI Machine Learning Repository, we may suspect that our data set from TEA is not perfectly "separable". For example, in our data set although some job seekers almost have the same specifications (age, gender, occupation, education, location of residence, etc.), some has better access to labour market and worked in at least 2 or more jobs for a longer period while others are unemployed for a long period. This clearly indicates that there are certain factors which ease accession to labour market for certain group of job seekers.

Second, as in most of the studies in literature, employment opportunities increase with higher levels of education. However, this is true when there are more white-collar jobs in labour market. Data set we observed from TEA includes high percentage of job seekers with lower levels of education. 59.8% of applicants have primary education or no degree at all and only 10% of applicants are university graduates in 2015 data. Moreover, most vacancies registered to TEA's database by employee seekers do not require high levels of education (only 4% of vacancies require bachelor's degree or more). In fact, majority of the open vacancies consists of daily jobs requiring muscle power. Thus, percentage of white-collar jobs registered to TEA's database is very low. We may infer that probability of an employment of a job seeker is higher if he/she is previously employed for a certain period of time (Transition from unemployment to employment is higher for that group). Hence, job seekers with similar or the same specifications has different employment opportunities which also lowers the quality of classification in our data set.

In sections 5.1 and 5.2, description of TEA data and a comprehensive data cleaning process is presented. Our data set is an administrative data. Unfortunately, it contains non-standard, non-verified and non-controlled information which have serious adverse effects on the quality of the data and services provided by the TEA. Majority of the information on the database of TEA can be obtained from other governmental bodies' database through web-services. However, as far as we understand from the given TEA data set only a small amount of information is obtained in that way. Address and family history of the applicant, information on social security (previous employment history including number of different jobs, period of employment, etc.), education level (primary, secondary, tertiary or higher education) and driving license are available in databases of governmental authorities who are responsible for these services. TEA should obtain this information through electronic channels to increase the efficiency of the services it provides. Moreover, there exists more useful information in other institutions' databases which is not taken by TEA, too. For example information on social assistance status and land registry system can provide significant information in prioritization of job applicants in directions to vacancies by TEA.

There are also certain issues with TEA's registry system which can be fixed by making minor modifications. Information entered in free-text fields should be converted to drop-down list or any type which can be categorized. For example, education level is defined as free-text field in TEA's system and according to given data, there are more than 10,000 university departments entered by the applicants. After data cleaning, this number reduces approximately to 3,300 departments. By using data cleaning process, we end up a "clean data set" that TEA does not have. Although our purpose is fully academic, a similar and a more professional data cleaning process should be applied by TEA to increase its data quality so that a more efficient JMS process can be employed.

Another structural issue is the type of diplomas provided by educational institutions or authorities in higher education, tertiary education or vocational high schools. Although there are many similar departments in these educational institutions, which run similar curriculums, they might give different diplomas for their graduates, which do not fit in labour market. For instance in TEA data set, there exist more than 77 university departments which include "computer" somewhere in department name entered by job applicants as highest degree acquired. However, majority of the applicants who can fill vacancies are graduates of 15 out of 77 departments which clearly show that rest of those applicants are not good matches for the vacancies (i.e. they do not have the skills that labour market needs). This is one serious issue to be solved for increasing the employment in Turkey. It is known that Higher Education Council of Turkey is the main responsible body for providing official permission to open new departments in universities. Hence, required measures should be taken by the co-operation of TEA, Ministry of National Education, Higher Education Council of Turkey and universities so as to meet the needs of labour market.

Education programs and curriculums provided by educational institutions, schools or universities in Turkey cannot efficiently provide new skills and qualifications required by the labour market. Thus, another measure is to redesign curriculums and programs by conducting labour force demand surveys with the co-operation of responsible institutions given above.

Future studies that we plan to do mainly focus on algorithms and methods that we propose in this thesis. First, we plan to modify CACDIFES Algorithm so that it can be applied to numerical or mixture type of data sets. Second, we will design a simulation study to compare the efficiency of CACDIFES algorithm by generating different data sets which include high number of attributes with high rate of dependence among them. Third, CACDIFES Algorithm will be used as a missing value imputation tool and its efficiency will be compared with other imputation methods.

As it is discussed in majority of the studies, many classification algorithms are originally designed to be used in binary classification problems and further modifications enabled them to be used in multi-class problems. As a future study we will compare the efficiency of our CACDIFES on different data sets which have multiple classes and imbalanced nature.

As discussed in section 3.1.2.1, there is not a hypothesis testing procedure to identify if $I\left(C;A_{j+1},A_1,\dots,A_j\right)-I\left(C;A_j,A_1,\dots,A_{j-1}\right)=0$. Also, its distribution is mathematically intractable. We believe that a proposal for such a hypothesis test is out of the scope of this thesis study, but we plan to develop one as a future study.

As it is given in section 3.1.2.1, exact distribution of symmetric uncertainty is unknown, and there is not a statistical test for checking whether the estimated symmetric uncertainty is different from 0 or above some certain value so that we may infer the existence of high dependence structure between any two attributes. Thus, taking into account the value of symmetric uncertainty which is between 0 and 1, user might decide on the dependence structure among attributes using $\mathbf{U}$ with a certain threshold value. As a future work, we plan to study the statistical properties of symmetric uncertainty and we will try to find the exact (or asymptotic) distribution of it. Besides, we will apply sensitivity analysis for the threshold value for determining the dependence level of attributes.

Another future study is to make suitable modifications in matching algorithm so that it can be applied to any type of matching problems which include numerical or mixture type of high dimensional data sets.

REFERENCES

Aggarwal, C. C. (2015). *Data classification : algorithms and applications*. *Series: Chapman & Hall/CRC data mining and knowledge discovery series ; 35*, http://www.worldcat.org/oclc/890721171\nhttps://www.chapters.indigo.ca/en-ca/books/data-classification-algorithms-and-applications/9781466586758-item.html (Retrieved on: 20.02.2016).

Al Aghbari, Z. (2010). Classification of Categorical and Numerical Data on Selected Subset of Features, Bayesian Network, Ahmed Rebai (Ed.), InTech, 355-367. Available from: http://www.intechopen.com/books/bayesian-network/classification-of-categorical-and-numerical-data-on-selected-subset-of-features

Aha, D. W. (1992). Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, *36*(2), 267–287.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-Based Learning Algorithms. *Machine Learning*, *6*(1), 37–66.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, 267–281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactionson Automatic Control*, *19*(6), 716–723.

