DATA MINING METHODS FOR CLUSTERING POWER QUALITY DATA COLLECTED VIA MONITORING SYSTEMS INSTALLED ON THE ELECTRICITY NETWORK

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

BY

MENNAN GÜDER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2009

Approval of the thesis:

DATA MINING METHODS FOR CLUSTERING POWER QUALITY DATA COLLECTED VIA MONITORING SYSTEMS INSTALLED ON THE ELECTRICITY NETWORK

submitted by **MENNAN GÜDER** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Müslim Bozyiğit _____
Head of Department, **Computer Engineering**

Assoc. Prof. Dr. Nihan Kesim Çiçekli _____
Supervisor, **Computer Engineering Dept., METU**

Prof. Dr. Işık Çadırcı _____
Co-Supervisor, **Electrical and Electronics Eng. Dept., Hacettepe University**

Examining Committee Members:

Assoc.Prof.Dr. Halit Oğuztüzün _____
Computer Engineering Dept., METU

Assoc. Prof. Dr. Nihan Kesim Çiçekli _____
Computer Engineering Dept., METU

Assoc.Prof.Dr. Ali Doğru _____
Computer Engineering Dept., METU

Asst. Prof. Dr. Tolga Can _____
Computer Engineering Dept., METU

Dr. Özgül Salor _____
TÜBİTAK

Date: 10 / 09 / 2009

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: Mennan Güder

Signature

ABSTRACT

DATA MINING METHODS FOR CLUSTERING POWER QUALITY DATA COLLECTED VIA MONITORING SYSTEMS INSTALLED ON THE ELECTRICITY NETWORK

Güder, Mennan

M.S., Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Nihan Kesim Çiçekli

Co-Supervisor: Prof. Dr. Işık Çadırcı

September 2009, 70 pages

Increasing power demand and wide use of high technology power electronic devices result in need for power quality monitoring. The quality of electric power in both transmission and distribution systems should be analyzed in order to sustain power system reliability and continuity. This analysis is possible by examination of data collected by power quality monitoring systems. In order to define the characteristics of the power system and reveal the relations between the power quality events, huge amount of data should be processed. In this thesis, clustering methods for power quality events are developed using exclusive and overlapping clustering models. The methods are designed to cluster huge amount of power quality data which is obtained from the online monitoring of the Turkish Electricity Transmission System. The main issues considered in the design of the clustering methods are the amount of the data, efficiency of the designed algorithm and queries that should be supplied to the domain experts. This research work is fully supported by the Public Research grant Committee (KAMAG) of TUBITAK within the scope of National Power quality Project (105G129).

Keywords: Power Quality, Power Quality Event, K-Means Clustering, Fuzzy C-Means Clustering, Data Mining, Monitoring System

ÖZ

ELEKTRİK SİSTEMİ ÜZERİNE KURULAN İZLEME SİSTEMLERİ TARAFINDAN TOPLANAN GÜÇ KALİTESİ OLAYLARI İÇİN VERİ MADENCİLİĞİ TEKNİKLERİ

Güder, Mennan

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi: Doç. Dr. Nihan Kesim Çiçekli

Ortak Tez Yöneticisi: Prof. Dr. Işık Çadırcı

Eylül 2009, 70 sayfa

Enerji talebi ve yüksek teknoloji elektronik aletlerin kullanımındaki artış elektriksel güç kalitesinin izlenmesini zorunlu kıldı. Sistemin devamlılığını ve güvenilirliğini sağlamak için hem dağıtım ve hem iletim sistemlerindeki elektriksel güç kalitesinin analiz edilmesi gerekmektedir. Bu analiz güç kalitesi izleme sistemleri ile toplanan verilerin incelenmesi ile yapılabilir. Güç sisteminin karakteristiklerini tanımlamak ve güç kalitesi olayları arasındaki ilişkileri tanımlayabilmek için büyük miktardaki verinin incelenmesi gerekmektedir. Bu tezde, ayrık ve çakışan sınıflandırma teknikleri kullanılarak, güç kalitesi olayları için sınıflandırma yöntemleri geliştirilmiştir. Yöntemler Türkiye Elektrik Sisteminin çevrimiçi izlenmesinden elde edilen büyük miktardaki güç kalitesi olaylarını sınıflandırmak için geliştirildiler. Veri miktarı, algoritmanın verimliliği ve alan uzmanlarının sistemden sorgulayabilecekleri veriler, sistem tasarlanırken üzerinde durulan en önemli konulardı. Bu araştırma çalışması, 105G129 kodlu Güç Kalitesi Milli Projesi kapsamında TÜBİTAK Kamu Araştırmaları Destekleme Programı (KAMAG) tarafından desteklenmektedir.

Anahtar Sözcükler : Güç Kalitesi, Güç Kalitesi Olayı, K-Means Sınıflandırma, Fuzzy C-Means Sınıflandırma, Veri Madenciliği, İzleme Sistemi

To My Family

ACKNOWLEDGMENTS

I express my sincerest thanks and my deepest respect to my supervisor, Assoc. Prof. Dr. Nihan Kesim Çiçekli, for her guidance, technical and mental support, encouragement and valuable contributions during my graduate studies.

I would like to thank to my co-supervisor, Prof. Dr. Işık Çadırcı, for her guidance, support and patience during my graduate studies.

I express my sincerest thanks to Dr. Özgül Salor, for her boundless knowledge transfer, guidance, support and encouragement during my graduate studies.

Special appreciation goes to Dilek Küçük, Tolga İnan and Serkan Pakhuylu for sharing their knowledge and valuable times with me during my studies.

Special thanks to National Power Quality Project of Turkey (Project No: 105G129) and TÜBİTAK UZAY for support.

Thanks to TÜBİTAK Bilim İnsanı Destekleme Programı for their support.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES

LIST OF TABLES

TABLES

CHAPTER 1

INTRODUCTION

Electrical power quality (PQ) of an electricity network is defined according to the results of the measurements of various PQ parameters, such as frequency, flicker, and harmonics. These parameters provide an insight on the quality of power supplied from that network. As a result of the increasing power demand and use of costly investments on delicate power electronic devices which are prone to be affected by the PQ, PQ monitoring has gained importance in the last decades. Monitoring PQ supplies indicative and representative data for the electricity network, and also enables early warning decisions and makes the analysis of the power quality of the whole electricity network possible. After the characteristic data about the system is obtained, the need for automatic pattern and information extraction is inevitable for detecting the problems and taking the necessary countermeasures. Several PQ monitoring systems have been developed in order to minimize the effects of PQ problems on the electronic equipments used in both industry and house loads, where PQ problem is defined as deviations from nominal power system behavior. However, the domain of PQ monitoring still suffers from the lack of a knowledge extraction system for a long period and from a wide geographical range of data.

Data mining is the most promising tool to transform PQ data into information by revealing intrinsic groupings and interesting relations in the unlabeled PQ data set. Enhancements of the electricity system and reliability decisions may be based on the results of the applied data mining methods. Domain experts can use the data mining results in the electrical decision making process to define the problems and their possible causes; and to propose solutions to the identified problems.

In order to improve the PQ in energy generation, transmission and distribution systems, real-time and long-period data have to be investigated. A widespread and long term PQ monitoring is required to make such data available. A corresponding system has been developed as a subproject of the National Power Quality Project by the Power Quality Department of TÜBİTAK (Turkish Scientific Technical Research Organization) [13]. The developed PQ monitoring system aims to monitor the PQ parameters in transformer substations from all over the country, handle the required PQ measurements and analyze the PQ archive with the use of domain specific data mining techniques. The National Power Quality Project enables users to access two types of data; daily average data and event data. The main concern of the thesis is the event data, which contains detected voltage sags, swells and interruptions. The amount of available data is huge, the current number of events is 1.060.000 which is calculated from (6776 events per week)*(52week per year)*(3 Years), where each event is defined as 3.5MB of raw data. The 10 minutes of the remaining PQ parameters such as, frequency, harmonics, inter-harmonics and flicker are stored in the PQ database. The overall structure of the system is given in the third chapter. The aim of this thesis is to fulfill the PQ monitoring system requirements by enabling data mining methods for the PQ domain.

The motivation of the thesis is based on the analysis requirement of the data collected by the National PQ Monitoring System. As the amount of data and the number of features of the data items increase, the experts are no longer able to make analysis on the entire data. The amount, structure and properties of the data make

also the available data mining tools and algorithms impossible to be applied directly. The algorithm dealing with the power quality data should take the amount of data, characteristics of the system and features of the data items into consideration. The data amount is crucial because of the memory bound problems. The characteristics of the system and the features of the data items are significant in the design of the domain specific calculations and measurements. From power quality domain perspective there are analysis requirements for defining characteristics of transformer substations, revealing relations between power quality events, modeling event distributions on time and location, load foresting and predicting electricity system behavior. In order to fulfill these requirements, the power quality data should be examined via data mining models. The most promising method is clustering since the data is unlabeled and the requirements are based on revealing the distribution details and inter relations between events. The best fit clustering algorithms to the domain are the ones which enable modifications for domain specific calculation details, supply graphically representable results for further analysis and manage the memory bound problem. The algorithm should also not contain complex calculations on the high dimension feature vectors of the power quality data items. The power quality events may trigger each other, so when the events are clustered the resulting clusters are expected to be globular.

In order to fulfill the analysis requirements of the power quality data, two clustering methods are proposed. The first proposed method designs a k-means based algorithm. K-means clustering algorithm that employs Lloyd's algorithm [15] is selected to be modified according to the requirements of the power quality domain and to be applied on the power quality events data. Lloyd's algorithm is an unsupervised and iterative clustering algorithm. When the data items are expected to be clustered into globular sets, Lloyd's algorithm may produce tighter clusters than the other clustering algorithms. The Lloyd's algorithm is also fast and simple compared to the other clustering methods. Simplicity of the algorithm makes it easily modifiable. The k-means algorithm is a hard clustering example, the data items

should be assigned to a definite cluster. However the membership degrees of a data item for all clusters may supply detailed information for characterizing the occurring locations of the data item. As a result of the requirement, fuzzy c-means algorithm is selected as the base of the second clustering model.

The author of this thesis has worked on the development of the PQ monitoring and the PQ mining system. The work includes developing a monitor for collecting the required power quality parameters, implementing calculations carried on the collected data, storing the calculated parameters on a relational database and developing clustering models for PQ domain.

The rest of the thesis is organized as follows. Chapter 2 summarizes the literature about power quality data mining and related concept definitions. Chapter 3 describes the National Power Quality System. Chapter 4 defines the details of the proposed k-means based method; its design and application results. The proposed k-means method focuses on the root-mean-square (rms) values of the voltage waveform for clustering the power quality events. The clustered units are 3-second rms voltages of the event data of three phases forming a (3x300) matrix. Clustering is achieved on the rms event data matrices. Chapter 5 defines the details of the proposed fuzzy c-means based method. In this method, the power quality events are clustered by chunk based fuzzy c-means algorithm and the membership values for each event and cluster pair are examined for discovering the relations between the variables in the power quality data items. Both of the proposed methods aim understanding the characteristics of the monitored electricity network. These methods enable both revealing interactions between the events and defining their causes, locations and consequences. Methods also aim at categorizing the transformer substations according to the characteristics of the events occurred at that substation. The final chapter summarizes the contributions of the thesis, provides the concluding remarks, and recommends future work.

CHAPTER 2

RELATED WORK

The increase in the power consumption results in complex electricity networks. The state of electricity systems can be defined as constant flux because of the continuous increase in its size and complexity. As a result of this growth, examining the electrical infrastructure of the system becomes crucial in order to sustain reliability of the system. Long term, widespread and accurate power quality parameter data is required to examine the infrastructure of the focused electricity network. The capability of the power quality monitoring systems has evolved with the improvements in storage, digital signal processing and computation capabilities of the computers. Power, frequency, harmonics, inter-harmonics, flicker and power quality events are the main power quality parameters that are used to define the infrastructure, problems and characteristics of the monitored electricity network. Events are used to define the instant problems occurred in the system, the remaining parameters are also results of the continuous monitoring of the system.

The power quality monitoring has evolved from just trying to sustain the system to improving the power quality. The long period and widespread data may be used to predict the future behavior of the system. The data may also be used to enforce the customers to obey the legal standards by the authority. Knowledge extraction

methodologies should be applied to the collected data in order to achieve these purposes. In the following sections of the chapter, related data mining concepts and power quality data mining models in the literature are described.

2.1 OVERVIEW OF RELATED DATA MINING METHODS

Wide availability of huge amount of high dimension and complex data and analysis requirements for defining characteristics of a domain, make use of data mining methods inevitable. Data mining is extracting useful and unknown knowledge from data which cannot be analyzed by direct observations of domain experts because of the complexity and amount. The main knowledge extraction steps are data cleaning, integration, data selection, data transformation, mining, pattern evaluation and knowledge representation. Data cleaning is defined as the process of eliminating incomplete, noisy and inconsistent data. Integration includes data and schema integration steps. Data selection is based on the binning and regression analysis. Data transformation is handled via smoothing, aggregation, generalization and normalization algorithms.

Classification, clustering, association rule mining and frequent pattern mining are the main data mining concepts. Performance, pattern evaluation, diversity, handling noise and handling incomplete data are the major issues to be considered in each data mining concept. Efficiency and scalability are the most important performance factors for data mining algorithms. Parallel and distributed processing capabilities of the data mining algorithms are the other criteria for the evaluation. Well designed data mining algorithm should be able to handle relational and complex data from heterogeneous databases and different information systems. The mining algorithm should also be able to use domain information to guide the discovery process. Discovered knowledge should be expressed visually in order to enable the domain experts to analyze the results effectively.

Classification is the process of modeling predefined data groups. According to the model developed by the classification results, the data items are assigned into the corresponding classes. Decision trees, neural networks, k-nearest neighbour classification, support vector machines and Bayesian classification are the main methods for constructing classification models [28]. The classification models require data to be labeled. In the power quality event data case, the event classes and their properties are not obviously defined which makes classification models not applicable without preprocessing the data.

Clustering power quality events is the focus of the thesis, where clustering is the process of grouping data items according to the domain and data feature specific similarity measures. Clustering algorithms try to maximize both the similarity between the data items in the same cluster and the distance between the data items in different clusters. The success criteria for the clustering algorithms are scalability, ability to deal with different types of attributes, ability to discover clusters with arbitrary shape, requiring minimum domain knowledge, ability to deal with noisy data, incremental clustering ability, being insensitive to the order of input records, ability to handle high dimension data, ability to take constraints into consideration, interoperability and usability. The general memory based clustering algorithms deals with two main data structures; data matrix and dissimilarity matrix. Interval scaled, binary, categorical variables and vector objects are the most observed variable types to be considered in the distance measure calculations of dissimilarity matrix. In power quality domain, the voltage levels are categorical values, event rms values are vectors of numeric variables. The features stored in rms vectors are interval scaled standardized variables, which are formed via continuous measure of a scale.

The most important clustering techniques are partitioning, hierarchical, density based and constraint based methods [28]. Partitioning methods partitions the data set into selected number of initial clusters and iteratively changes cluster assignments to improve the partitioning. Partitioning based methods are well suited for clustering

medium sized databases however in order to handle huge amount of data, the methods should be modified. Main partitioning algorithms are k-medoids, k-means and fuzzy c-means. The k-medoids algorithm represents the cluster centroid with the data item that is nearest to the center of the cluster. K-means algorithm is the centroid based of partitioning method, in which each cluster center is modeled by the mean of the feature vectors of the data items in that cluster. The algorithm partitions data set of n items into k clusters by minimizing inter cluster similarity and maximizing intra cluster similarity. Fuzzy c-means algorithm is another partitioning algorithm in which each data item has fuzzy membership labels for the belonging degrees to the clusters.

Hierarchical clustering methods are based on bottom-up or top-down hierarchical decomposition of the data set. Bottom-up approach merges the data items and forms the final clusters on the other hand top-down approach divides the data set into sub-clusters and obtains final clusters. Both hierarchical clustering approaches use linkage analysis in the merge and split steps. There are three types of linkage calculations; complete linkage, single linkage and average linkage. Complete linkage is based on maximum pairwise distance, single linkage is based on difference between two closest members and average linkage is based on the average pairwise distance. Bottleneck of the hierarchical clustering algorithms is rigid split and merge steps, which may decrease the computations but decrease flexibility of the algorithms.

Partition based methods could only form spherical-shaped clusters. Density based clustering methods propose a solution to that limitation. Each cluster is formed by adding data items until the density achieves defined threshold value in density based clustering. Constraint based clustering models are based on the user specified or application oriented constraints where constraints include specifications for the final clusters.

The methods given in the thesis are based on the k-means and fuzzy c-means algorithms. Both of these algorithms are based on the minimum sum of squared distance clustering problem. There are many performance optimization studies for these algorithms. The selection of the initial cluster center may have a big impact on the accuracy and performance levels of both k-means and fuzzy c-means clustering. Barakbah and Helen [29] propose an approach to optimize the initial centroid selection. The algorithm is based on spreading the initial centroids in the feature space and maximizing the distance between them as far as possible. The method first computes the center of all data set, c_{center} and finds the nearest data to the center, $c_{nearest}$. Then it selects the next nearest data c_{next} by satisfying the following formula $d(c_{nearest}, c_{next}) \geq d(c_{nearest}, c_{center})$. This optimization also covers noisy data since outliers would be far from the mean [29].

K-means++ is a modified version of the classical k-means algorithm. It is proposed as an accuracy optimization to k-means clustering algorithm. K-means++ is the combination of an initial cluster centroid assignment step and k-means algorithm. The algorithm starts with selecting one random cluster center. After that, it iterates through selecting another cluster center according to D(x) calculations. D(x) is the distance between data item x and the closest cluster center selected so far. When the cluster selection process is over, the k-means algorithm is applied on the data set with the selected cluster centers [30]. The k-means++ algorithm is implemented in the k-means based event clustering method proposed in Chapter 4.

Kolen and Hutcheson [23] propose a method for reducing time complexity of the fuzzy c-means algorithm. Main computations in fuzzy c-means algorithm are cluster centroid and membership matrix calculations. In [23], the membership matrix is eliminated and the two separate computations are combined into one step. The effects of this optimization are observed in both data access and runtime. The optimization method is implemented in the fuzzy c-means based power quality event clustering method proposed in Chapter 5.

The memory usage optimization is another crucial concept in clustering since the data amount is generally huge. Data structures that consume less space, like Trie which is an ordered tree data structure used to store an associative array, are used in order to minimize memory requirements of clustering algorithms.