Alba, A., Arranz, J. M., & Muñoz-bullón, F. (2012). Re-employment probabilities of unemployment benefit recipients. *Applied Economics*, *44*(28), 3645–3664.

Alfons, A. (2012). cvTools: Cross-validation tools for regression models. R package version 0.3.2. https://CRAN.R-project.org/package=cvTools. (Retrieved on: 03.05.2016).

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. Academic Press, New York.

Andrews, M. J., Bradley, S., & Upward, R. (2001). Estimating the probability of a match using microeconomic data for the youth labour market. *Labour Economics*, *8*(3), 335–357.

Aranganayagi, S., & Thangavel, K. (2010). Incremental Algorithm to Cluster the Categorical Data with Frequency Based Similarity Measure, *4*(1), 1251–1259.

Archana, S., & Elangovan, K. (2014). Survey of Classification Techniques in Data Mining. *International Journal of Computer Science and Mobile Applications*, *2*(2), 65–71.

Arcidiacono, P., Beauchamp, A., & Mcelroy, M. (2016). Terms of Endearment: An Equilibrium Model of Sex and Matching. *Quantitative Economics*, *2524*(5), 1–42.

Bartel, A., & Borjas, G. (1981). Wage growth and job turnover: An empirical analysis. *Studies in Labor Markets*, *I*, 65–90. http://www.nber.org/chapters/c8908.pdf. (Retrieved on: 03.03.2016).

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is “nearest neighbor” meaningful? In *Database Theory—ICDT ’99, 7th international conference, Jerusalem, Israel, Proceedings*, volume 1540 of *Lecture Notes in Computer Science*, 217–235. London, UK: Springer.

Bigi, B. (2003). Using Kullback-Leibler distance for text categorization. *Proceeding ECIR’03 Proceedings of the 25th European Conference on IR Research*, 305–319.

Blanchard, O. J., & Diamond, P. (1989). The Aggregate Matching Function. Massachusetts institute of technology, *NBER Working Paper*, (No: w3175). https://dspace.mit.edu/bitstream/handle/1721.1/63290/aggregatematchin00blan.pdf?sequence=1. (Retrieved on: 03.08.2015).

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*(1-2), 245–271.

Bogdanovic, D., Nikolic, D., & Ivana, I. (2012). Mining method selection by integrated AHP and PROMETHEE method. *Anais Da Academia Brasileira de Ciencias*, *84*(1), 219–233.

Boone, J., Sadrieh, A., & van Ours, J. C. (2009). Experiments on unemployment benefit sanctions and job search behavior. *European Economic Review*, *53*(8), 937–951.

Boser, B. E., Laboratories, T. B., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory (COLT' 92)*, 144–152. http://dl.acm.org/citation.cfm?id=130401. (Retrieved on: 03.07.2016).

Brillinger, D. (2004). Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics*, *18*(January), 163–182. www.redeabe.org.br/bjpspublishedpapers_volume18_2_pp163-182.pdf (Retrieved on: 15.05.2016).

Bryman, A., & Cramer, D. (1997). *Quantitative data analysis with SPSS for Windows: A Guide for Social Scientists*. Routledge, New York, NY.

Burda, M., & Wyplosz, C. (1994). Gross worker and job flows in Europe. *European Economic Review, 38*(6), 1287–1315.

Bufardi, A., Gheorghe, R., Kiritsis, D., & Xirouchakis, P. (2004). Multicriteria decision-aid approach for product end-of-life alternative selection. *International Journal of Production Research, 42*(16), 3139-3157.

Carbonera, J. L., & Abel, M. (2014a). An Entropy-Based Subspace Clustering Algorithm for Categorical Data. *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, 272–277.

Carbonera, J. L., & Abel, M. (2014b). Categorical Data Clustering: A Correlation-Based Approach for Unsupervised Attribute Weighting. *In 2014 IEEE 26th International Conference on Tools with Artificial Intelligence,* 259–263.

Caucutt, E. M., Guner, N., & Knowles, J. (2002). Why do women wait? Matching, wage inequality and incentives for fertility delay. *Review of Economic Dynamics*, *5*(4), 815–855.

Christensen, R. (1997). *Log-linear models and logistic regression, Second Edition*. Springer-Verlag, Inc., New York.

Christensen, B. J., Lentz, R., Mortensen, D. T., Neumann, G. R., & Werwatz, A. (2005). On-the-Job Search and the Wage Distribution. *Journal of Labor Economics, 23*(1), 31–58.

Costa, E. P., Postal, C., Lorena, A. C., & Ad, R. S. (2007). A Review of Performance Evaluation Measures for Hierarchical Classifiers. *Evaluation Methods for Machine Learning II: Papers from the AAAI-2007 Workshop*, 1–6.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Cramér, Harald. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press, New Jersey, 282.

Emerson, J.W., & Kane, M.J. (2016). biganalytics: Utilities for 'big.matrix' Objects from Package 'bigmemory'. R package version 1.1.14. https://CRAN.R-project.org/package=biganalytics. (Retrieved on: 03.05.2016).

Erceg, C. J., & Levin, A. T. (2014). Labor Force Participation and Monetary Policy in the Wake of the Great Recession. *Journal of Money, Credit, and Banking, 46*(2), 3–49.

Eurostat. (2008). NACE Rev. 2 Statistical classification of economic activities in the European Community. *Methodologies and Working papers*. *Luxembourg: Office for Official Publications of the European Communities*.

Fletcher, R. H., Fletcher, S. W., & Fletcher, G. S. (2012). *Clinical epidemiology: the essentials*. Lippincott Williams & Wilkins, Philadelphia, 118.

Flinn, C., & Heckman, J. (1982). New Methods for Analyzing Structural Models of Labor Force Dynamics. *Journal of Econometrics, 18*(1), 115–168.

Flinn, C. J. (2002). Interpreting minimum wage effects on wage distributions: a cautionary tale. *Annales d'E´ conomie et de Statistique*, (67/68), 309–355.

Flinn, C. J. (2006). Minimum wage effects on labor market outcomes under search, matching, and endogenous contact rates. *Econometrica*, *78*(4), 1013-1062.

Foster, L., Haltiwanger, J. C., & Krizan, C. J. (2001). Aggregate productivity growth. Lessons from microeconomic evidence. In *New developments in productivity analysis*, University of Chicago Press, National Bureau of Economic Research, Chicago, 303-372.

Frederiksen, A. (2008). Gender differences in job separation rates and employment stability: New evidence from employer-employee data. *Labour Economics*, *15*(5), 915–937.