2.2 POWER QUALITY DATA MINING MODELS IN THE LITERATURE

The measurement of power quality parameters are based on IEC-61000-4-30 [11] standard. IEC-61000-4-30 standard defined the methods for measurement and interpretation of results for power quality parameters in 50/60 Hz A.C. power supply systems. The standard is a performance specification for the collected and computed power quality data collected via monitoring systems [11]. The knowledge extraction methods that have been proposed for examining the PQ data are classification, clustering, and fuzzy neural network for rule generation [1-3]. Generally, the power quality data mining research uses data mining tools on a restricted number of power quality parameters. For example in [1], automatic clustering is applied on the 3^{rd}, 5^{th} and 7^{th} harmonics data collected from three year simultaneous measurements of eight sites in a transformer substation. SNOB [24] and AutoClass [25] data mining tools are used to cluster the collected data, where SNOB implements unsupervised learning using minimum message length principle and AutoClass implements Bayesian classification. The resulting classes of the data are not strict, the membership is a probability value, and partial membership is also possible in the research. The resulting clusters are examined by the utility engineers and the reasons, properties and characteristics of the clusters are decided according to the domain knowledge [1].

Another power quality data mining research is carried on the voltage raw data. In [2], a data mining and knowledge discovery approach is mentioned. The research is carried on a one year period voltage raw data. First data processing by using a phase

corrected wavelet transform is applied to extract relevant features. After that step the features and if then ruled fuzzy neural classifier are used to classify the short duration transient PQ disturbance patterns. Fuzzy multilayered perception is used to determine the class membership values of the input patterns. The trained fuzzy neural network is also used for rule generation. The defined rules describe the association between the input feature, satisfying criteria and the class assignment [2].

Load forecasting is an important purpose for the data mining research on power quality domain. The research in [3], uses ACPro clustering software in order to build predicting models for load forecasting and to discover the relationships between the input and output variables. The data that research is applied includes 10 minute voltage and current readings of the fundamental, THD, and 3^{rd}, 5^{th} and 7^{th} harmonics over a period of two weeks for each of eight sites. ACPro is used to identify the classes in the harmonic data, where ACPro implements an unsupervised clustering with Gaussian mixture model. After the classes are identified link analysis is applied to merge the obtained clusters into super-groups. The C5.0 algorithm, a decision tree generation algorithm, is used to define the rules behind each super-group. Clementine is the data mining software used to utilize the required algorithms. The links between the clusters are visualized using KNOT. An appriori algorithm of association rules in Clementine is applied to categorize the variables at different sites for interrelated super-groups.

The other researches are based on signal processing techniques [4-10]. In [4], covariance behavior of several features derived from the event data is used for PQ event detection and classification. Classification of PQ events such as harmonics, sags, and capacitor switching is achieved using time-frequency analysis of the voltage and current waveforms in [5]. Neural networks have been used by [6] for PQ disturbance classification, while fuzzy-expert systems are used by [7] for the same purpose. Wavelets are used in [8-10] for PQ event classification. In these types of

systems [4-10], PQ events are characterized by several features and these features are classified for single phase voltages.

The power quality data mining research is generally based on the 10 minute average power quality monitoring of a restricted area for a limited period of time. The research has a deficit for the examining event data. Power quality events may cause shut down of processes run by electronics devices. Therefore it is important to detect and classify PQ events occurring on a specific site to take countermeasures against the potential PQ problems. Data mining methodologies on the power quality event data may be used to identify the correlations between the events, sites and transformer substations. The cause and location of any event may also be identified with the use of collected data. The resulting knowledge may be used to avoid the problem in the future.

CHAPTER 3

THE NATIONAL POWER QUALITY MONITORING SYSTEM

In order to define characteristics of an electricity network, the network should be monitored for a long period of time and a comprehensive examination should be carried on the monitoring results. PQ monitoring system that is able to fulfill the monitoring, computation, storage and analysis requirements, is crucial to enable characteristic definition of a network. Well designed PQ monitoring systems should supply high sampling rates, backup facilities, reliable network connections and high computational power. In this chapter, a nation-wide power quality monitoring system is described. The system has sensor computation, storage and analysis capabilities.

The National PQ monitoring system is being developed as a subproject of National Power Quality Project by the Power Quality Department of TÜBİTAK. The system consists of three main units: National PQ monitors, National Power Quality Monitoring Center and client software. The system architecture of the National PQ Monitoring System is given in Figure 3.1.

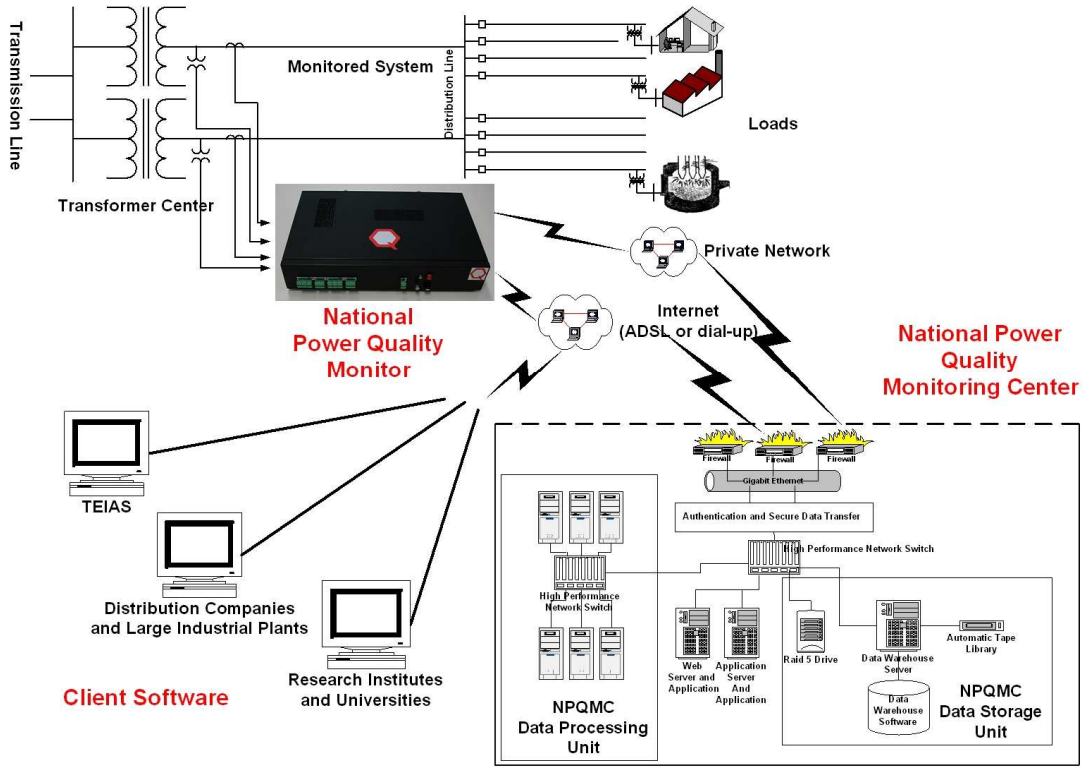


Figure 3. 1 System Architecture of the National PQ Monitoring System

3.1 NATIONAL POWER QUALITY MONITOR

PQ monitors are installed on the transformer substations to continuously monitor the required PQ parameters. 154 National PQ monitors have been installed on the 70 critical transformer substations since March 2006. The distribution of the monitored transformer substations over the country is displayed in Figure 3.2. The number of the monitors will be increased up to 2000 in order to cover the entire electricity network of Turkey.

The National PQ monitor has measurement, error detection, error logging and system monitoring functionalities. The most important property of the monitor for this thesis is the event logger capability. As an event logger, National PQ monitor measures

events (sags, swells and interruptions) and creates event packages, which include 3-second raw data, event time and location of all voltage and current channels. A new event can be detected every three seconds. Event packages are saved to the hard disc of the monitor and sent to a central database via internet connection as soon as event is detected.

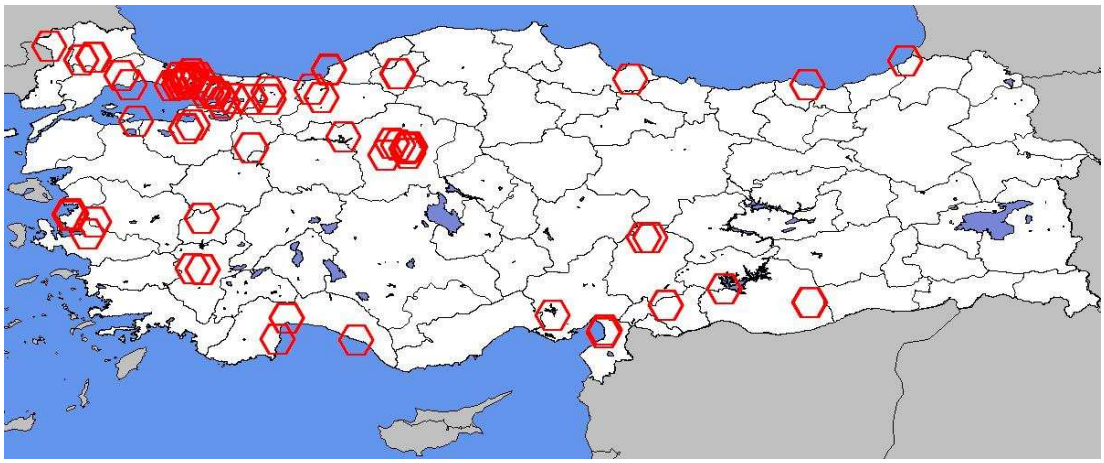


Figure 3. 2 Distribution of the Monitored Transformer Substations

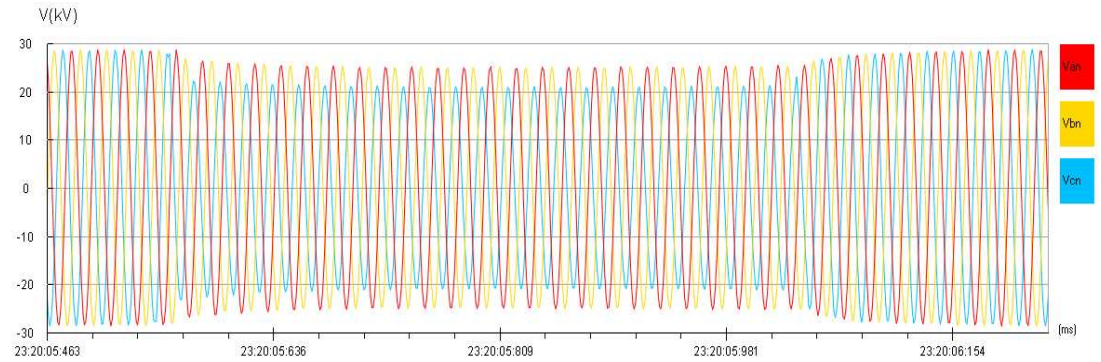
Main components of the National PQ Monitor are the data acquisition unit, processing software and the user interface. The first component of the National Power Quality Monitor is the data acquisition unit which is an analog to digital converter. It samples six voltage and six current waveforms from three-phases of two feeders at each transformer substation. The sampling rate of the unit is 25.6 kHz per channel.

The second component of the National Power Quality Monitor, Data Processing Software contains signal analysis and PQ parameter computation units. Signal analysis software buffers 12 channels of voltage and current data every three seconds. The functions of the parameter computation unit are computing 10 minute average of the PQ parameters and running the defined algorithms. It is also able to

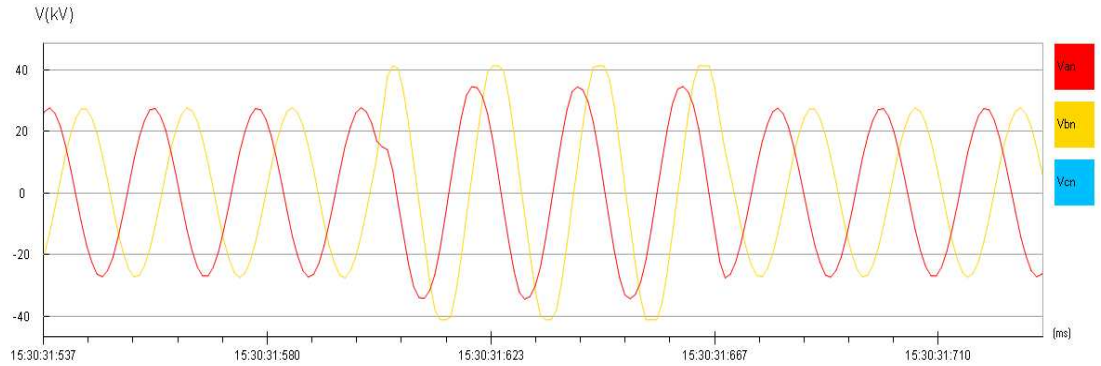
detect and label the anomalies in the system. During these computations of PQ parameters, IEC-61000-4-30 [11] standard is used. The 10 minute averages of the PQ parameters are sent to the National Power Quality Monitoring Center (NPQMC) every 24 hours. The definitions of the main power quality parameters that are used to quantify the PQ are given below:

- **Events**: Events are the main concern of the data mining methodologies described in the third chapter. The main event labels are sag, swell, unbalance, and interruption.
  - **Voltage sag** is label of the under-voltage case
  - **Voltage swell** is the label of the over-voltage case,
  - **Unbalance** is the measure of how much the amplitude and angles of the three phases differ from each other.
  - **Interruption** is an under-voltage situation where the voltage level is very close to zero signal level.

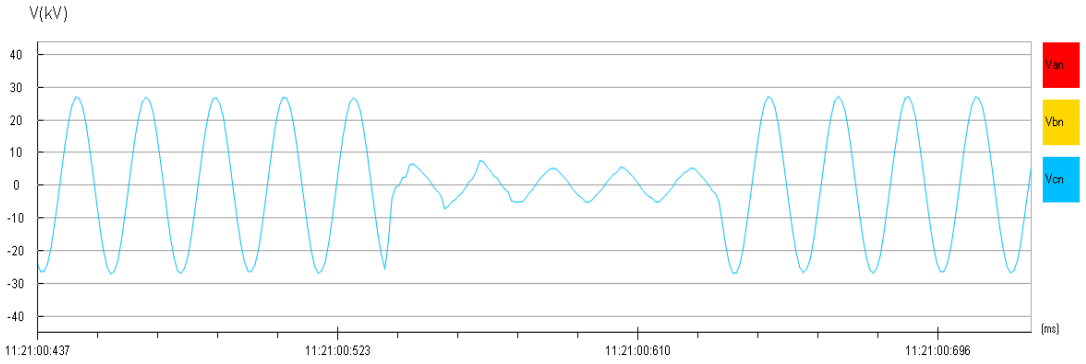
  Example of voltage sag, voltage swell and voltage interruption are given in Figure 3.3.
- **Power**: Electric power is defined as the amount of work done by an electric current, or the rate at which electrical energy is transferred. The power parameter includes power factor, active, reactive and apparent power components.
- **Frequency**: Frequency is the fundamental frequency component measured by dividing the number of integral cycles counted during one second time clock interval by the cumulative duration of the integer cycles.
- **Harmonics**: Harmonic frequencies are the multiples of the fundamental frequency existing in the frequency spectrum of the waveform.
- **Inter-harmonics**: Inter-harmonics are the frequencies existing between the multiples of the fundamental frequency.
- **Flicker**: Flicker is the fluctuation of the voltage signal envelope at frequencies from 0.5Hz to 30Hz.



(a) Voltage versus Time Graph of Typical Voltage Sag of three phases



(b) Voltage versus Time Graph of Typical Voltage Swell of two phases



(c) Voltage versus Time Graph of Typical Voltage Interruption of one phase

Figure 3. 3 Voltage versus Time Graph for Event Types

Power, frequency, harmonics, inter-harmonics and flicker are the 10 minutes average data that are sent to the center daily. In addition to the daily average data, PQ events such as voltage sags, swells and interruptions are detected at every half cycle as defined in the IEC-61000-4-30 [11] standard. 3-sec raw data of the events is send to the NPQMC as soon as the event is detected. The 3 second period contains 0.5 seconds before and 2.5 seconds after the event start. The starts of the events are determined according to the result of supply voltage levels and comparison of the thresholds where the thresholds are reconfigurable parameters. The event information contained in the 3-second raw data is given in Figure 3.4. This raw data is the main concern of the data mining methods described in the thesis, because it contains all the required information for an individual event.

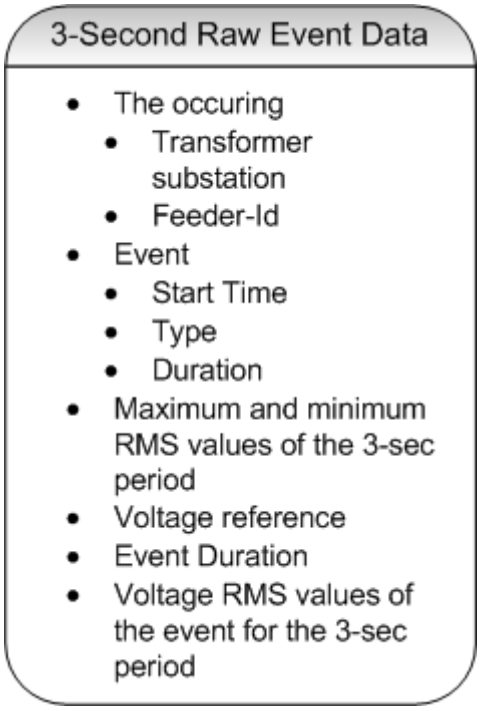


Figure 3. 4 Event Information Contained in the 3-Second Raw Data

The third component of the National PQ Monitor is the User Interface. Current and voltage waveforms at the connected feeders can be monitored instantaneously by

means of this user interface. The Interface has the following modules: Configuration, Event Log, Waveform, Harmonics and Power. Configuration module enables users to view and change the configuration parameters of the monitors. Waveform module enables users to view the 6 voltage and 6 current channels. The harmonic module shows the voltage and current channel harmonic distributions of the monitored buffer. Event log module displays the events occurred in the last three days. Power module enables users to monitor power factor, active, reactive and apparent power values of the connected feeder instantaneously.