Fujikawa, Y., & Ho, T. (2002, May). Cluster-based algorithms for dealing with missing values. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer Berlin Heidelberg, 549-554.

Gan, G., Ma, C., & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. *ASASIAM Series on Statistics and Applied Probability* (Vol. 20). SIAM, Philadelphia, ASA, Alexandria, VA.

Ganti, V., Gehrke, J., & Ramakrishnan, R. (1999). CACTUS-clustering categorical data using summaries. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '99*, 73–83.

Gordon, A. (1999). *Classification*, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC, London-New York.

Gray, M. C., & Hunter, B. H. (2002). A Cohort Analysis of the Determinants of Employment and Labour Force Participation: Indigenous and Non-Indigenous Australians, 1981 to 1996. *Australian Economic Review*, *35*(4), 391–404.

Guha, S., Rastogi, R., & Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. *ACM SIGMOD Record*, *27*(2), 73–84.

Guyon, I., & Elisseeff, a. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(2003), 1157–1182.

Hartigan, J. A. (1975). Clustering Algorithms. *Information Retrieval Data Structures and Algorithms, 2,* 419–442.

Hinneburg, A., & Keim, D. A. (1999). Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. *International Conference on Very Large Databases (VLDB),* 506–517.

ILO (1982). *Statistics of Labour Force, Employment, Unemployment and Underemployment*. Thirteenth International Conference of Labour Statisticians, Geneva, 18-29 October 1982. International Labour Organization. (Retrieved from http://embargo.ilo.org/public/libdoc/ilo/1982/82B09_438_engl.pdf).

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

Jiang, L., & Li, C. (2013). An augmented value difference measure. *Pattern Recognition Letters, 34*(10), 1169–1174.

Jiang, L., & Li, C. (2011). An empirical study on class probability estimates in decision tree learning. *Journal of Software*, *6*(7), 1368–1373.

Jiang, L., LI, C., Zhang, H., & Cai, Z. (2014). a Novel Distance Function: Frequency Difference Metric. *International Journal of Pattern Recognition and Artificial Intelligence, 28*(02), 1451002.

Jolivet, G., Postel-Vinay, F., & Robin, J. M. (2006). The empirical content of the job search model: Labor mobility and wage distributions in Europe and the US. *European Economic Review*, *50*(4), 877-907.

Kahn, L. M. (2012). Temporary jobs and job search effort in Europe. *Labour Economics*, *19*(1), 113–128.

Kane, M.J., Emerson, J. W., & Weston, S. (2013). Scalable Strategies for Computing with Massive Data. *Journal of Statistical Software*, 55(14), 1-19. URL http://www.jstatsoft.org/v55/i14/. (Retrieved on: 03.05.2016).

Kasif, S., Salzberg, S., Waltz, D., Rachlin, J., & Aha, D. W. (1998). A probabilistic framework for memory-based reasoning. *Artificial Intelligence*, *104*(1-2), 287–311.

Kaufman, L., and Rousseeuw, P. (1990). *Finding Groups in Data- An Introductory to Cluster Analysis*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Kiefer, N. M., & Neumann, G. R. (1979). An empirical job-search model, with a test of the constant reservation-wage hypothesis. *The Journal of Political Economy*, 89-107.

Klosters, D. (2014). *Matching Skills and Labour Market Needs: Building Social Partnerships for Better Skills and Better Jobs*. World Economic Forum Global Agenda Council on Employment. Switzerland, 22-25 January. (Retrieved from http://www3.weforum.org/docs/GAC/2014/WEF_GAC_Employment_MatchingSkillsLabourMarket_Report_2014.pdf)

Koçak, O. & Akman, A. C. (2011). İşsizlikle Mücadelede İş Danışmanlık Hizmetleri ve Yalova Örneği, *İş, Güç Endüstri İlişkileri ve İnsan Kaynakları Dergisi*, 13(2), 135- 154.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273–324.

Kullback, S. (1959). *Statistics and Information Theory*. (Vol. 1). John Wiley & Sons, Inc., New York.

Kullback, S., and Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1):79–86.

Kumas, H. (2010). Türkiye İş Kurumu Faaliyetleri ve İşgücü Piyasası İhtiyaçları Arasındaki Uyum. *Sosyoekonomi*, *11*(11), 131-166.

Kyyrä, T., & Ollikainen, V. (2008). To search or not to search? The effects of UI benefit extension for the older unemployed. *Journal of Public Economics*, *92*(10-11), 2048–2070.

Lentz, R., & Mortensen, D. T. (2008). An Empirical Model of Growth through Product Innovation. *Econometrica*, *76*(6), 1317–1373.

Li, C., & Li, H. (2011). One dependence value difference metric. *Knowledge-Based Systems*, *24*(5), 589–594.

Li, C., Jiang, L., Li, H., & Wang, S. (2013). Attribute Weighted Value Difference Metric. In *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, 575-580.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Lindeboom, M., Ours, J. Van, & Renes, G. (1994). Matching Employers and Workers: An Empirical Analysis on the Effectiveness of Search. *Oxford Economic Papers*, *46*(1), 45–67.

Liu, Y., & Kender, J. R. (2003). Fast video segment retrieval by sort-merge feature selection, boundary refinement, and lazy evaluation. *Computer Vision and Image Understanding*, *92*(2-3), 147–175.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1*(14), 281–297.

Mergias, I., Moustakas, K., Papadopoulos, A., & Loizidou, M. (2007). Multi-criteria decision aid approach for the selection of the best compromise management scheme for ELVs: The case of Cyprus. *Journal of Hazardous Materials*, *147*(3), 706-717.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*(2), 181–204.

Mincer, J. (1997). The production of human capital and the lifecycle of earnings: Variations on a theme. *Journal of Labor Economics*, *15*(1), 26–47.

Mincer, J. (1986). *Wage changes in job changes*. National Bureau of Economic Research, Cambridge, Mass., USA.

Mitchell, T. M. (1997). Instance-based learning. *Machine Learning*, *1*, 231-236.

Mortensen, D. T. (1986). Chapter 15: Job search and labor market analysis. *Handbook of Labor Economics*, *2*(11), 849–919.

Mortensen, D. T., & Pissarides, C. A. (1999). Chapter 39: New developments in models of search in the labor market. *Handbook of Labor Economics*, *3*(2), 2567–2627.

Motwani, R., & Raghavan, P. (1995). *Randomized Algorithms*. Cambridge University Press, New York.

Novomestky, F. (2012). matrixcalc: Collection of functions for matrix calculations. R package version 1.0-3. https://CRAN.R-project.org/package=matrixcalc. (Retrieved on: 03.05.2016).

OECD (2013). OECD Skills Outlook 2013: First Results from the Survey of Adult Skills. OECD, Paris, France. available at: http://dx.doi.org/10.1787/9789264204256-en.

OECD (2016a), Unemployment rate (indicator). doi: 10.1787/997c8750-en (Retrieved on 02.01.2017)

OECD (2016b), Youth unemployment rate (indicator). doi: 10.1787/c3634df7-en (Retrieved on 08.11.2016).

OECD (2016c). Long-term unemployment rate (indicator). doi: 10.1787/76471ad5-en (Retrieved on: 17.05.2016).

Öz, S. (2010). “Nobel Ekonomi Ödülü: Arama ve Eşleştirme Modelleri”. Politika Notu: 10-17, Kasım 2010. Ekonomik Araştırma Forumu (EAF) TÜSİAD-Koç Üniversitesi.

Patrick, E. M. (2014). infotheo: Information-Theoretic Measures. R package version 1.2.0. https://CRAN.R-project.org/package=infotheo. (Retrieved on: 03.05.2016).

Phyu, T. N. (2009). Survey of Classification Techniques in Data Mining. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, *1*, 18–20.

Pissarides, C.A. (2000). *Equilibrium Unemployment Theory*, Second Edition, MIT Press, Cambridge, London, England.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies 2*(1): 37–63.

Quintini, G. (2011). *Right for the Job: Over-qualified or Under-skilled?*. DELSA/ELSA/WD/SEM, Employment and Migration Working Papers (No. 120). OECD Publishing, Paris, France.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1988). *Numerical recipes example book (C)* (Vol. 2). Cambridge University Press, New York, NY.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahm, E., & Do, H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, *23*(4), 3–13.

Romanski, P. & Kotthoff, L. (2014). FSelector: Selecting attributes. R package version 0.20. https://CRAN.R-project.org/package=FSelector. (Retrieved on: 03.05.2016).

Sakshi, and Khare, S. (2015). A Comparative Analysis of Classification Techniques on Categorical Data in Data Mining. *International Journal on Recent and Innovation Trends in Computing and Communication*, *3*(8), 5142 – 5147.

Saaty, T. L. (1990). How to make a decision: the analytic hierarchy process. *European journal of operational research*, *48*(1), 9-26.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, *1*(1), 83.

Saranya, V.M., Uma, S., Sherin, A., & Saranya, K. (2014). Survey on Classification Techniques Used in Data Mining and their Recent Advancements. *International Journal of Science*, Engineering and Technology Research, *3*(9), 2278 – 7798.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*(4), 623-656.

Smith, T. E., & Zenou, Y. (2003). A discrete-time stochastic model of job matching. *Review of Economic Dynamics, 6*(1), 54–79.

Soundarya, M., & Balakrishnan, R. (2014). Survey on Classification Techniques in Data mining. *International Journal of Advanced Research in Computer and Communication Engineering, 3*(7), 7550-7552.

Srinivasa, S. (2003). A Review on Multivariate Mutual Information. *University of Notre Dame, Notre Dame, Indiana, 2,* 1–6. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.6038&amp;rep=rep1&amp;type=pdf. (Retrieved on 06.08.2015).

Stanfill, C., & Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, *29*(12), 1213–1228.

Steven Davis, J. H., Schuh, S., & John. (1996). Job Creation and Destruction. *MIT Press*, *87*, 174–183. http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=5745. (Retrieved on: 03.08.2015).

Sun Han, T. (1980). Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, *46*(1), 26–45.

Sujatha, M., Prabhakar, S., & Devi, D. G. L. (2013). A Survey of Classification Techniques in Data Mining. *International Journal of Innovations in Engineering and Technology (IJIET)*, *2*(4), 86-92.

Taşçı, H. M., & Tansel, A. (2005). Youth unemployment duration in Turkey. *METU Studies in Development*, *32*(2), 517-545.

Taşçı, H. M. (2008). Search and determinants of job search intensity in Turkey1. *METU Studies in Development*, *35*(2), 399-425.

Taşçı, H. M. & Darıcı, B. (2009). Türkiye'de İşsizliğin Mikro Veri İle Farklı Tanımlar Altında, Cinsiyet Ayırımına Göre Analizi, *Elektronik Sosyal Bilimler Dergisi*, *8*(28), 139-159.

Tatsiramos, K. (2009). Unemployment Insurance in Europe: Unemployment Duration and Subsequent Employment Stability. *Journal of the European Economic Association*, *7*(6), 1225–1260.

TEA. (2013a). Türkiye İş Kurumu Genel Müdürlüğü 7. Genel Kurul Çalışma Raporu. İŞKUR, Ankara, Turkey. Available at http://www.iskur.gov.tr/tr-tr/kurumsalbilgi/raporlar.aspx#dltop._(Retrieved on: 15.02.2015).

TEA. (2013b). Türkiye İş Kurumu Genel Müdürlüğü 2013 Yılı Faaliyet Raporu. İŞKUR, Ankara, Turkey. Available at http://www.iskur.gov.tr/tr-tr/kurumsalbilgi/raporlar.aspx#dltop. (Retrieved on: 15.02.2015).

TEA. (2014a). Türkiye İşgücü Piyasası Analizi, (2014 1. Dönem). İŞKUR, Ankara, Turkey. Available at http://www.iskur.gov.tr/tr-tr/kurumsalbilgi/raporlar.aspx#dltop. (Retrieved on: 17.02.2015).

TEA. (2014b). Türkiye İş Kurumu Genel Müdürlüğü 2014 Yılı Faaliyet Raporu. İŞKUR, Ankara, Turkey. Available at http://www.iskur.gov.tr/tr-tr/kurumsalbilgi/raporlar.aspx#dltop. (Retrieved on: 05.11.2015).

Tuszynski, J. (2014). caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.. R package version 1.17.1. https://CRAN.R-project.org/package=caTools. (Retrieved on: 03.05.2016).

Veracierto, M. (2008). On the cyclical behavior of employment, unemployment and labor force participation. *Journal of Monetary Economics*, *55*(6), 1143-1157.

Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, *6*(1997), 1–34.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3-36.

World Bank. (2014). Turkey’s Transitions: Integration, Inclusion, Institutions. World Bank Report No: 90509-TR.,Ankara, Turkey.

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003) Washington DC*, *3*, 1–8.