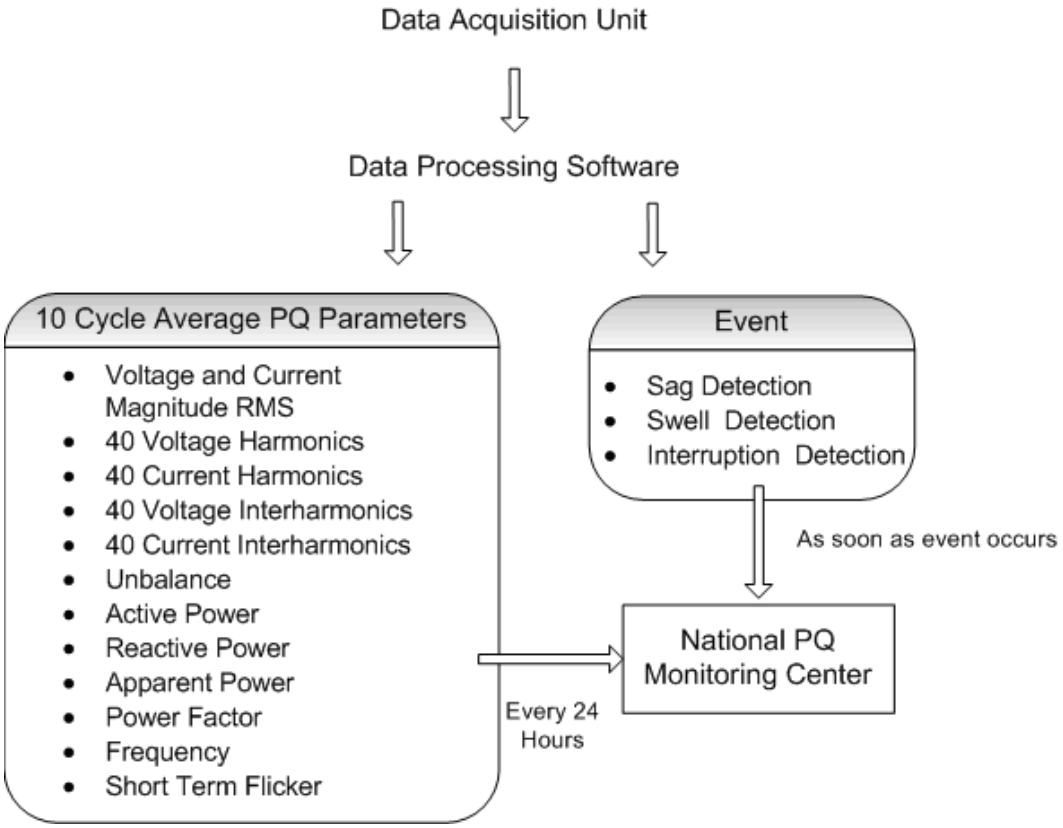


Figure 3. 5 Data Flow of the National PQ Monitoring System

Uninterruptible power supply (UPS) component of National PQ monitor is required in order to prevent the data losses in case of power outage. Each power quality

monitor has its own data storage to be used for saving computation results and for storing the backup data. Whenever there is a problem to send the data to the NPQMC, the data for that period is kept in the internal data storage of the corresponding monitor until the communication problem is solved. When the connection is reestablished, the stored data is started to be sent to the NPQMC. The data flow of the national power quality monitoring system is given in Figure 3.5. The communication between the monitors and the center is IP based ADSL over telephone line. Each monitor also has a time-synchronizer component; as a result the data stored on the NPQMC contains time and geographical location information.

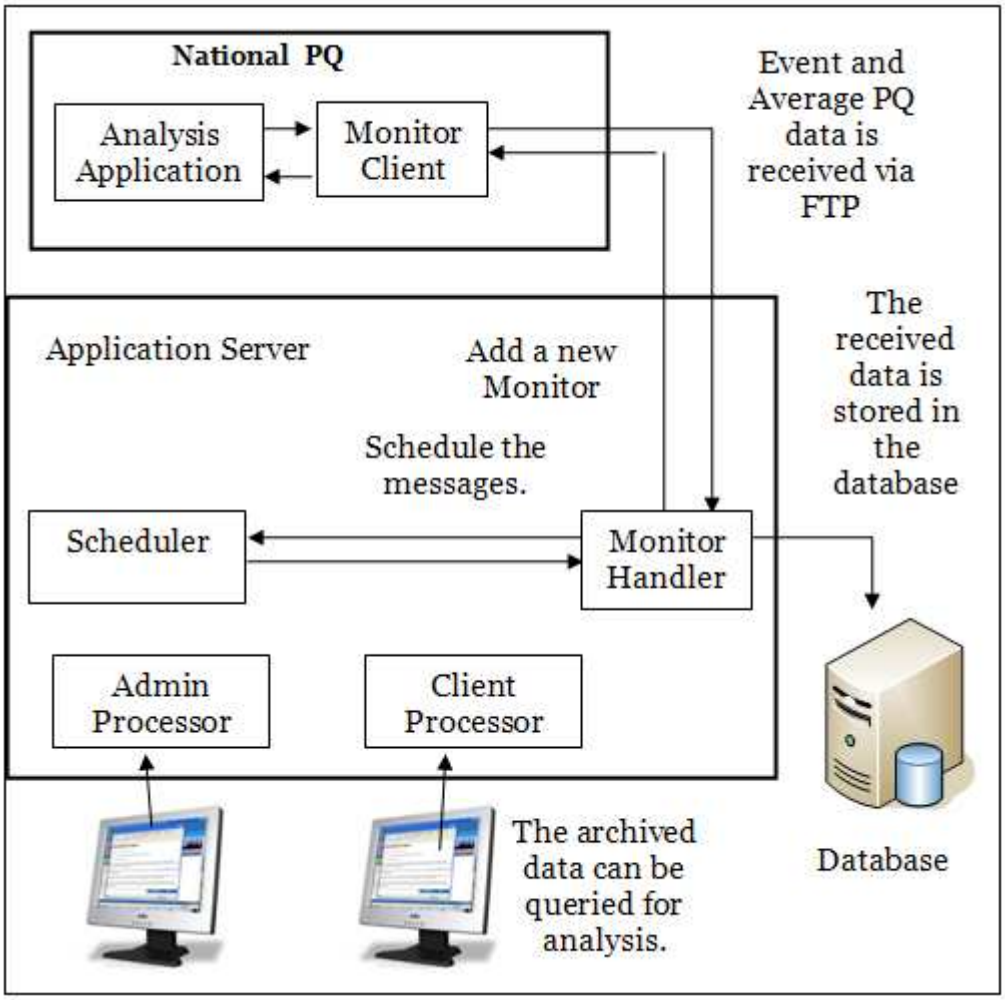


Figure 3. 6 Relations between the Components of the System and NPQMC

3.2 NATIONAL POWER QUALITY MONITORING CENTER (NPQMC)

NPQMC has the following capabilities: communication, data storage, and analysis. Relations between the Components of the System and NPQMC are given in Figure 3.6. Monitor Handler module receives the daily PQ parameter and event data from all connected National PQ monitors. Monitor Handler module is responsible for all communication between the monitors, database and the scheduler. The received data is stored in the database. PostgreSQL [27] is used as the database management system. The client processor module communicates with the clients and handles their analysis and search queries. Admin processor is responsible for the authorization level assignments. The modules of the application server described above are implemented in C programming language.

3.3 PQ DATA MINING AND VISUALIZATION SOFTWARE

The methodologies developed in this research are implemented in the PQ data mining and visualization software in order to utilize the visualization of the results. The PQ data mining and visualization software also enables users to run the proposed methodologies under different configurations. The user interface of the software is given in Figure 3.7. The interface enables users to select the methodology to be applied and time interval, voltage level and the transformer substations to be examined. The methodology specific parameters are also collected by the software. The overview of the monitoring system can also be examined from the application. The monitored transformer substations are displayed on the country map and the feeders in the monitored transformer substations are listed in the tree view.

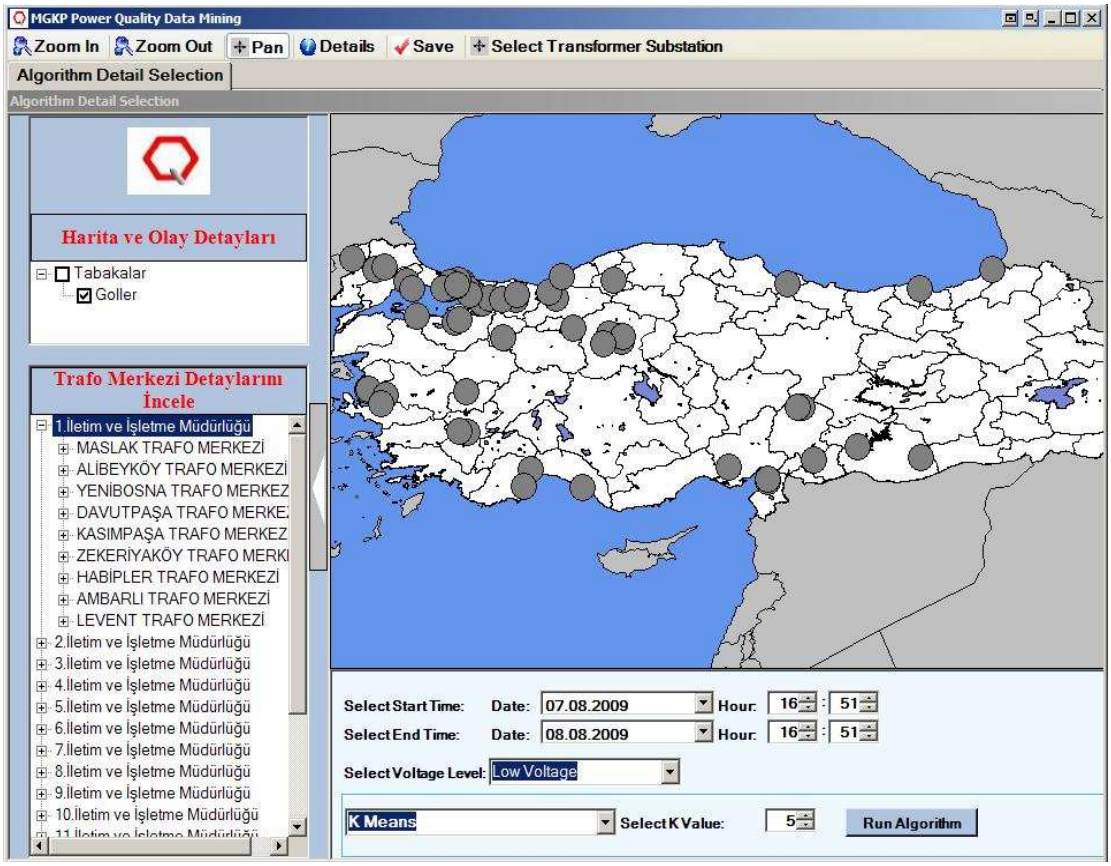


Figure 3. 7 PQ Data Mining and Visualization Software

In order to get secure access to PQ parameters, PQ Parameter Analysis software is developed. Users could insert search and analysis queries on the PQ database. Both the event data and the 10 minutes average PQ parameters can be examined by using the software. Experts examine the data mining methodologies described in this research with PQ Parameter Analysis software in order to identify the problems and the characteristics. The event analysis user interface of the Parameter Analysis software is given in Figure 3.8.

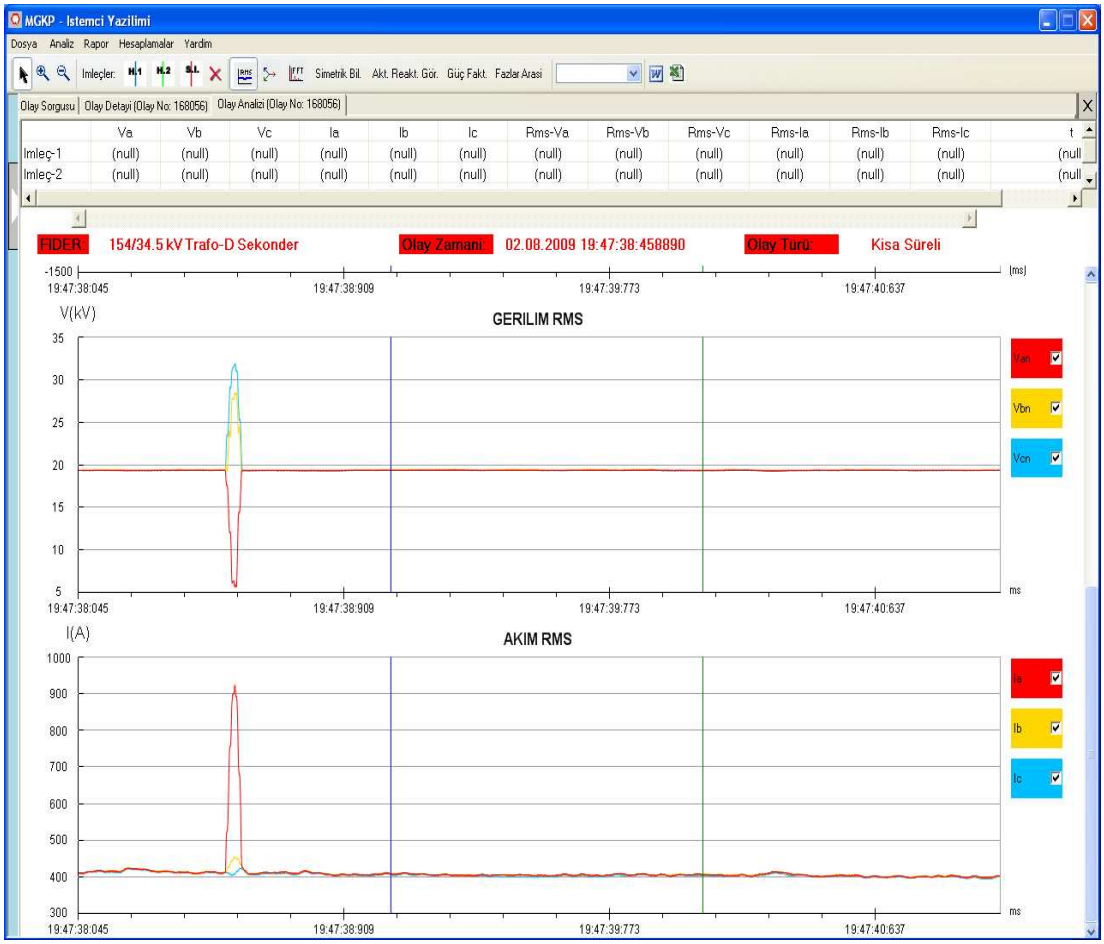


Figure 3. 8 PQ Parameter Analysis Software.

The PQ data mining and visualization software and the PQ Parameter Analysis software is implemented in C# programming language by using Microsoft Visual Studio development environment. MapObject [18] map library is used for the mapping applications and Xceed [19] graphic library is used for the graphic representations.

CHAPTER 4

K-MEANS BASED CLUSTERING METHODOLOGY FOR POWER QUALITY EVENT DATA

Clustering is the process of assigning unlabeled data into the intrinsic grouping. Clustering algorithms should be able to deal with different types of attributes, noise and outliers. The algorithms should also be able to handle large number of complex high dimensioned data items in the data set. In case of distance based clustering, the domain should be investigated and the distance should de defined according to the domain characteristics and clustering aims. There are four main clustering types: exclusive, overlapping, hierarchical and probabilistic clustering. Exclusive clustering algorithms assign a definite cluster to a data item; on the other hand overlapping clustering algorithms assign partial membership values to the data sets. Hierarchical clustering defines the clusters by considering the links between the data sets of the clusters. The probabilistic clustering defines a model for each of the clusters and assigns the data items to these clusters by maximizing the fit between the data item and the model.

In this chapter an exclusive clustering algorithm, k-means based method is described. The proposed method includes both the general exclusive clustering algorithm steps and the average linkage based cluster merging step of the hierarchical clustering. A

version of the proposed method and its application results are presented in the Knowledge Discovery and Data Mining (KDD2009) Conference [31]. The version given in this chapter is a combination of initial cluster selection, k-means, and average linkage algorithms together with further examination on the final clusters.

The proposed clustering method implements a modified version of K-means++ which is an exclusive clustering algorithm that divides a given data set into k clusters after applying an initial clustering method. The run time of K-means++ clustering algorithm increases linearly with the size of the data set, which is the main reason to choose the algorithm for power quality event clustering. After the cluster centers are selected, K-means algorithm is applied to cluster the PQ event data with error based outlier elimination modifications. K-means clustering has three main algorithm types: Lloyd's [15], swap and Hybrid. Heuristics based modifications are applied to the Pure Lloyd's algorithm. The modifications deal with the cluster center swapping strategy. Thus the proposed PQ event clustering method uses a hybrid algorithm for k-means clustering. After the clusters are formed average linkage calculation results are used to merge the similar clusters. The rule based event classification is carried on the resulting clusters by domain experts.

The proposed method helps to manage the event data to come up with PQ assessments for the specific measurement points and to make comparisons of various measurement points in terms of PQ of the electricity network. The method aims to cope with the huge event data size and cluster the event types thus the run time of the selected algorithm is a crucial aspect.

4.1 DESCRIPTION OF THE PROPOSED METHOD

The proposed method deals with the PQ events data that is stored as raw data in the database. The data flow of the proposed PQ event clustering method is given in Figure 4.1. The first three steps are described in Chapter 3. After these three steps are

handled, the event data become ready to be used in clustering. The client software enables users to run the proposed k-means based PQ event clustering method after selecting the time interval, transformer substations, voltage level and the k value parameters. The selection of the time period and transformer substation enables the users to examine events occurred in a precise period of time, in a specific area. It is possible to query the events in the selected voltage level of the transformer substations occurred in the selected time interval from the National PQ Database.