Zhang, P., Wang, X., & Song, P. X.-K. (2006). Clustering Categorical Data Based on Distance Vectors. *Journal of the American Statistical Association, 101*(473), 355–367.

APPENDIX A

NACE REV. 1.1 ECONOMIC ACTIVITY CODES

Table 30. NACE Rev. 1.1 Economic Activity Codes (Eurostat, 2008)

L1	Description	L2	Description
A	Agriculture, Hunting and Forestry	AA	Agriculture, Hunting and Forestry
B	Fishing	BA	Fishing
C	Mining and Quarrying	C	Mining and quarrying
		CA	Mining and quarrying of energy producing materials
		CB	Mining and quarrying, except of energy producing materials
D	Manufacturing	D	Manufacturing
		DA	Manufacture of food products, beverages and tobacco
		DB	Manufacture of textiles and textile products
		DC	Manufacture of leather and leather products
		DD	Manufacture of wood and wood products
		DE	Manufacture of pulp, paper and paper products; publishing and printing
		DF	Manufacture of coke, refined petroleum products and nuclear fuel
		DG	Manufacture of chemicals, chemical products and man-made fibers
		DH	Manufacture of rubber and plastic products
		DI	Manufacture of other non-metallic mineral products
		DJ	Manufacture of basic metals and fabricated metal products
		DK	Manufacture of machinery and equipment n.e.c.
		DL	Manufacture of electrical and optical equipment
		DM	Manufacture of transport equipment
		DN	Manufacturing n.e.c.

Table 30. (Continued)

L1	Description	L2	Description
E	Electricity, Gas and Water Supply	EA	Electricity, Gas and Water Supply
F	Construction	FA	Construction
G	Wholesale and Retail Trade: Repair of Motor Vehicles, Motorcycles and Personal and Household Goods	GA	Wholesale and Retail Trade: Repair of Motor Vehicles, Motorcycles and Personal and Household Goods
H	Hotels and Restaurants	HA	Hotels and Restaurants
I	Transport, Storage and Communications	IA	Transport, Storage and Communications
J	Financial Intermediation	JA	Financial Intermediation
K	Real Estate, Renting and Business Activities	KA	Real Estate, Renting and Business Activities
L	Public Administration and Defense; Compulsory Social Security	LA	Public Administration and Defense; Compulsory Social Security
M	Education	MA	Education
N	Health and Social Work	NA	Health and Social Work
O	Other Community, Social and Personal Services Activities	OA	Other Community, Social and Personal Services Activities
P	Activities of Private Households as Employers and Undifferentiated Production Activities of Private Households	PA	Activities of Private Households as Employers and Undifferentiated Production Activities of Private Households
Q	Extraterritorial Organizations and Bodies	QA	Extraterritorial Organizations and Bodies

APPENDIX B

VARIABLE TYPES FOR VACANCY AND APPLICANT DATASET

Table 31. Variable Types for Vacancy Dataset

No	Variables	Data Type
1	Vacancy ID	Numerical
2	Employee Seeker Type (Public or Private)	Categorical
3	Name of the Firm	Categorical
4	Vacancy Announcement Date	Numerical
5	**Gender**	**Categorical**
6	**Min. Age Requirement for Males**	**Numerical**
7	**Max Age limit for Males**	**Numerical**
8	**Min. Age Requirement for Females**	**Numerical**
9	**Max Age limit for Females**	**Numerical**
10	Driving License	Categorical
11	**Min. Education Level**	**Categorical**
12	Highest Education Level Required	Categorical
13	Economic Activity Description	Categorical
14	Economic Activity Code (NACE Rev. 1.1)	Categorical
15	**Occupation**	**Categorical**
16	Previous Experience (in Years)	Numerical
17	Previous Experience (in Months)	Numerical
18	**Experience Type**	**Categorical**
19	Certifications	Categorical
20	Foreign Language Requirement	Categorical
21	Period of Work	Categorical
22	Type of the Contract	Categorical
23	Number of Total Vacancies	Numerical
24	Vacancy Location (Country)	Categorical
25	**Vacancy location (Province)**	**Categorical**
26	Vacancy location (District)	Categorical
27	Residential Preference (Province - District)	Categorical

Table 32. Variable Types for Applicant Dataset

No	Variables	Data Type
1	Applicant ID	Numerical
2	Date of Birth	Numerical
3	**Age**	**Numerical**
4	Marital Status	Categorical
5	Gender	Categorical
6	**Residence (Province)**	**Categorical**
7	Residence (District)	Categorical
8	**State of Workforce**	**Categorical**
9	First Registration Date	Numerical
10	Registration Renewal Date	Numerical
11	State of Job Search	Categorical
12	Social State	Categorical
13	Disability Status	Categorical
14	Condemned Status	Categorical
15	Priority Status	Categorical
16	**#of Unemployment Benefits Received**	**Numerical**
17	**Total Payment of Unemployment Benefits**	**Numerical**
18	**#of Days that Unemployment Benefits Received**	**Numerical**
19	Driving License	Categorical
20	**Education Level**	**Categorical**
21	**Education Level (School Type)**	**Categorical**
22	**Education Level (Department)**	**Categorical**
23	Education Level (Graduation Year)	Categorical
24	**Occupation**	**Categorical**
25	**Occupation (Experience Type)**	**Categorical**
26	Occupation (Experience in Years)	Numerical
27	Occupation (Experience in Months)	Numerical
28	Additional Information (courses and certificates)	Categorical
29	#of Job Applications	Numerical
30	#of Appointments with CJOs	Numerical
31	# of Vacancy Fillings	Numerical
32	Last Filled Vacancy	Categorical
33	Last Filled Vacancy (Vacancy ID)	Numerical
34	Last Filled Vacancy (Province of Vacancy)	Categorical
35	Last Filled Vacancy (District of Vacancy)	Categorical
36	Last Filled Vacancy (Occupation)	Categorical
37	Last Filled Vacancy (Economic Activity Code)	Categorical
38	Last Filled Vacancy (Date of Application)	Numerical
39	Last Filled Vacancy (Date of Vacancy Filling)	Numerical
40	Location preference for work (Province)	Categorical
41	Preference of Period of Work	Categorical
42	Preference of Type of the Contract	Categorical
43	Preference of Economic Activity	Categorical
44	Opt for Vocational Training	Categorical
45	**Previous Job Experience**	**Categorical**
46	Previous Job Experience (Firm Name)	Categorical
47	Previous Job Experience (Economic Activity Code)	Categorical
48	Previous Job Experience (Vacancy level - Position)	Categorical
49	Previous Job Experience (Date of Entrance)	Numerical
50	Previous Job Experience (Date of leaving)	Numerical