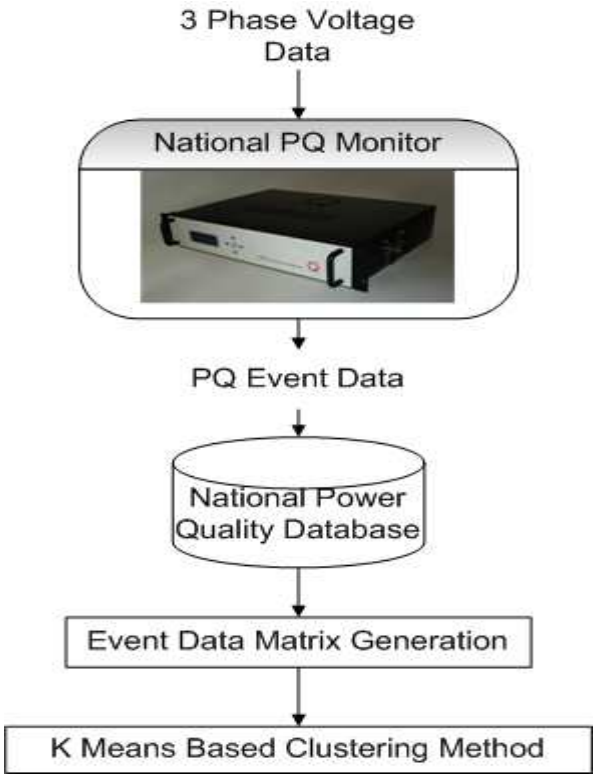


Figure 4. 1 Data Flow of the Proposed PQ Event Clustering Algorithm

After the selection of the method parameters, the event data collected for the selected measurement points is downloaded from the NPQMC. The downloaded data include three-phase voltages of sag, swell and interruption events sampled at a sampling rate of 25.6 kHz. Each event data package includes three-second raw data, thus each event has 25600*3 samples for each of three phases in the raw data from which the

required representative parameters could be selected, formed and computed. In the proposed clustering algorithm, each event is represented by a matrix of size (3x300) given in Figure 4.2.

$$Event = \begin{bmatrix} V_A[0] & V_A[1] & V_A[2] & \cdots & V_A[299] \\ V_B[0] & V_B[1] & V_B[2] & \cdots & V_B[299] \\ V_C[0] & V_C[1] & V_C[2] & \cdots & V_C[299] \end{bmatrix}$$

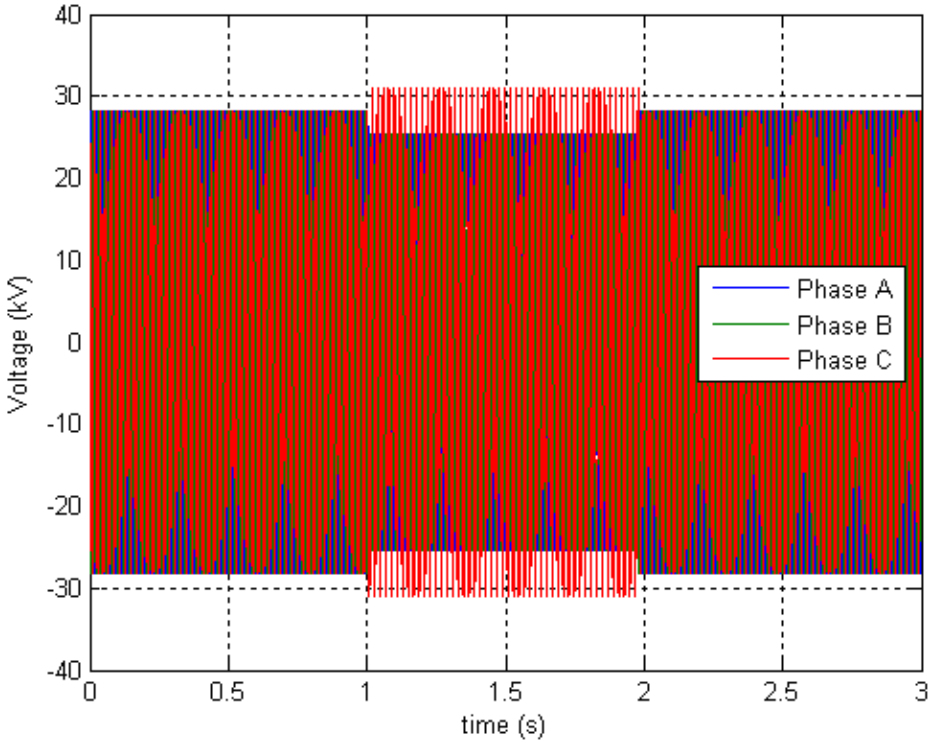
Figure 4. 2 Event Matrix Representation

The event matrices are calculated from rms values of the received raw data. The root mean square is a measure of the magnitude of a varying quantity. The rms value of $x_1, x_2, ..., x_n$ are calculated with the Formula 4.1.

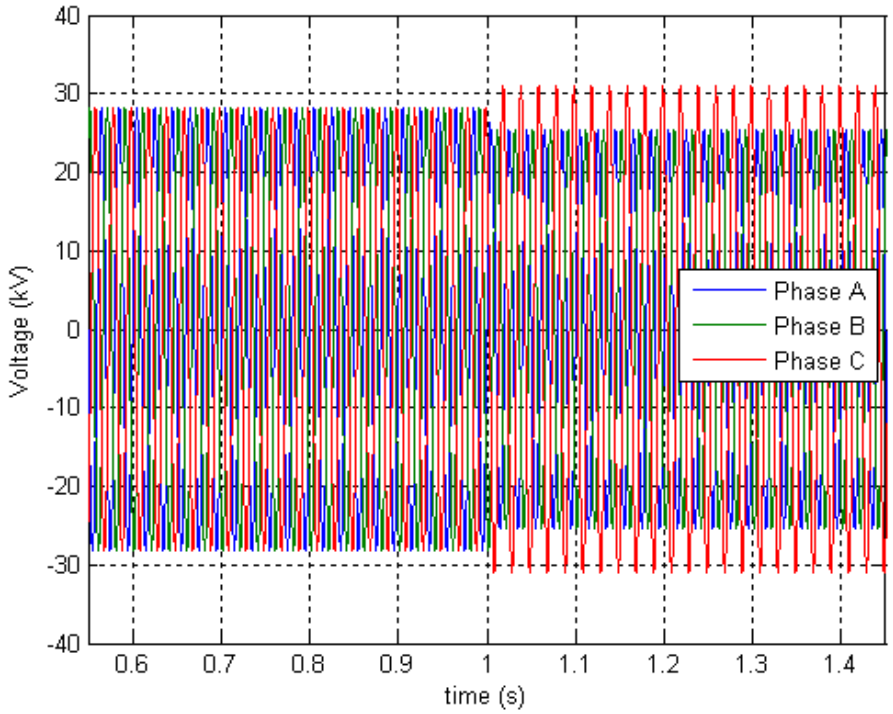
$$x_{rms} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \qquad \textbf{(4.1)}$$

The sampling rms values of the three-phase voltages computed for 512-sample windows shifted by 256 samples are obtained for each event data package as recommended in [11]. This process results in rms voltage matrices of size (3x300). After the matrices are computed, a normalization process is applied so that all rms voltages are represented as the percentages of their nominal value.

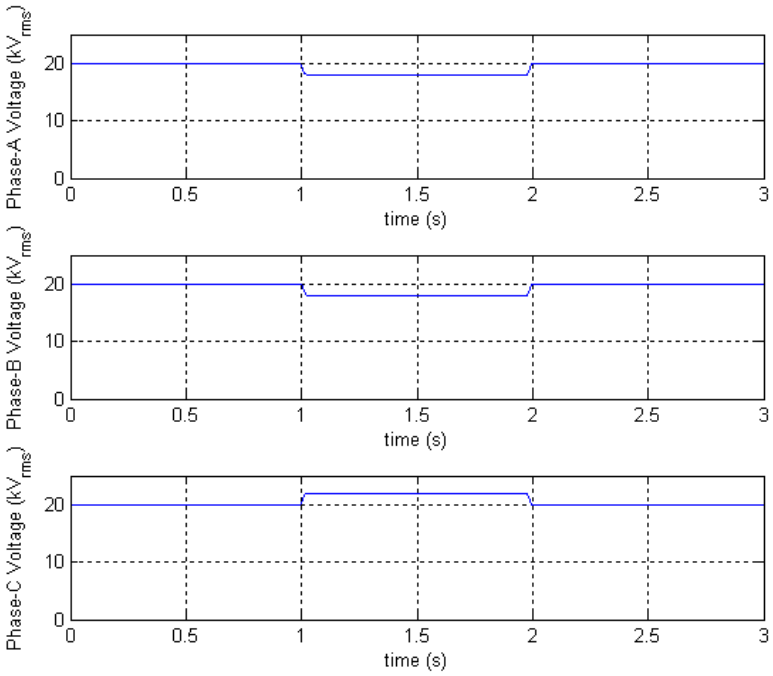
The rms value computation results are illustrated in Figure 4.3. The voltage time graph of the event's raw data is given in Figure 4.3 (a). Figure 4.3 (b) shows the event zoomed version of the event's raw data voltage time graph. After the rms calculations in Formula 4.1 are applied, the resulting representation of the event becomes as given in Figure 4.3 (c).



(a) Voltage-Time Graph of an Event Raw Data



(b) Zoomed Version of the Voltage-Time Graph of the Event



(c) Voltage-Time Graph of the Event RMS Values

Figure 4. 3 Raw Data and RMS Representation

Event raw data is stored on disc; the PQ database contains only the summary of the event. After the raw data is received from the network for each event, a data cleaning step should be carried on the data. The data cleaning step should eliminate the events with undefined or erroneous event time, voltage level, event type, event duration, maximum and minimum rms values.

The proposed clustering method is described to implement a modified version of K-means++ clustering algorithm. K-means++ algorithm requires a distance measure for the comparisons of two items in the data set. The distance measure used in the proposed method is selected as the Euclidean distance between the defined event matrices. The distance function which calculates the distance between the events shown in Figure 4.4 is defined as in Formula 4.2. *B* is the bias coefficient of the distance measure, which emphasizes the one-second part of the event data starting

from the 0.5th second. This biasing, which is for $B > 1$, is applied to obtain a clustering more dependent on the first one-second part of the event. After the distance measure $d_{1,2}$ is calculated, the distance value is calculated and decreased by one precision in order to prevent the small but continuous differences from being dominant during the cluster determination in the k-means algorithm.

$$E1 = \begin{bmatrix} V1_A[0] & V1_A[1] & V1_A[2] & \cdots & V1_A[299] \\ V1_B[0] & V1_B[1] & V1_B[2] & \cdots & V1_B[299] \\ V1_C[0] & V1_C[1] & V1_C[2] & \cdots & V1_C[299] \end{bmatrix}$$

$$E2 = \begin{bmatrix} V2_A[0] & V2_A[1] & V2_A[2] & \cdots & V2_A[299] \\ V2_B[0] & V2_B[1] & V2_B[2] & \cdots & V2_B[299] \\ V2_C[0] & V2_C[1] & V2_C[2] & \cdots & V2_C[299] \end{bmatrix}$$

Figure 4. 4 Representation of events E1 and E2

$$\begin{aligned} d_{1,2} = \sum_{i=A}^{C} \Big\{ & \sum_{j=0}^{49} \left(V1_i[j] - V2_i[j]\right)^2 \\ & + B \sum_{j=50}^{149} \left(V1_i[j] - V2_i[j]\right)^2 \\ & + \sum_{j=151}^{299} \left(V1_i[j] - V2_i[j]\right)^2 \Big\} \end{aligned} \tag{4.2}$$

The detailed block diagram of the proposed PQ event clustering method is given in Figure 4.5. After the rms matrices are formed and distance measure is determined, the cluster assignment phase starts. The cluster assignment phase includes four steps; initial cluster center selection, Lloyd's algorithm application, heuristics cluster correction and average linkage cluster merging.

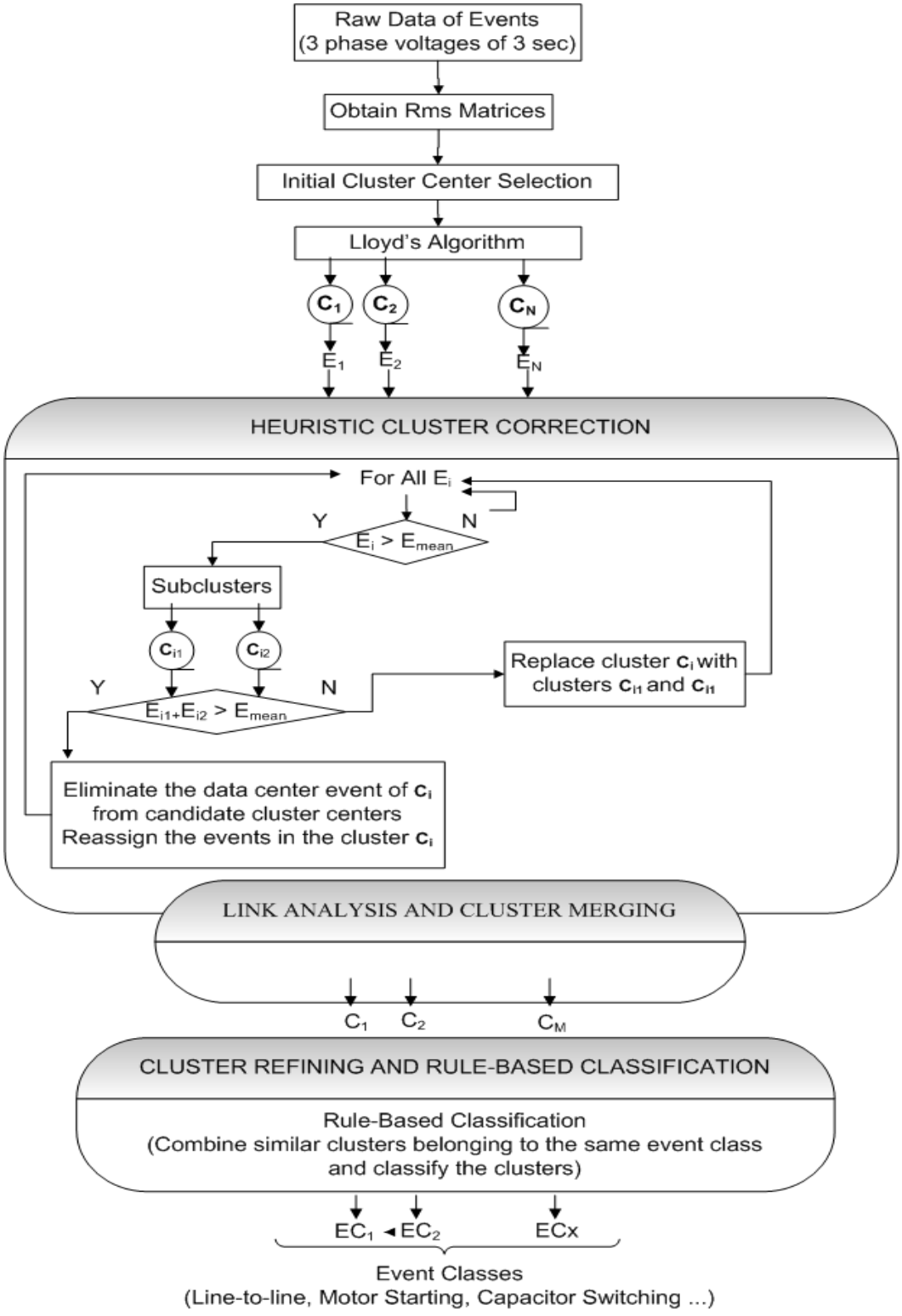


Figure 4. 5 Detailed Block Diagram of the Proposed PQ Event Clustering Method

Lloyd's algorithm may result in clusters far from the optimal, thus modifications are applied to ensure better clustering. Initial cluster center selection may have a big impact on the accuracy and performance of the algorithm. The employed cluster center selection algorithm aims to spread the initial centroids in the feature space in a way that the centroids are as far as possible from each other. The initial cluster center selection algorithm starts with computing the center of the whole data set. After the center is computed, the algorithm selects a random data item as the first cluster center. Afterwards, it iterates through selecting the next cluster center according to the distance between the current data item and the closest cluster center selected so far [29]. The algorithm selects the cluster centers by not only maximizing the distance between them, but also considering their distance from the overall center. In this way, selecting outliers as the cluster center is prevented.

When the cluster centers are defined, the Lloyd's algorithm assigns events to one of these clusters. After each item assignment, the centers of the clusters are recomputed and the algorithm continues until all the items are assigned to one of the clusters. When all of the events are placed into k clusters, each event is checked to be in the cluster which has the closest centroid to that event. If the event is not in the cluster with the closest centroid, it is assigned to the cluster with the closest centroid.

The heuristics cluster correction step swaps cluster centers and the selected candidates. The search heuristic used in this swapping is based on the mean square distance calculations. The total mean square distance between the center of the i^{th} cluster of the event number n and the events belonging to i^{th} cluster is computed as $E_i = E_{i1} + E_{i2}+... + E_{in}$. If $E_i > E_{mean}$ then the cluster C_i is refined. This refining process results in either the elimination of the center data of that cluster or the subdivision of the cluster into two clusters so that the mean error of each new class is less than E_{mean}. In case of sub clustering C_i is divided into C_{i1} and C_{i2} and the cluster correction step continues with the next cluster check. However the swapping requires

eliminating the previous cluster center, defining another cluster center and reassigning the events in the current cluster to the new cluster center combination.

After the heuristic clustering correction step, the proposed clustering method implements the average linkage cluster merging method. The average linkage is a method of calculating distances between clusters. According to the results of average linkage calculations, the clusters that are close according to link analysis are merged. Once linkage analysis is finished, rule-based clustering is applied to obtain event classes representing specific classes such as motor starting, line-to-line fault, line-to-neutral fault, capacitor switching, etc. by the domain experts.

4.2 PQ EVENT CLUSTERING RESULTS AND DISCUSSIONS

PQ Data Mining and Visualization Software user interface enables domain experts to select the methodology to be applied and time interval, voltage level and the transformer substations to be examined. After the k-means based clustering method is selected the results may also be seen in another page of the client software.

Let us query the events between 08.08.2007 21:27 and 08.08.2009 21:27. The selected transformer substations are Payas, Habas, Alibeykoy and Kaynaşlı and the voltage level is “All” levels. The user selects k as 20 and runs the proposed k-means based clustering methodology. The current configuration of the interface is given in Figure 4.6. After the method computes the results, they are displayed in the k-means clustering results page of the client software as shown in Figure 4.7. The clustering configuration is given in the left pane, and the resulting clusters are displayed with the details in the interface. The total distance of the events to cluster center versus cluster and cluster versus event counts are also given in the interface.