Table 32. (Continued)

No	Variables	Data Type
51	Foreign Language	Categorical
52	Vocational Trainings Attended	Categorical
53	Vocational Trainings Attended (Training ID)	Numerical
54	Vocational Trainings Attended (Date of Start)	Numerical
55	Vocational Trainings Attended (Date of Finish)	Numerical
56	Vocational Trainings Attended (Training Type)	Categorical
57	Vocational Trainings Attended (Occupation Type)	Categorical
58	Vocational Trainings Attended (Min. Age Requirement)	Numerical
59	Vocational Trainings Attended (Max. Age Requirement)	Numerical
60	Vocational Trainings Attended (Training Location - Province)	Categorical
61	Vocational Trainings Attended (Training Location - District)	Categorical
62	#of Job Interviews	Numerical
63	Final Interview date	Numerical

Variables in bold letters, which are listed in Table 31 and Table 32, show the attributes which contain statistically significant information with respect to their class variables, after IFS is employed.

APPENDIX C

DETAILED EXPERIMENTAL RESULTS OF UCI DATA SETS

Table 33. Performance Estimates of Classification on Breast Cancer Data

Breast-cancer data	**With IFS**				**Without IFS**			
For *k*=9	**IWVDM**	**VDM**	**FDM**	**OM**	**IWVDM**	**VDM**	**FDM**	**OM**
Accuracy	**75.58**	62.79	72.09	74.42	67.44	**70.93**	69.77	68.6
Sensitivity	93.22	66.10	91.53	91.53	87.50	91.07	85.71	92.86
Specificity	37.04	55.56	29.63	37.04	30.00	33.33	40.00	23.33
Pos Pred Value	76.39	76.47	73.97	76.06	70.00	71.83	72.73	69.33
Neg Pred Value	71.43	42.86	61.54	66.67	56.25	66.67	60.00	63.64
Prevalence	68.60	68.60	68.60	68.60	65.12	65.12	65.12	65.12
Detection Rate	63.95	45.35	62.79	62.79	56.98	59.30	55.81	60.47
Detection Prevalence	83.72	59.30	84.88	82.56	81.40	82.56	76.74	87.21
Balanced Accuracy	65.13	65.13	60.58	64.28	58.75	62.20	62.86	58.10

Table 34. Performance Estimates of Classification on Credit-g Data

Credit-g	**With IFS**				**Without IFS**			
For *k*=7	**IWVDM**	**VDM**	**FDM**	**OM**	**IWVDM**	**VDM**	**FDM**	**OM**
Accuracy	**71.48**	67.68	69.58	68.82	74	73.67	72	**74.3**
Sensitivity	39.33	35.96	31.46	32.58	44.00	45.33	37.33	38.67
Specificity	87.93	83.91	89.08	87.36	84.00	83.11	83.56	86.22
Pos Pred Value	62.50	53.33	59.57	56.86	47.83	47.22	43.08	48.33
Neg Pred Value	73.91	71.92	71.76	71.70	81.82	82.02	80.00	80.83
Prevalence	33.84	33.84	33.84	33.84	25.00	25.00	25.00	25.00
Detection Rate	13.31	12.17	10.65	11.03	11.00	11.33	9.33	9.67
Detection Prevalence	21.29	22.81	17.87	19.39	23.00	24.00	21.67	20.00
Balanced Accuracy	63.63	59.93	60.27	59.97	64.00	64.22	60.44	62.44

Table 35. Performance Estimates of Classification on Diabetes Data

Diabetes	**With IFS (*)**				**Without IFS**			
For *k*=3	**IWVDM**	**VDM**	**FDM**	**OM**	**IWVDM**	**VDM**	**FDM**	**OM**
Accuracy	80	80.87	80	**81.3**	**74.35**	72.61	72.17	71.13
Sensitivity	90.12	90.74	90.12	91.36	82.12	78.81	78.15	78.15
Specificity	55.88	57.35	55.88	57.35	59.49	60.76	60.76	58.23
Pos Pred Value	82.95	83.52	82.95	83.62	79.49	79.33	79.19	78.15
Neg Pred Value	70.37	72.22	70.37	73.58	63.51	60.00	59.26	58.23
Prevalence	70.43	70.43	70.43	70.43	65.65	65.65	65.65	65.65
Detection Rate	63.48	63.91	63.48	64.35	53.91	51.74	51.30	51.30
Detection Prevalence	76.52	76.52	76.52	76.96	67.83	65.22	64.78	65.65
Balanced Accuracy	73.00	74.05	73.00	74.36	70.81	69.78	69.45	68.19

(*) Observed performance increase due to FS in VDM, FDM and OM

Table 36. Performance Estimates of Classification on Ionosphere Data

Ionosphere	**With IFS**				**Without IFS**			
For *k*=5	**IWVDM**	**VDM**	**FDM**	**OM**	**IWVDM**	**VDM**	**FDM**	**OM**
Accuracy	**93.33**	90.48	88.57	81.63	**76.96**	71.3	73.48	76.09
Sensitivity	88.10	83.33	73.81	80.95	81.46	76.10	78.15	82.12
Specificity	96.83	95.24	98.41	98.41	68.35	62.03	64.56	64.56
Pos Pred Value	94.87	92.11	96.87	97.14	83.11	79.31	80.82	81.58
Neg Pred Value	92.42	89.55	84.93	88.57	65.85	57.65	60.71	65.38
Prevalence	40.00	40.00	40.00	40.00	65.65	65.65	65.65	65.65
Detection Rate	35.24	33.33	29.52	32.38	53.48	50.00	51.30	53.91
Detection Prevalence	37.14	36.19	30.48	33.33	64.35	63.04	63.48	66.09
Balanced Accuracy	92.46	89.29	86.11	89.68	74.91	69.09	71.35	73.34

Table 37. Performance Estimates of Classification on SPECT Data

SPECT	**With IFS**				**Without IFS**			
For *k*=5	**IWVDM**	**VDM**	**FDM**	**OM**	**IWVDM**	**VDM**	**FDM**	**OM**
Accuracy	**88.89**	82.54	84.13	79.92	79.41	75	**86.76**	83.23
Sensitivity	22.22	33.33	0.00	0.00	20.00	30.00	30.00	20.00
Specificity	100.00	90.74	98.15	93.24	89.66	82.76	96.55	94.83
Pos Pred Value	100.00	37.50	0.00	0.00	25.00	23.08	60.00	40.00
Neg Pred Value	88.53	89.09	85.48	81.21	86.67	87.27	88.89	87.30
Prevalence	14.29	14.29	14.29	13.57	14.71	14.71	14.71	14.71
Detection Rate	3.18	4.76	0.00	0.00	2.94	4.41	4.41	2.94
Detection Prevalence	3.18	12.70	1.59	1.51	11.77	19.12	7.35	7.35
Balanced Accuracy	61.11	62.04	49.07	46.62	54.83	56.38	63.28	57.41