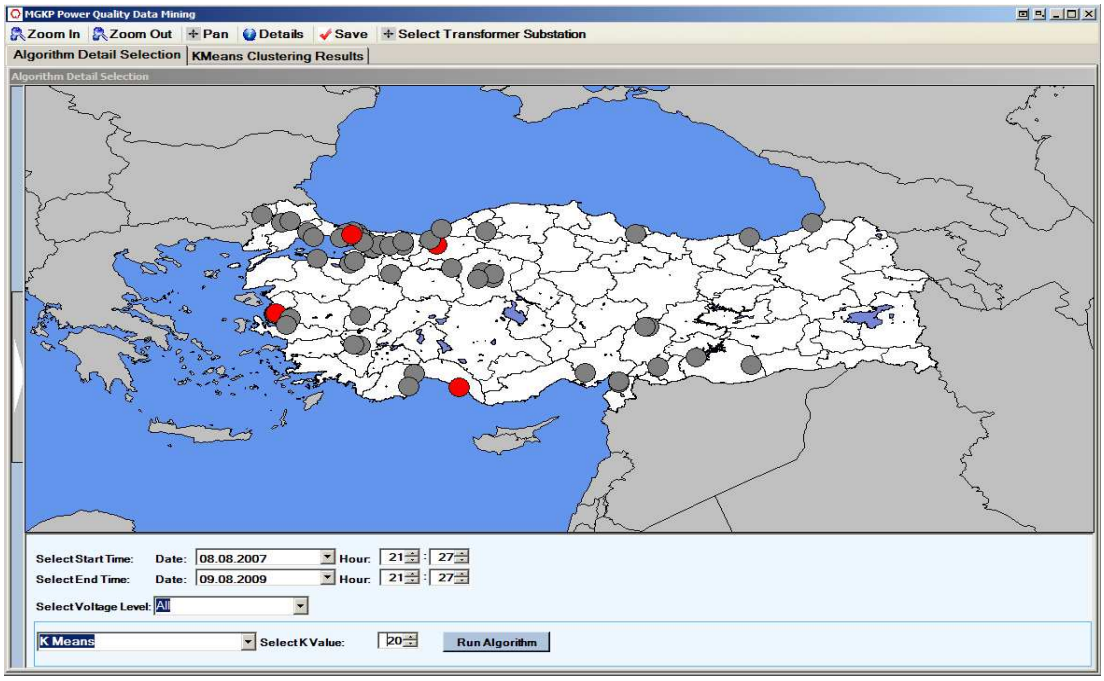


Figure 4. 6 Algorithm Selection User Interface of the Client Software with the Selected Configuration

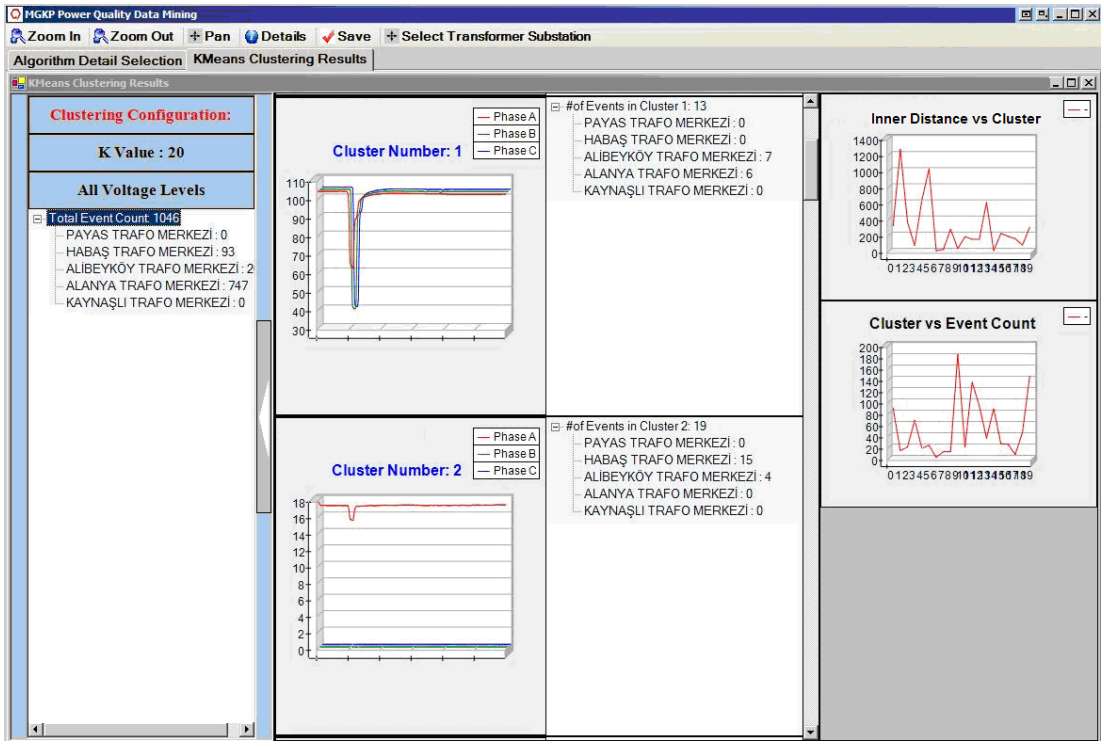
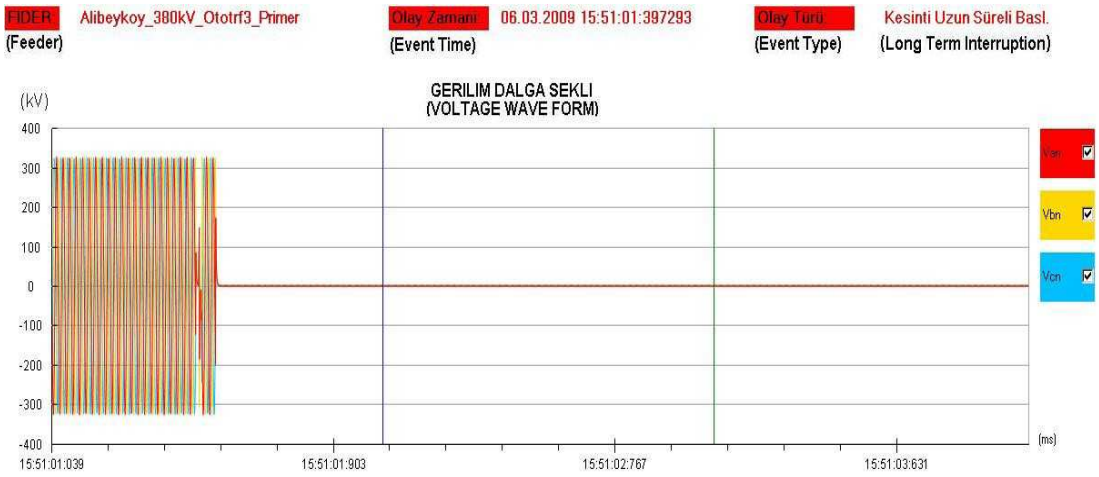
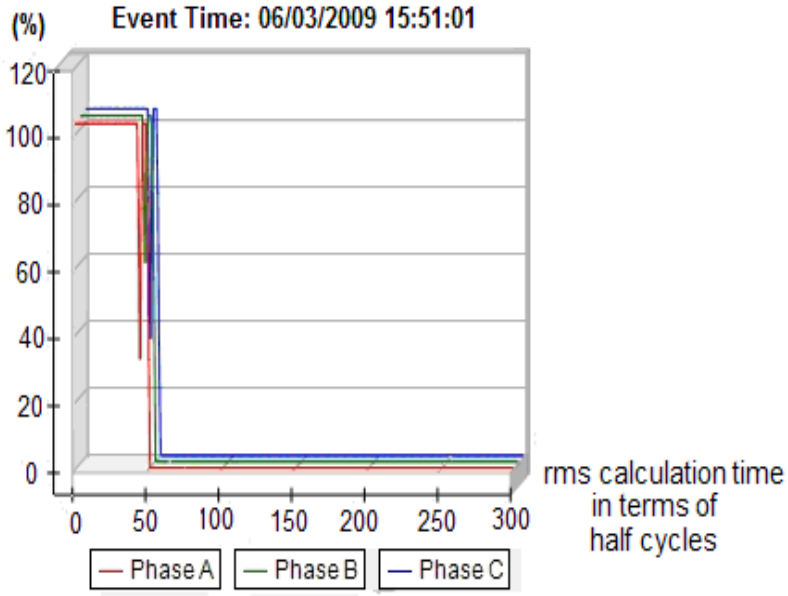


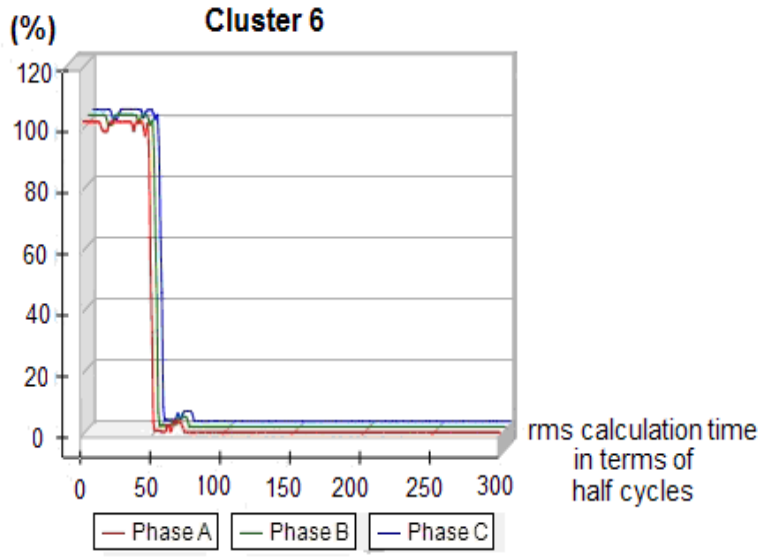
Figure 4. 7 Results of the Clustering with the Selected Configuration



(a) Voltage – Time Graph of the Raw Data of the Voltage Interruption Event



(b) Voltage – Sample Graph of the Rms Values of the Event



(c) Voltage – Sample Graph of the Cluster that the Event Belongs to

Figure 4. 8 Raw Data, Event and Cluster Graphs

The examination of the clustering result is done by the domain experts. The detailed examination of the cluster is done by examining one sample event from the cluster. The raw data voltage graph, cluster voltage graph and event voltage graph are examined together to label the cluster. Examples of these graphs are given in Figure 4.8.

The final step of the proposed power quality clustering method is rule based classification. The classification is done by the experts after the examination of the cluster according to electrical power related concerns. There are some examples of the event clusters which are determined as the basic fault classes after examining the results of the proposed clustering application on the selected configuration.

The cluster number 8 (C8) is a typical cluster for line-to-ground fault in Phase-A as observed in Figure 4.9. A sag event of approximately 0.5 seconds is observed in Phase-A, and swells are observed in Phase-B and Phase-C. This is the common

behavior of the transmission system during a line-to-ground fault. The voltage level for the sag is approximately 45%, which shows that the fault does not occur at exactly the measurement point, because if an event is in the same transformer substation where the fault occurred, the voltage level would be expected to decrease to 0%.

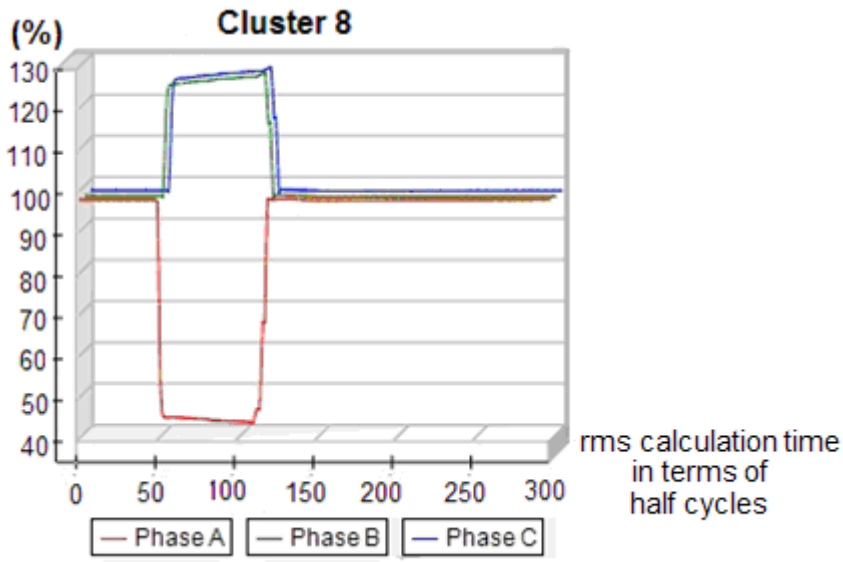


Figure 4. 9 Cluster Number 8, a typical class for line-to-ground fault at Phase-A.

In Figure 4.8.c, Cluster 6 (C6), which is a typical case of an interruption in all phases, is illustrated. This can be the beginning of either a short term interruption or a sustained interruption according to the IEEE Standard [16] depending on the duration of the event.

Cluster 7 (C7) is shown in Figure 4.10. As observed, this is a sag event in all three phases. Since the duration of the event is very short (approximately one cycle of the 50 Hz power signal), it can be defined as an “instantaneous” sag event according to [16]. It is observed that after the recovery of the event another voltage drop is observed but it is immediately recovered before it reaches to a sag level.

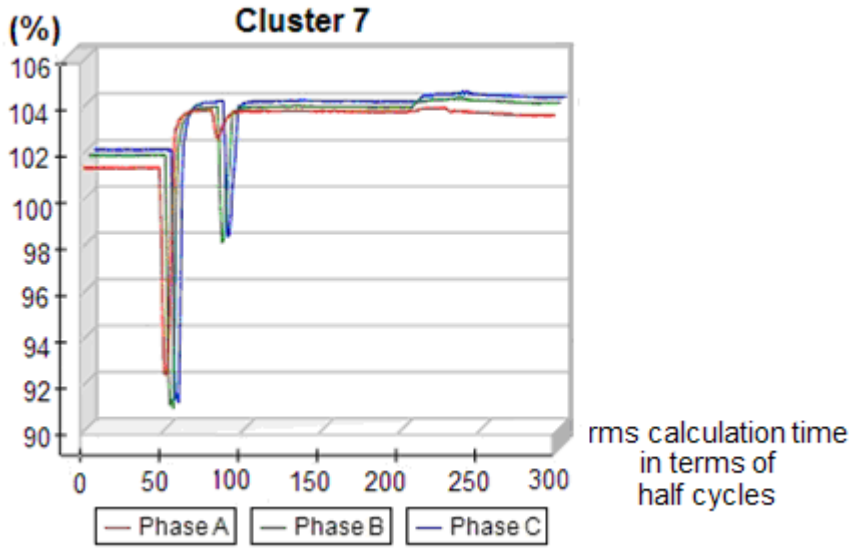


Figure 4. 10 Cluster Number 7, a typical instantaneous sag event in all three phases.

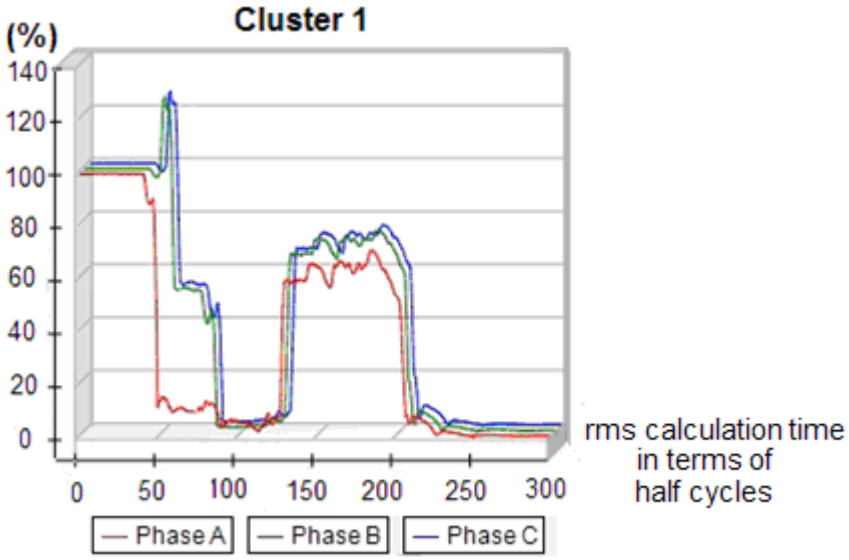


Figure 4. 11 Interruption in Phase-A followed by instantaneous swell, sag and interruption in Phase-B and Phase-C.

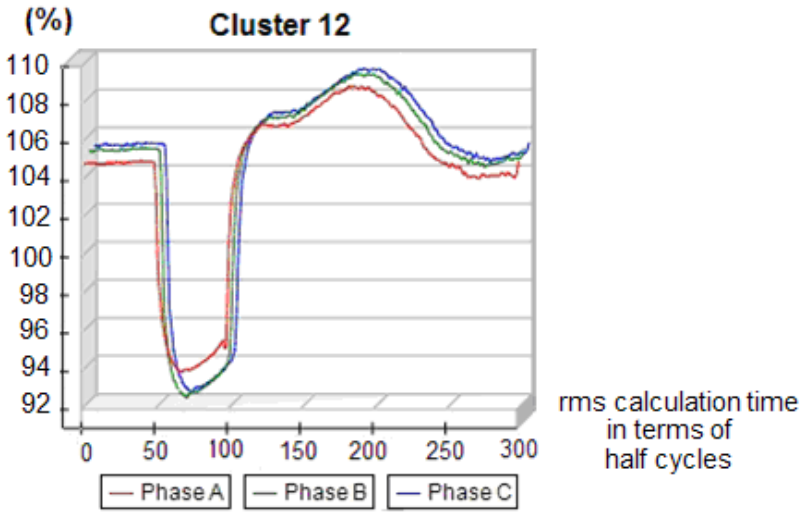


Figure 4. 12 Event in all three phases.

In Figure 4.11, Cluster 1 (C1) is shown. This cluster is among the most interesting clusters, since it includes various consecutive events in all phases. The event starts with an interruption in Phase-A, followed by instantaneous swells, sags, and interruptions in Phase-B and Phase-C. Then a recovery attempt is observed in the power system, in which voltages of all phases rise up to approximately 70 % of the nominal voltage; however, the attempt comes out to be unsuccessful and finally all phases end up with an interruption level.

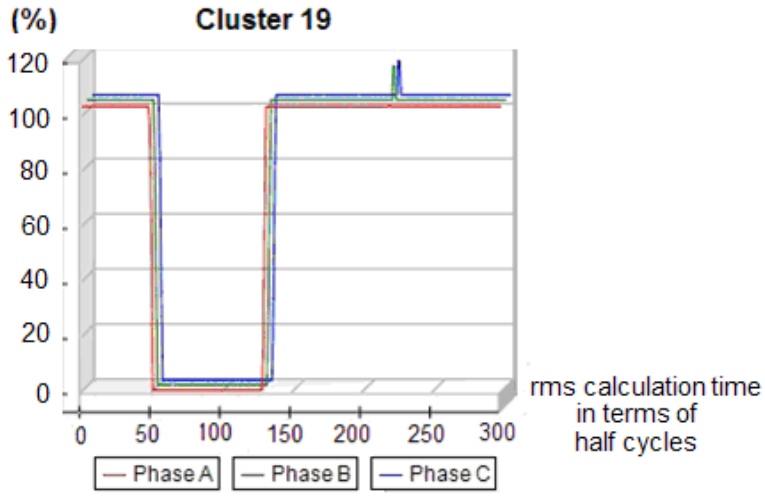


Figure 4. 13 Voltage interruptions in all three phases.

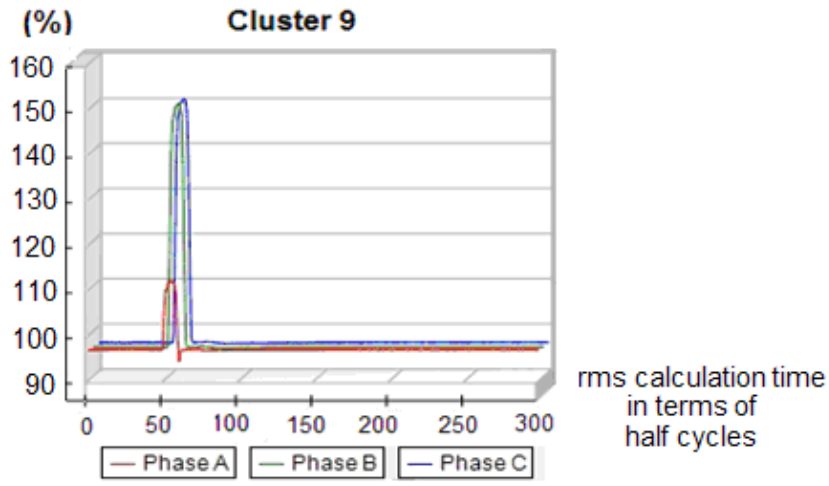
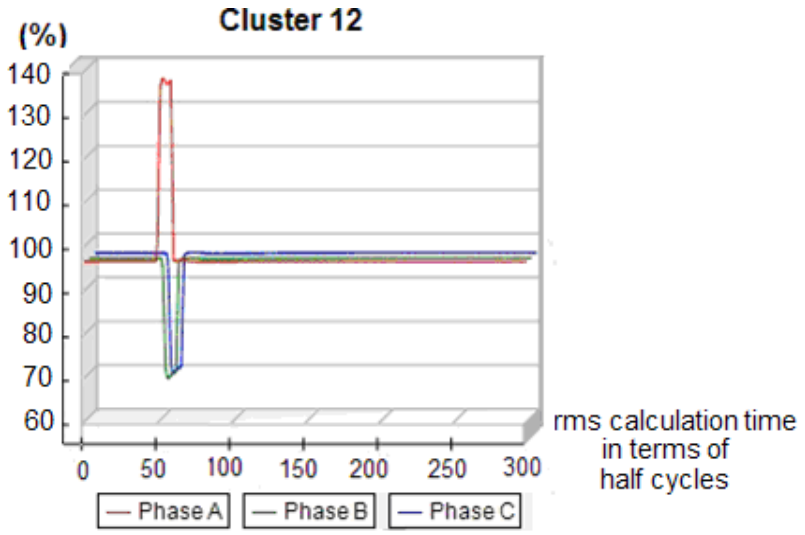
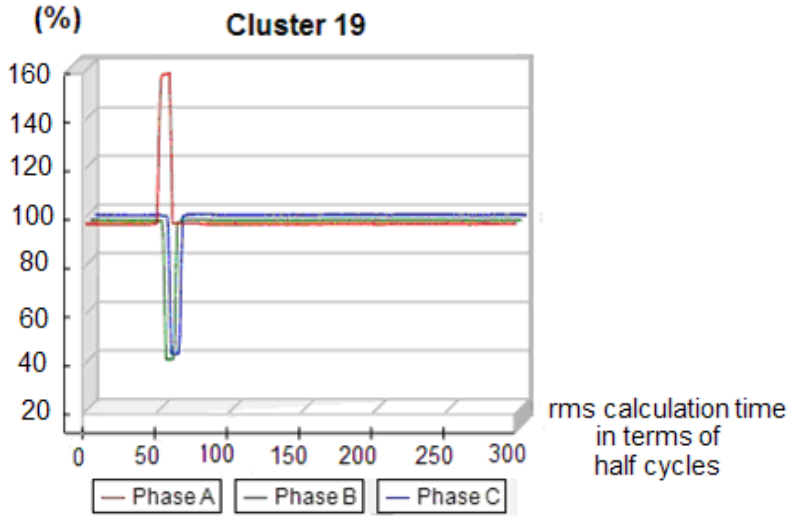


Figure 4. 14 Voltage swells in all phases.

Cluster 12 (C12) in Figure 4.12 consists of swells in all three phases, which correspond to the start of a high power consuming device. In Figure 4.13, Cluster 19 (C19) represents an interruption in all three phases. In Figure 4.14, Cluster 9 (C9) is a typical example for instantaneous swells of all phases. The clusters shown in Figure 4.8 to Figure 4.14 are obtained by selecting k value in the clustering algorithm as 20. Almost all of the clusters represent a specific power quality event class. When the number of clusters is increased to 30, some clusters belong to the same type of event classes are split into different clusters. These similar clusters give clues about the distance from the occurrence of the event. In Figure 4.15, such an example is illustrated for the total of 30 clusters. Cluster 12 (C12) and Cluster 19 (C19), in (a) and (b), represent swell in Phase-A and sag in the other phases.



(a) Voltage swell in Phase-A and sag in Phase-B and Phase-C



(b) Voltage swell in Phase-A and sag in Phase-B and Phase-C

Figure 4. 15 Voltage Sag and Swell Cluster Examples

The event cluster in 4.15 (b) represents events occurring at a point further than the point in which the events represented by the cluster in 4.15 (a) occurs, because the sag and swell levels are lower in Cluster 12 (C12), which means that the effect of the

sags and swells have less influence on the quality of power at the measurement point of the events in Cluster 12.

4.3 DESIGN ISSUES FOR THE PQ EVENT CLUSTERING METHOD

The running time, computational complexity and accuracy of the proposed algorithm are the main design issues. The current number of events is 1.060.000, where each event is defined as a 3.5MB of raw data. As a result, the proposed algorithm should be able to handle a huge amount of data by considering run time and memory bound problems. The design strategies of the proposed algorithm for these problems are discussed in the following.

The running time of a method depends on the computational complexity and the number of iterations of the required algorithm. The properties and the amount of the data may also have a big impact on the running time of the method. In the proposed PQ data mining method, the data of concern is high-dimensional and huge. The defined distance measure requires complex computations. The number of iterations may also become large in case of false selections of initial clustering centers. The iteration number has a big impact on the run time since the distances of all events to every cluster center are calculated until the convergence is achieved. In case of n events, and k clusters, the complexity of this computation is O(k*n). For p iterations, the complexity is O(k*n*p). The worst case occurs if all the events are replaced to new clusters, that is p is approximately O(n) and the algorithm is then $O(k*n^2)$. The proposed algorithm has heuristics cluster correction and average linkage cluster merging steps; however the applied versions of these steps allow limited number of iterations for these algorithms.

The solutions developed for solving long running time problem is based on reducing the computation complexity. The computational complexity may be reduced by

modifying the flow of the algorithm, forecasting results and performing less floating point calculations.

The first countermeasure for the run time problem is decreasing the number of iterations of both k-means clustering and heuristics cluster correction algorithms. In order to decrease the iteration number the clustering errors should be minimized. The appropriate assignment of the initial cluster centers results in decreases in the iteration number of both k-means and heuristics cluster correction algorithms. In order to utilize better cluster center selection, before the clustering algorithm starts, the average distance between two events in the data set is calculated according to the defined distance measure. Then the Lloyd's algorithm selects K cluster center so that the distance between any pair of centers is higher than the average of all distribution.

$$
\begin{aligned}
d_{1,2} = \sum_{i=A}^{C} \{ & \sum_{j=0}^{49} \left| V1_i[j] - V2_i[j] \right| \\
& + K \sum_{j=50}^{149} \left| V1_i[j] - V2_i[j] \right| \\
& + \sum_{j=151}^{299} \left| V1_i[j] - V2_i[j] \right|
\end{aligned}
\tag{4.3}
$$

The second countermeasure for the run time problem is decreasing the computational complexity. The most complex computation in the proposed PQ event clustering method is the distance measure calculation. The distance measure calculation that is described in Formula 4.2 is based on Euclidean distance. In Euclidean distance, each distance calculation requires computing the square of 300*3 numbers. Thus the distance measure is reassigned as in the Formula 4.3 in order to eliminate the square calculations; this distance measure is called as Manhattan Distance [32].

The third countermeasure for the run time problem is parallel computation. Iterative computations of distance measures result in a long running time. Current computers

generally have more than one processor, which enables parallel computation of the distance measure. After the events are assigned to initial clusters, parallel computations of the events and cluster distances start. Events are distributed to the available processors. Each processor is able to calculate the distance of the event to the cluster centers. The distances of an event from each cluster center are calculated, and decided to replace the event into another cluster or not. After the replacements, the cluster centers are computed iteratively. The parallel computation and iterative assignment continues until convergence.

The proposed PQ event clustering method deals with a large data set, mentioned as 1.060.000 events for now. This number of high dimensional data cannot be loaded simultaneously into memory. Therefore the regular memory based programming approach cannot be applied for this case. The events are retrieved via internet thus exchanges of the events are costly. The only improvement is running the method in the Power Quality Monitoring Center by enabling the use of the event data directly from the network without using database connection. The method is still bounded by input output operations. The exchanges are handled in order to pass over the data once.

CHAPTER 5

FUZZY CLUSTERING METHODOLOGY FOR POWER QUALITY EVENT DATA

Clustering is the process of grouping unlabeled data according to an appropriate similarity measure. Depending on the data domain and expectations, different types of similarity measures and algorithms may be used. Similarity measure is based on the distance and connectivity between the data points. Clustering algorithm types are mentioned to be exclusive, overlapping, hierarchical and probabilistic clustering. K-means clustering based methodology described in Chapter 4 is an example of exclusive clustering. In the exclusive clustering, a definite cluster is assigned to each data item. However a data point can belong to all defined clusters with a membership degree in overlapping clustering algorithms. Partial membership knowledge may be used to identify further relations between the data items. In this chapter an overlapping clustering example, fuzzy c-means clustering method for power quality domain is described.

The aim of clustering power quality event data is to define characteristics and types of the event clusters. Fuzzy c-means clustering algorithm is selected not to lose the details of relations between events and clusters by keeping all membership degrees. In Fuzzy c-means clustering algorithm [20, 21], data items may be assigned to two or

more clusters with different membership values. The algorithm reveals not only the relation between the events and its best fit cluster but also the relations between the event and other clusters. The knowledge of membership degrees of the events to all clusters enables domain experts to investigate characteristics of the events from inter-cluster view. The distribution of the events to the clusters for different time periods enables defining characteristics of the transformer substations. Fuzzy membership values supply flexible knowledge compared to definite clustering results of the exclusive clustering. The proposed fuzzy c-means clustering based method is designed to cluster the huge number of events by dividing the data set into chunks and combining the results obtained for each chunk. Memory problem, running time and accuracy of the clustering are the critical aspects considered during the design and implementation of the algorithm.

In order to define the fuzzy rules for data domain, features of the input data set should be examined and supplied to the algorithm. Inputs of the algorithm are the unlabeled data set, number of clusters, maximum number of iterations, fuzziness degree and accuracy threshold. Data set may be represented as $X = \{X_1, X_2, ..., X_n\}$ where X_k is a vector of p features; p is the number of features in the vector representing each data item. The maximum number of iterations and the accuracy threshold are used to define the stopping criteria of the algorithm. Fuzziness degree (weighting exponent) m is used to make the clustering more immune to the noisy data items. It also provides weighting centrally and densely located cluster centroid vectors. After the initial parameters are obtained, fuzzy c-means algorithm iterates through the data set and selects the membership values by minimizing the total weighted mean square error [22].

The main elements in the fuzzy c-means clustering method described in this chapter are proposed method description, features analysis on the data set, cluster analysis and classifier design. Method description covers the algorithm design and implementation details. Feature analysis includes data preprocessing, extraction and

selection processes. Cluster analysis contains the validity check and labeling steps which is applied on the resulting clusters of the fuzzy c-means algorithm. Classifier design includes further investigation on the results of the clustering such as estimations and predictions for the data domain. PQ event clustering results and design issues of the fuzzy c-means PQ event clustering method are the other two important topics discussed in this chapter. The description of the proposed method, details of cluster analysis, classifier design, the result of the applied clustering method and design issues of the fuzzy c-means PQ event clustering method are explained in the following sections.

5.1 DESCRIPTION OF THE PROPOSED METHOD

The method includes power quality data and fuzzy c-means clustering algorithm. Power quality event data is the concern of the proposed fuzzy c-means clustering method. PQ events are stored in the database as summary of its details and on disc as raw data. For the initial event analysis, the summary is retrieved from the database and examined to extract the occurring time, place and down sampled rms values. In order to examine a power quality event by considering all of its features, the 3.5MB raw data should be retrieved from disc. The raw data amount is 1.000.000 events * 3.5 MB per event. The summary of the event is stored in the database and raw data is stored on disc in order to enable both fast and organized access to the data. Raw data could be queried from the database and retrieved via internet access. Direct disc access to the event data is also possible by running the analysis program on the same storage. The proposed algorithm is designed to retrieve the data from disc.

The data flow of the proposed method is shown in Figure 5.1. After the PQ event data is formed by the power quality monitors, it is stored on the database and on the disc. The proposed fuzzy c-means algorithm starts with the formation of the event vectors from the event summary data, system details and event raw data. The method

performs chunk clustering and forms the final clustering structure from the results of the fuzzy c-means clustering applications on all of the chunks.

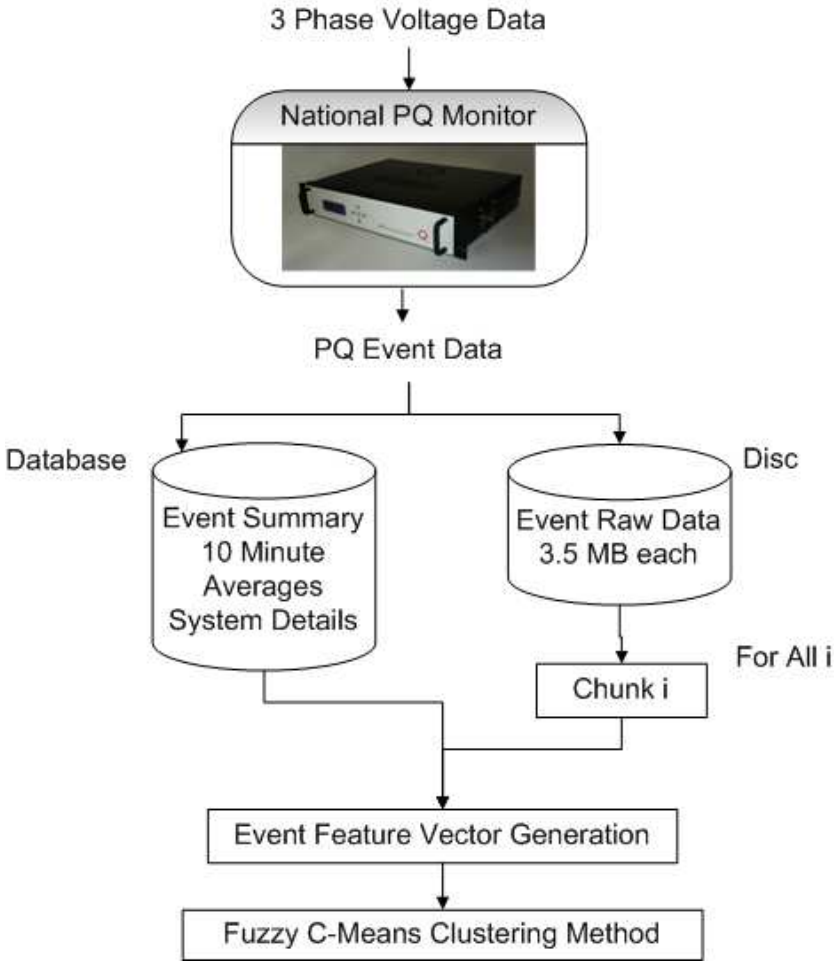


Figure 5. 1 Data Flow of the Proposed Fuzzy C-Means PQ Event Clustering Algorithm

The formation of the event feature vector includes data normalization, feature construction and feature selection steps. Normalization of the PQ event data is based on the min-max normalization and the voltage differences between the occurring places of the events. The scales of the rms values are defined according to the voltage differences between the occurring transformer substations of the events. Let the power quality event data set is represented as $E = \{E_1, E_2, ..., E_n\}$ where $E_k \in \Re^p$

and p is the number of features in the vector representing each event. Since the chunks are loaded into the memory separately, normalization is performed for each chunk of the data independently. The general normalization formula is given in Formula 5.1.

$$E_k^{p\prime} = \left[\frac{E_k^{p} - Min_{Chunki}^{p}}{Max_{Chunki}^{p} - Min_{Chunki}^{p}}\right] * [1-0] * \left[\frac{100 * \sqrt{3}}{VoltageLevelofE_k}\right] \quad \textbf{(5.1)}$$

Feature construction and selection are the other steps in the formation of the event feature vector. These steps deal with feature extraction and selection processes from the raw data. Raw event data contains 25600 voltage and current measurements for each of three phases that are collected in three second neigbourhood of the event. In Chapter 4 each event is represented with a (3x300) matrix. In the sampling described in the fourth chapter, rms values of the three-phase voltages computed for 512-sample windows shifted by 256 samples are obtained for each event data package as recommended in [11]. In order to reduce the calculations and memory requirements, sampling is done from the continuous 512 samples without overlapping the samples. Since each sample measurement is considered once in this under sampling, each event may be represented by a (3x150) matrix in this data modeling as illustrated in Figure 5.2.

$$Event = \begin{bmatrix} V_A[0] & V_A[1] & V_A[2] & \cdots & V_A[149] \\ V_B[0] & V_B[1] & V_B[2] & \cdots & V_B[149] \\ V_C[0] & V_C[1] & V_C[2] & \cdots & V_C[149] \end{bmatrix}$$

Figure 5. 2 Event Matrix Representation

The block diagram of the proposed fuzzy c-means PQ event clustering method is given in Figure 5.3. The first step in the method is collecting the user specified parameters; such as the number of final clusters c, maximum number of iterations T,

fuzziness degree m and termination threshold ε. The cluster number c should satisfy the inequality $2 \leq c \leq n$, where n is the total number of events to be clustered. The initial condition on fuzziness degree is $1 \leq m$. The other user specified parameters are the constraints for the events to be considered in the clustering such as possible transformer substations, possible voltage levels and time intervals those events could belong to.