Table 38. Performance Estimates of Classification on Vote Data

Vote	**With IFS**				**Without IFS**			
For *k*=7	**IWVDM**	**VDM**	**FDM**	**OM**	**IWVDM**	**VDM**	**FDM**	**OM**
Accuracy	**96.95**	96.95	96.95	88.55	**95.42**	93.13	95.42	86.29
Sensitivity	96.51	97.67	97.67	95.35	95.35	94.19	94.19	95.35
Specificity	97.78	95.56	95.56	75.56	95.56	91.11	97.78	68.89
Pos Pred Value	98.81	97.67	97.67	88.17	97.62	95.29	98.78	85.42
Neg Pred Value	93.62	95.56	95.56	89.47	91.49	89.13	89.80	88.57
Prevalence	65.65	65.65	65.65	65.65	65.65	65.65	65.65	65.65
Detection Rate	63.36	64.12	64.12	62.60	62.60	61.83	61.83	62.60
Detection Prevalence	64.12	65.65	65.65	70.99	64.12	64.89	62.60	73.28
Balanced Accuracy	97.14	96.61	96.61	85.45	95.45	92.65	95.98	82.12

Table 39. Performance Estimates of Classification on UC Census-Income Data (Without using DC)

UC Census	**With IFS**						**Without IFS**					
For *k*=13	**IWVDM**					**Acc**	**IWVDM**					**Acc**
Class No	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**	
Sensitivity	92.4	95.9	96.4	97.0	100.0	**94.4**	97.7	69.7	45.0	27.0	31.3	74.9
Specificity	99.7	96.6	98.7	99.1	99.0		97.1	88.2	87.7	95.7	99.4	
Pos Pred Value	99.7	89.7	93.0	91.5	72.0		96.9	68.0	39.4	37.0	55.6	
Neg Pred Value	92.9	98.7	99.4	99.7	100.0		97.9	89.0	90.0	93.3	98.5	
Prevalence	50.1	23.3	15.1	9.1	2.4		47.8	26.5	15.1	8.5	2.2	
Detection Rate	46.3	22.4	14.5	8.8	2.4		46.7	18.5	6.8	2.3	0.7	
Detection Prevalence	46.4	25.0	15.6	9.6	3.4		48.2	27.1	17.2	6.2	1.2	
Balanced Accuracy	96.1	96.3	97.6	98.1	99.5		97.4	79.0	66.4	61.3	65.3	
For *k*=13	**VDM**					**Acc**	**VDM**					**Acc**
Class No	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**	
Sensitivity	92.4	95.4	95.5	94.0	100.0	93.9	96.3	96.4	91.0	95.2	87.5	89.8
Specificity	98.9	96.3	98.9	99.1	99.0		97.7	97.6	98.9	99.4	99.7	
Pos Pred Value	98.8	88.7	93.8	91.3	72.0		97.4	93.5	93.5	93.8	87.5	
Neg Pred Value	92.9	98.6	99.2	99.4	100.0		96.7	98.7	98.4	99.6	99.7	
Prevalence	50.1	23.3	15.1	9.1	2.4		47.8	26.5	15.1	8.5	2.2	
Detection Rate	46.3	22.3	14.4	8.5	2.4		46.0	25.5	13.7	8.1	1.9	
Detection Prevalence	46.8	25.1	15.3	9.4	3.4		47.2	27.3	14.7	8.7	2.2	
Balanced Accuracy	95.7	95.8	97.2	96.6	99.5		97.0	97.0	94.9	97.3	93.6	
For *k*=13	**FDM**					**Acc**	**FDM**					**Acc**
Class No	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**	
Sensitivity	91.6	95.4	92.8	95.5	100.0	93.2	94.0	89.7	86.5	73.0	31.3	88.6
Specificity	99.7	96.5	98.7	98.4	98.6		97.4	95.2	96.3	96.4	99.9	
Pos Pred Value	99.7	89.1	92.8	85.3	64.3		97.1	87.1	80.7	65.7	83.3	
Neg Pred Value	92.2	98.6	98.7	99.5	100.0		94.7	96.3	97.6	97.5	98.5	
Prevalence	50.1	23.3	15.1	9.1	2.4		47.8	26.5	15.1	8.5	2.2	
Detection Rate	45.9	22.3	14.0	8.7	2.4		44.9	23.7	13.0	6.2	0.7	
Detection Prevalence	46.0	25.0	15.1	10.2	3.8		46.3	27.3	16.2	9.5	0.8	
Balanced Accuracy	95.7	95.9	95.8	96.9	99.3		95.7	92.5	91.4	84.7	65.6	
For *k*=13	**OM**					**Acc**	**OM**					**Acc**
Class No	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**	
Sensitivity	91.6	95.9	96.4	97.0	100.0	94.0	90.9	93.3	91.0	85.7	75.0	90.8
Specificity	99.7	96.3	98.7	99.0	99.0		98.7	94.3	97.1	98.1	99.9	
Pos Pred Value	99.7	88.7	93.0	90.3	72.0		98.5	85.5	84.9	80.6	92.3	
Neg Pred Value	92.2	98.7	99.4	99.7	100.0		92.2	97.5	98.4	98.7	99.4	
Prevalence	50.1	23.3	15.1	9.1	2.4		47.8	26.5	15.1	8.5	2.2	
Detection Rate	45.9	22.4	14.5	8.8	2.4		43.4	24.7	13.7	7.3	1.6	
Detection Prevalence	46.0	25.2	15.6	9.8	3.4		44.1	28.9	16.2	9.1	1.8	
Balanced Accuracy	95.7	96.1	97.6	98.0	99.5		94.8	93.8	94.1	91.9	87.4	

Table 40. Performance Estimates of Classification on UC Census-Income Data (With using DC)