The second step of the method is forming the cluster centers. Because of the characteristics of the fuzzy c-means clustering, the initial cluster center selection is a crucial subject. The sensitivity of the fuzzy c-means algorithm to the initial cluster selections may result in slow convergence and local minimum problems. The algorithm for initial cluster selection described in Chapter 4 requires passing over the entire events once and calculating distance values between all event pairs before the clustering step starts. In order to eliminate these access requirements and calculations, the initial cluster selection is randomized in the fuzzy c-means clustering method. The cluster centers are selected from all data items by forming a distributed sampling. Selecting cluster centers randomly may lead the algorithm to result in local minimum. However multiple runs of the algorithm could be investigated in order to get a good overview for the data domain.

The third step is dividing the data into chunks. The fuzzy c-means algorithm should be designed to apply on the huge number of events. All of these events could not be loaded into the memory simultaneously. Therefore chunk based clustering is required to be employed. Chunks are formed by sampling from all of the selected feeders. Events in the feeders are distributed to chunks by sampling.

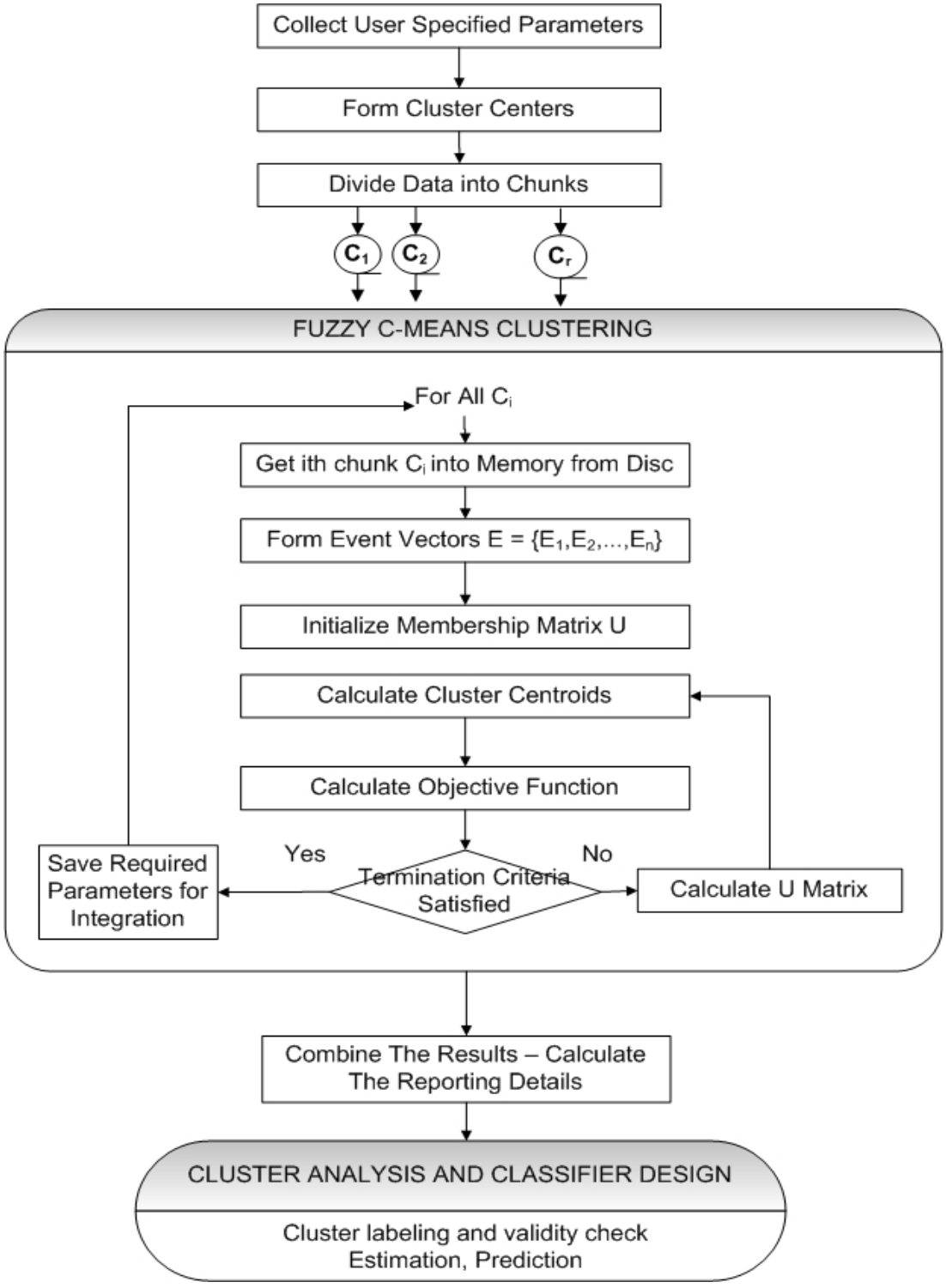


Figure 5. 3 Block Diagram of the Proposed Fuzzy C-Means PQ Event Clustering Method

The next step in the method is iteratively applying fuzzy c-means clustering on the designed chunks of events. The fuzzy c-means clustering algorithm is based on minimizing the objective function given in Formula 5.2 which is also called as performance index. The objective function is the degree of similarity between cluster centroids and the power quality events. U is (c x n) membership matrix which contains membership degrees of each event to the clusters; the membership matrix should satisfy the equation in Formula 5.6 U_{ik} is the membership degree of event E_k for the ith cluster, calculation of U_{ik} is given in Figure 5.4. The membership matrix is one of the main memory consumption reasons since its size depends on the event number linearly. d_{ik} is the distance between ith cluster centroid and the kth event. The proposed fuzzy c-means method employs the Manhattan norm in order to decrease the complexity of the distance measure calculations. The details of the distance measure are given in Formula 5.3 C_i is the center of ith cluster. The Formula 5.5 describes the calculation of the cluster centroid.

$$J_m = \sum_{i=1}^{c} \sum_{k=1}^{n} U_{ik}^{\ m} \ * d_{ik}^{\ 2} \tag{5.2}$$

$$d_{ik}^{\ 2} = \left\| E_k - C_i \right\|_1^{\ 2} \text{ Where } \left\| E \right\|_1 = \sum_{i=1}^{p} E_i \tag{5.3}$$

$$d_{ik} = \sum_{i=0}^{C} \left\{ \sum_{j=0}^{49} \left| E_k[j] - C_i[j] \right| + K \sum_{j=50}^{149} \left| E_k[j] - C_i[j] \right| \right.$$

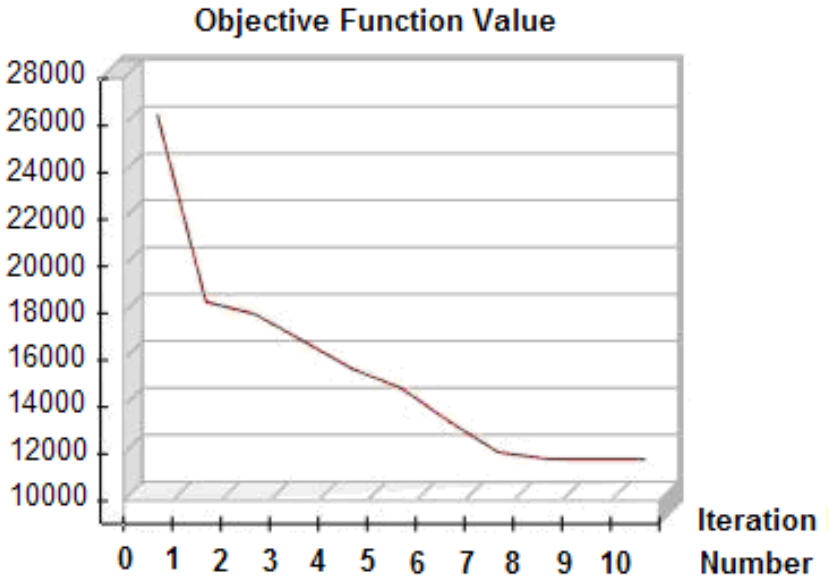
$$U_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{jk}}{d_{ik}} \right)^{\frac{2}{m-1}}} \quad \forall k \text{ And } \forall i \tag{5.4}$$

$$C_i = \left(\frac{\sum_{k=1}^{n} U_{ik}^{\ m} * X_k}{\sum_{k=1}^{n} U_{ik}^{\ m}} \right) \quad \forall i \tag{5.5}$$

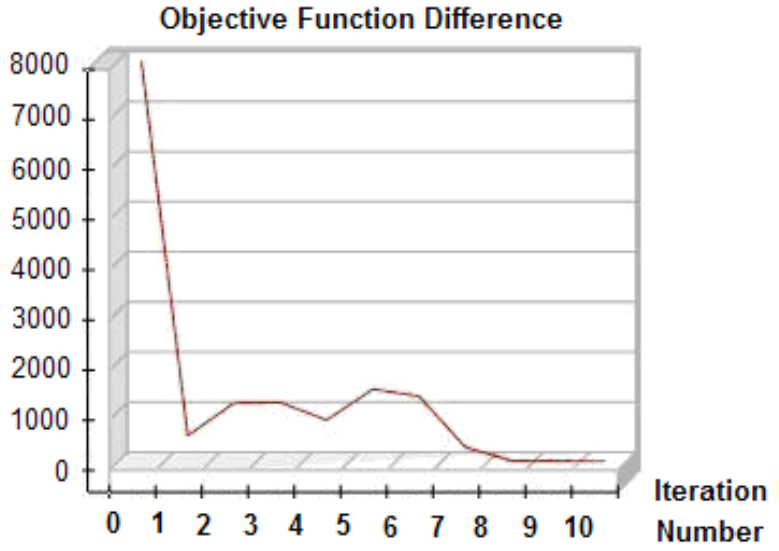
$$\sum_{i=1}^{c} U_{ik} = 1 \quad \forall k \tag{5.6}$$

The objective function calculations are the main step in the fuzzy c-means clustering algorithm. The algorithm employs a fuzzy partitioning through an iterative optimization of the objective function J_m. In each step, the membership matrix and the cluster centroids are recalculated according to the new configuration. The iteration stops when the termination threshold is achieved. This iteration converges to a local minimum or a saddle point of J_m [23]. The change of the objective function with the iterations for a chunk is given in Figure 5.4 (a). As it is seen from the figure, for this chunk and configurations the saddle point of J_m is between 10.000 and 12.000. The change in the objective function between the consecutive iterations for a sample chunk is given in Figure (b). As expected the difference between the consecutive J_m values decreases below the termination threshold Э when the clustering of the chunk is finished.

As the last step of the fuzzy c-means clustering, after all chunks are clustered, the results are combined and the reporting details are calculated.



(a) Objective Function versus Iterations Graph



(b) Objective Function Difference versus Iterations Graph

Figure 5. 4 Objective Function Graphs

5.2 CLUSTER ANALYSIS AND CLASSIFIER DESIGN

Cluster analysis includes fuzzy labeling of the cluster membership degrees, cluster correction and cluster labeling steps. At the end of the chunk based fuzzy c-means clustering, the distribution of the events to clusters and cluster center rms values are obtained. The cluster membership degrees are labeled according to the stored values in the membership matrix U. The labeling is done according to the rules defined below:

- If $0 < U[i,k] < 0.3$ then event E_k is defined to belong cluster C_i, SOMEWHAT
- Else If $0.3 \leq U[i,k] < 0.6$ then event E_k is defined to belong cluster C_i, MODERATELY
- Else If $0.6 \leq U[i,k] < 0.9$ then event E_k is defined to belong cluster C_i, MOSTLY
- Else If $0.9 \leq U[i,k]$ then event E_k is defined to belong cluster C_i, COMPLETELY

Cluster correction algorithm is applied on the final clusters which are formed after all chunks are clustered and results are combined. In Chapter 4, the average linkage cluster merging method is applied on the final clusters. However in this case in order to apply the average linkage clustering merge algorithm, the cluster membership information for all of the events and average distance between the events of all clusters are required. Only the summary of the membership matrix *U* is stored and known at the end of the algorithm, since the algorithm is chunk based. Even if the cluster membership degrees are kept in the raw data, retrieving these degrees and calculating the average linkage values for each event require multiple passes over all raw data. Because of these restrictions, eliminating empty clusters is the only applied run time cluster refining step. Further examinations on the clusters are handled by the domain experts.

In order to define the criteria for comparing accuracy levels between the clusters, an overall performance index is calculated for each cluster C_i. The calculation of the cluster performance index is given in Formula 5.7. Even if this measure does not reflect the accuracy of the cluster, the value may be used to compare the clusters with each other. The equation in Formula 5.9 gives a performance index for all clustering, where the first part is the sum of cluster fluctuation and the second part is the sum of the between fuzzy cluster fluctuations. For a better clustering both the first part and the second part should be minimized.

$$P(C_i) = \sum_{k=1}^{n} U_{ik}{}^{m} * \left\| E_k - C_i \right\|_1 - \left\| C_i - \overline{E} \right\|_1 \quad \textbf{(5.7)}$$

$$\overline{E} = \frac{1}{n} \sum_{k=1}^{n} E_k \quad \textbf{(5.8)}$$

$$\underbrace{\sum_{i=1}^{c} \sum_{k=1}^{n} U_{ik}{}^{m} \left(\left\| E_k - C_i \right\|_1 \right)}_{First-Part} - \underbrace{\sum_{i=1}^{c} \sum_{k=1}^{n} U_{ik}{}^{m} \left(\left\| C_i - \overline{E} \right\|_1 \right)}_{Second-Part} \quad \textbf{(5.9)}$$

The performance index results of the clusters and rms voltage graphs enable the domain experts to refine, correct and analyze the clusters. The domain experts examine and label the formed clusters. The formed class labels are Line-to-line, motor starting, capacitor switching, etc. From the labeled clusters and domain knowledge, classification rules could be formed. The classification rules are used in the classifier design by the experts.

5.3 PQ EVENT CLUSTERING RESULTS AND DISCUSSIONS

Domain expert selects the configuration to run the fuzzy c-means clustering by using PQ Data Mining and Visualization Software user interface. The configuration parameters are transformer substations, time interval, voltage level, fuzziness degree,

and cluster number. The results of the fuzzy c-means clustering on power quality events satisfying the selected configuration are displayed as a report in a separate document. The document includes the configuration details, cluster voltage graphs, cluster event distribution counts and the performance index versus cluster number graph.

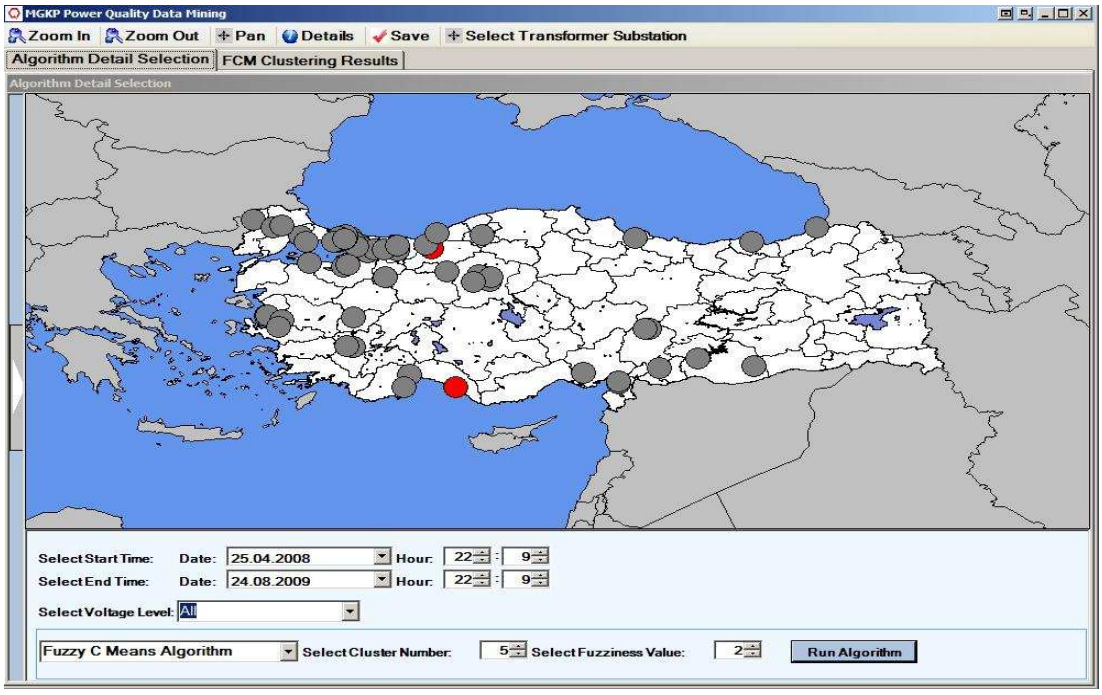


Figure 5. 5 Selected Configuration on the PQ Data Mining and Visualization Software user interface

In the following, an example application of the fuzzy c-means clustering is given. Let us query the events between 25.04.2008 22:09 and 24.08.2009 22:09. The selected transformer substations are Payas, Habas, Alibeykoy, Alanya and Kaynaşlı and the voltage level is "All" levels. The user selects cluster number as 5, fuzziness value as 2. The configuration selection user interface is given in Figure 5.5.

The distance measure calculation is modified in order to exclude the period before the start of the event and after the end of the event. This modification is applied to prevent event shifts to become effective in distance measure calculations.

Example voltage graphs of clusters formed by applying fuzzy c-means clustering on the selected configuration are shown in Figures 5.6 to 5.9.