UC Census	**With IFS**						**Without IFS**					
For *k*=13	**IWVDM**					**Acc**	**IWVDM**					**Acc**
Class No	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**	
Sensitivity	73.7	36.8	100.0	100.0	100.0	**67.3**	96.5	75.7	47.0	30.6	31.8	75.9
Specificity	78.8	97.0	90.7	93.8	96.1		98.7	88.8	88.6	94.6	99.4	
Pos Pred Value	66.7	87.5	69.2	57.1	33.3		98.5	69.7	43.5	34.5	63.6	
Neg Pred Value	83.9	72.7	100.0	100.0	100.0		96.9	91.5	90.0	93.6	97.9	
Prevalence	36.5	36.5	17.3	7.7	1.9		47.5	25.3	15.7	8.5	3.0	
Detection Rate	26.9	13.5	17.3	7.7	1.9		45.8	19.2	7.4	2.6	1.0	
Detection Prevalence	40.4	15.4	25.0	13.5	5.8		46.5	27.5	17.0	7.5	1.5	
Balanced Accuracy	76.2	66.9	95.4	96.9	98.0		97.6	82.3	67.8	62.6	65.6	
For *k*=13	**VDM**					**Acc**	**VDM**					**Acc**
Class No	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**	
Sensitivity	47.4	21.1	44.4	100.0	0.0	40.4	96.5	96.8	93.0	90.3	95.5	**95.5**
Specificity	60.6	69.7	93.0	93.8	96.1		99.0	97.6	99.0	99.0	99.6	
Pos Pred Value	40.9	28.6	57.1	57.1	0.0		98.8	93.2	94.7	88.9	87.5	
Neg Pred Value	66.7	60.5	88.9	100.0	98.0		96.9	98.9	98.7	99.1	99.9	
Prevalence	36.5	36.5	17.3	7.7	1.9		47.5	25.3	15.7	8.5	3.0	
Detection Rate	17.3	7.7	7.7	7.7	0.0		45.8	24.5	14.6	7.7	2.9	
Detection Prevalence	42.3	26.9	13.5	13.5	3.8		46.4	26.3	15.5	8.6	3.3	
Balanced Accuracy	54.0	45.4	68.7	96.9	48.0		97.8	97.2	96.0	94.6	97.5	
For *k*=13	**FDM**					**Acc**	**FDM**					**Acc**
Class No	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**	
Sensitivity	47.4	42.1	100.0	100.0	100.0	59.6	93.1	94.1	80.0	77.4	13.6	87.6
Specificity	81.8	84.9	90.7	93.8	94.1		99.2	94.9	95.9	95.2	99.6	
Pos Pred Value	60.0	61.5	69.2	57.1	25.0		99.1	86.1	78.6	60.0	50.0	
Neg Pred Value	73.0	71.8	100.0	100.0	100.0		94.1	97.9	96.3	97.8	97.4	
Prevalence	36.5	36.5	17.3	7.7	1.9		47.5	25.3	15.7	8.5	3.0	
Detection Rate	17.3	15.4	17.3	7.7	1.9		44.2	23.8	12.6	6.6	0.4	
Detection Prevalence	28.9	25.0	25.0	13.5	7.7		44.6	27.6	16.0	10.9	0.8	
Balanced Accuracy	64.6	63.5	95.4	96.9	97.1		96.2	94.5	88.0	86.3	56.6	
For *k*=13	**OM**					**Acc**	**OM**					**Acc**
Class No	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5**	
Sensitivity	63.2	10.5	88.9	100.0	0.0	50.0	91.1	92.4	92.2	83.9	95.5	91.1
Specificity	60.6	90.9	90.7	91.7	96.1		98.2	94.0	97.7	98.5	99.9	
Pos Pred Value	48.0	40.0	66.7	50.0	0.0		97.8	83.8	88.3	83.9	95.5	
Neg Pred Value	74.1	63.8	97.5	100.0	98.0		92.4	97.3	98.5	98.5	99.9	
Prevalence	36.5	36.5	17.3	7.7	1.9		47.5	25.3	15.7	8.5	3.0	
Detection Rate	23.1	3.8	15.4	7.7	0.0		43.2	23.4	14.5	7.1	2.9	
Detection Prevalence	48.1	9.6	23.1	15.4	3.8		44.2	27.9	16.4	8.5	3.0	
Balanced Accuracy	61.9	50.7	89.8	95.8	48.0		94.6	93.2	95.0	91.2	97.7	

APPENDIX D

DEFINITIONS OF PERFORMANCE MEASURES

Confusion matrix for a binary classification problem is given in Table 41.

Table 41. Confusion Matrix

Predicted Class	**True Class**	
	Event	No Event
Event	TP	FP
No Event	FN	TN

where TP: True Positive,

FP: False Positive,

FN: False Negative,

TN: True Negative.

Sensitivity	= TP / (TP + FN),
Specificity	= TN / (FP + TN),
Prevalence	= (TP + FN) / (TP + FP + FN + TN),
Pos.Pred.Value	= TP / (TP + FP),
Neg.Pred.Value	= TN / (FN + TN),
Detection Rate	= TP / (TP + FP + FN + TN),
Detection Prevalence	= (TP + FP) / (TP + FP + FN + TN),
Balanced Accuracy	= (Sensitivity + Specificity) / 2.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name	: Ortakaya, Ahmet Fatih
Nationality	: Turkish (TC)
Date and Place of Birth	: 21.04.1982, Ankara
Marital Status	: Married
Phone	: 0505 918 09 49

EDUCATION

Degree	**Institution**	**Year of Graduation**
M.Sc.	Statistics - METU	June 2009
B. Sc.	Statistics - METU	June 2005
High School	Aydınlıkevler Lisesi	June 2000

PROFESSIONAL EXPERIENCE

Year	**Place**	**Enrollment**
April 2009-June 2013/ July 2014-March 2015/ January 2016 - Present	Ministry of Family and Social Policy	Social Policy Expert
March 2015-December 2015	Ministry of Family and Social Policy	Consultant to the Minister of FSP
June 2013-July 2014	Undersecretariat of Public Order and Security	Expert
March 2006 – April 2009	General Directorate of Social Assistance and Solidarity	Junior Social Policy Expert

PUBLICATIONS

Ortakaya, A. F. & Yozgatlıgil T. C. (2013). Multivariate Time Series Modeling of the Number of Applicants for Conditional Cash Transfer Program in Turkey, *Journal of Social Policies, 7*(30), 101-116.

Coşkun, S., Güneş, S., & Ortakaya, A. F. (2011). A Proposal for a Minimum Income Support Program for Turkey (in Turkish), *Journal of Faculty of Economic and Administrative Science*, 13(3), 1-30, Gazi University.