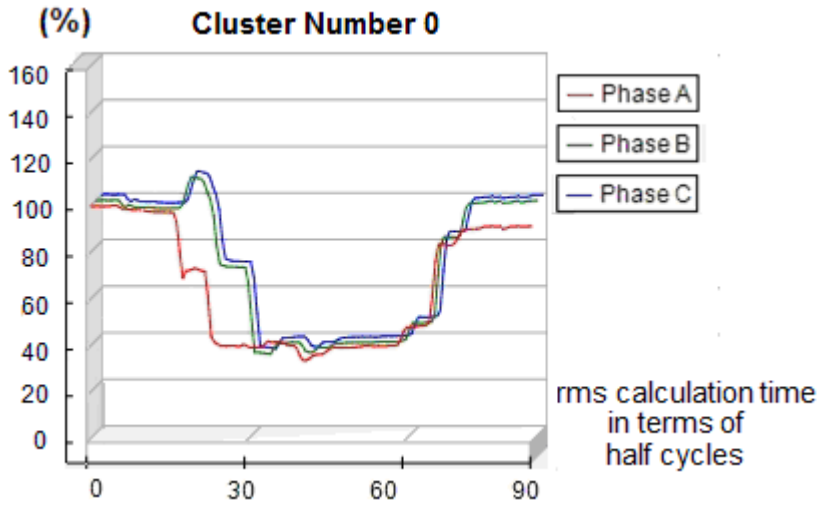


Figure 5. 6 Cluster Number 0

Table 5. 1 Cluster 0 Event Distribution

	Transformer Substation	Somewhat	Moderately	Mostly	Exactly
1	PAYAS	403	20	20	9
2	HABAŞ	49	30	14	10
3	ALİBEYKÖY	140	28	17	11
4	ALANYA	432	50	16	6
5	KAYNAŞLI	0	0	0	0

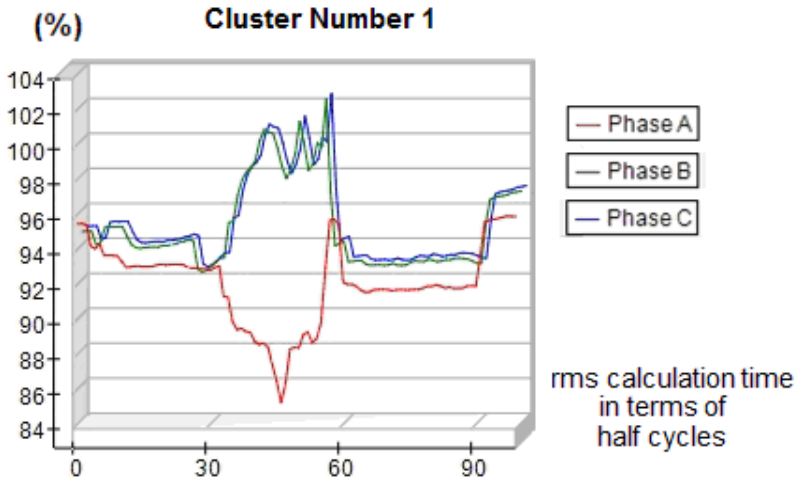


Figure 5. 7 Cluster Number 1

Table 5. 2 Cluster 1 Event Distribution

	Transformer Substation	Somewhat	Moderately	Mostly	Exactly
1	PAYAS	340	85	15	12
2	HABAŞ	3	14	11	0
3	ALİBEYKÖY	61	18	13	12
4	ALANYA	217	12	29	18
5	KAYNAŞLI	0	6	0	0

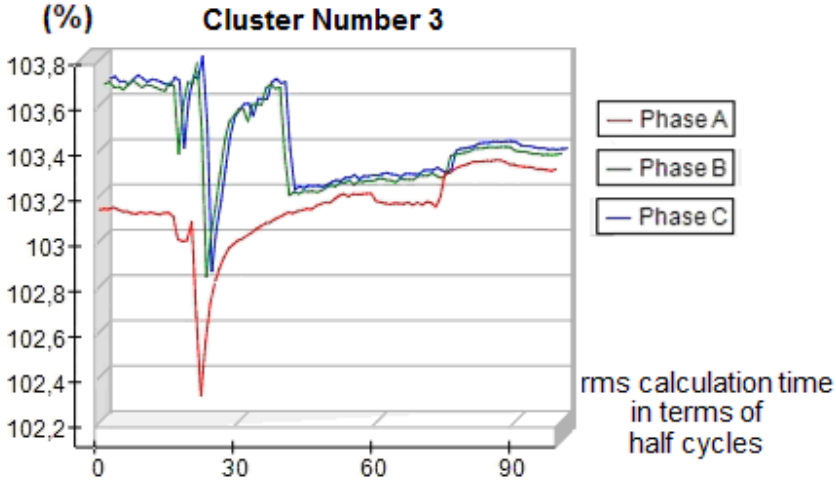


Figure 5. 8 Cluster Number 3

Table 5. 3 Cluster 3 Event Distribution

	Transformer Substation	Somewhat	Moderately	Mostly	Exactly
1	PAYAS	45	37	10	3
2	HABAŞ	36	57	0	0
3	ALİBEYKÖY	78	97	15	9
4	ALANYA	232	253	21	12
5	KAYNAŞLI	0	6	0	0

The examination of the application of the fuzzy c-means clustering on power quality data may be used to define the types of the events. The event displayed in Figure 5.10 is labeled as "Somewhat" because of its belonging degree to the cluster given in Figure 5.9. The event displayed in Figure 5.11 is labeled as "Moderately" because of its belonging degree to the cluster given in Figure 5.9. And the event displayed in Figure 5.11 is labeled as "Mostly" because of its belonging degree to the cluster given in Figure 5.9. Examination of these events results in the following implications:

- "Somewhat" label cannot be used for type description.
- "Moderately" label gives clues about the event type but not about the magnitude and duration details.
- "Mostly" label may be examined for magnitude, duration and location comparisons.

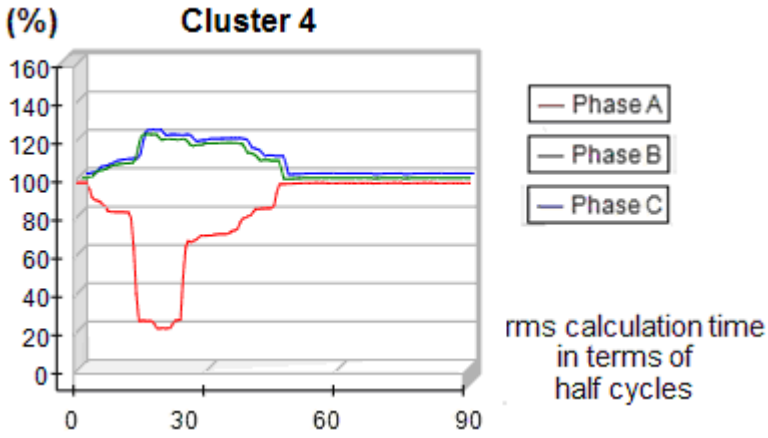


Figure 5. 9 Cluster Number 4

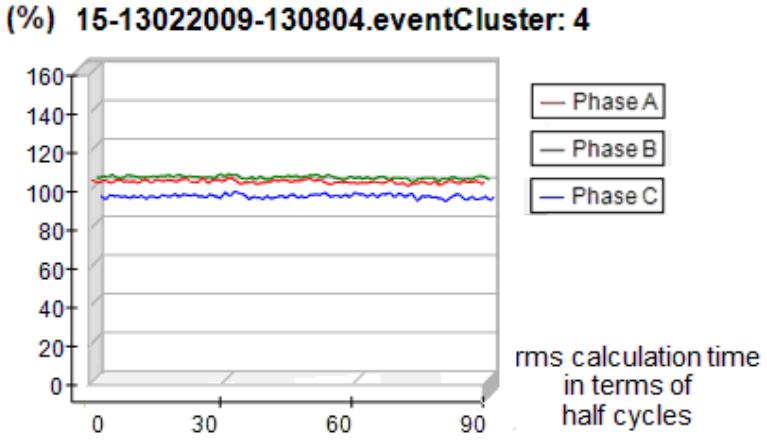


Figure 5. 10 Event belongs to Cluster 4 - Somewhat

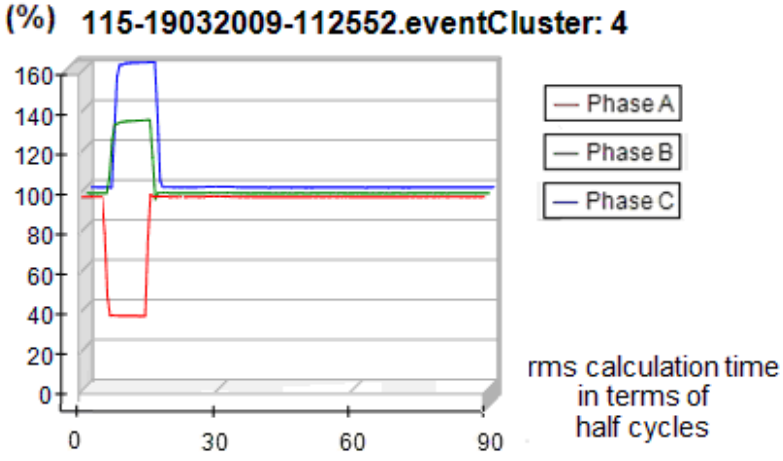


Figure 5. 11 Event belongs to Cluster 4 - Mostly

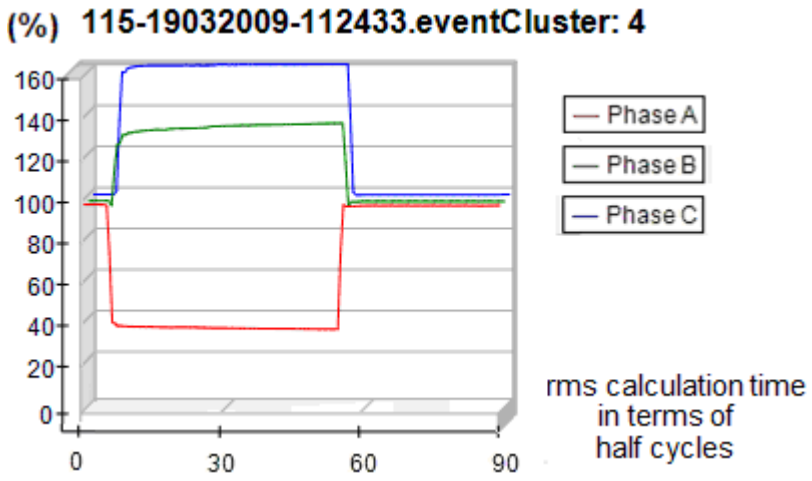


Figure 5. 12 Event belongs to Cluster 4 - Extremely

The examination of the applications of the fuzzy c-means clustering on power quality data may also represent the transformer substation characteristics as in Figure 5.13. The results define the long term characteristics of the transformer substations since the events are selected from a large period of time. By decreasing the time period, the characteristics for a defined period may also obtained.

Figure 5. 13 Transformer Substation Characteristics

5.4 DESIGN ISSUES FOR THE PQ EVENT CLUSTERING METHOD

Fuzzy c-means clustering method is designed to cluster the power quality events. There are approximately 1.060.000 power quality events to be considered. Feature vectors representing the events cannot be loaded into the memory simultaneously when the number of events to be clustered is increased, since the number of features for each event is also high. As a result the main issue in the design of the algorithm becomes solving the memory bound problem. In order to solve this problem, data is divided into chunks. Each of the chunks is formed by sampling from all of the selected feeders. The basic fuzzy c-means algorithm is modified to handle the data by iteratively applying the clustering to the chunks of data. The algorithm clusters individual chunks by the use of global cluster centroid and performance index calculations. The chunk based form of the fuzzy c-means algorithm solves the memory bound problem.

In order to decrease the running time of the algorithm, the flow of the algorithm should be investigated carefully. The chunk based clustering passes through each chunk twice. The first pass is for the initial clustering and the second pass is for refining the event cluster memberships. The pseudo code of the fuzzy c-means clustering algorithm for each pass is given below:

1. Initialize Membership Matrix $U^{(0)} = [u_{ik}]$ (Only for the First Pass)

2. For 1 to T (Maximum iteration number)

3. Calculate cluster centroid $C^{(s)} = [c_j]$ from Formula 5.4

4. Calculate the Objective Function from Formula 5.2

If the improvement over previous iteration is less than the termination threshold ɜ, go to Step 5 else quit

5. Calculate $U^{(s)} = [u_{ik}]$ from Formula 5.5, Return to step 2

Update Performance Index values for each cluster

The complexity of the $C^{(s)}$ calculation is $O(n*c*p)$ where n is the number of events, c is the number of clusters and p is the feature count of an event. The complexity of the $U^{(s)}$ calculation is $O(n*c^2*p)$. Thus the fuzzy c-means clustering algorithm is $O(n*c^2*p*T)$. By considering T as $O(n)$, the final complexity of the algorithm is $O(n^2*c^2*p)$. The $U^{(s)}$ calculation is the most time consuming part of the algorithm. Thus in order to reduce the run time of the fuzzy c-means clustering algorithm, $U^{(s)}$ calculation should be revised.

Eliminating the membership matrix is possible since membership matrix and the cluster centroid matrices are calculated from each other during the execution of the algorithm. Even if the Formula 5.5 contains U matrix access for the updates of C matrix, the numerator is the weighted sum of the feature values of the data point and the denominator is the sum of i^{th} row of membership matrix U. Both the numerator and the denominator calculations could be accumulated during the calculations for the membership matrix. After the calculations of the cluster center matrix the membership matrix rows are no longer required to be stored [23]. Thus the U matrix could be eliminated by modifying the algorithm flow. The complexity of the new matrix calculations is $O(n*c*p)$, thus the overall complexity of the algorithm is decreased to $O(n^2*c*p)$. The elimination of membership matrix calculation decreases both run time and memory requirement of the algorithm. The disadvantage of this modification and chunk based structure of the algorithm is that parallel processing is no longer applicable to be used for decreasing the run time.

The fuzzy c-means algorithm contains fractional exponentiation and complex distance measure calculations. The algorithm should also deal with the huge number of event features. As a result, the calculations in the algorithm are time consuming.

In order to decrease the computational complexity of the algorithm, the distance measure is designed for minimum computation. Manhattan Norm is used for the distance calculations.

The initial cluster center selection is another important point in the run time considerations. The iterations of the algorithm for each chunk stop when a local minimum or saddle point is achieved. The initial cluster centers should be selected to utilize fast convergence to the local minimum and saddle point. However the initial cluster selection algorithm requires an initial pass over the entire data set, thus implementing this algorithm does not reduce the run time directly.

The events data is stored as raw data on the disc. In order to avoid internet access to the data via database connection, the fuzzy c-means method is designed to run on the same storage with the raw data.

CHAPTER 6

CONCLUSION AND FUTURE WORK

The system described in this thesis uses exclusive and overlapping clustering models to design two clustering methods for the power quality data collected via online monitoring of the Turkish Electricity Transmission System. The exclusive clustering based method is a modified version of the basic k-means algorithm whereas the overlapping clustering based method is a modified version of the fuzzy c-means clustering algorithm. The developed methods reveal the intrinsic groupings and relations in the unlabeled power quality event data. The time and location based queries enable domain experts to examine the events and the event distribution for specific periods of time at specific locations. By the use of these queries and the domain knowledge; the characteristics of the transformer substations and the relations between the events can be obtained. The results of the clustering are used to define the classes of the events and the classification rules are inferred from the defined clusters and domain knowledge. The distribution of the events and the location of the initiating events are investigated for their possible reasons.

A power quality data mining management system is the outcome of this thesis. Important power quality data mining issues which are considered in this thesis can be listed as follows:

- The data mining algorithms are modified and a chunk based clustering method is proposed for clustering the power quality events. Chunk-based approach solves the memory bound problem; however it results in the requirement of integration problems.
- Performance issues are emphasized such that optimization strategies are applied in order to reduce the running time developed throughout this thesis.
- A widespread power quality monitoring system is integrated with the power quality data mining interface.
- The power quality parameters are investigated and data mining concepts are applied on the power quality events.
- Applying data mining on power quality events fulfills the requirement of defining causes of the events and relationships among various events. Transformer substations may be classified according to their long period characteristics by use of the proposed power quality data mining methods.

There are two main enhancements required to be applied on the system. The first enhancement is enlarging the application range. Currently, 154 national PQ monitors are installed on the transmission system, however in order to cover the whole Turkish Electricity Transmission System 2000 monitors should be installed. When this monitor number is achieved, the whole system could be monitored for the propagation of the problems. The effects of a single event to the remaining parts of the system could be revealed. The propagation of an event could be modeled, which enables future event predictions. For future work, an event prediction and load forecasting system could be developed.

The second enhancement should be applied on the event comparison algorithms. The designed algorithm should be pattern based. The comparison of two events should be independent of the magnitudes of event rms values, time and duration shifts in order to detect type similarities.

REFERENCES

[1] A. Asheibi, D. Stirling, D. Soetanto, "Analyzing Harmonic Monitoring Data Using Data Mining", 5th Australian Data Mining Conference, AusDM, 2006.

[2] P. K. Dash, I. L.W. Chun, M. V. Chilukuri, "Power Quality Data Mining Using Soft Computing and Wavelet Transform", Conference on Convergent Technologies for Asia-Pacific Region, TENCON 2003.

[3] A. Asheibi, D. Stirling, D. Robinson, "Identification of Load Power Quality Characteristics Using Data Mining", IEEE CCECE/CCGEI, 2006.

[4] O. N. Gerek, D. G. Ece, A. Barkana, "Covariance Analysis of Voltage Waveform Signature for Power-Quality Event Classification", IEEE Trans. on Power Delivery, vol. 21, no. 4, Oct. 2006.

[5] M. Wang, G. I. Rowe, A. V. Mamishev, "Classification of Power Quality Events Using Optimal Time-Frequency Representations – Part 2: Application", IEEE Trans. on Power Delivery, vol. 19, no. 3, July 2003.

[6] M. Uyar, S. Yildirim, M. T. Gencoglu, "An expert system based on S-transform and neural network for automatic classification of power quality disturbances", Expert Systems with Applications, Elsevier, 2008.

[7] Y. Liao, J. B. Lee, "A fuzzy expert system for classifying power quality disturbances", Electrical Power and Energy Systems, vol. 26, pp. 199-205, 2004.

[8] G.S. Hu, F.F. Zhu, Z. Ren, "Power quality disturbance identification using wavelet packet entropy and weighted support vector machines", Expert Systems with Applications, vol. 35, pp. 143-149, 2008.

[9] M. Uyar, S. Yildirim, M. T. Gencoglu, “An effective wavelet-based feature extraction method for classification of power quality disturbance signals”, Electric Power Systems Research, vol. 78, pp. 1747-1755, 2008.

[10] B. Biswal, P. K. dash, J. B. V. Reddy, “Power Signal Classification Using Dynamic Wavelet Network”, Applied Soft Computing, vol. 9, pp. 118-125, 2009.

[11] IEC Standard 61000-4-30, Electromagnetic Compatibility (EMC) – Part 4-30: Testing and Measurement Techniques – power Quality Measurement Methods.

[12] T. Demirci, A. Kalaycioglu, Ö. Salor, et al., “National PQ Monitoring Network for Turkish Electricity Transmission System”, IEEE Conference on Instrumentation and Measurement Technology, IMTC 2007.

[13] http://www.guckalitesi.gen.tr. 5 September 2009.

[14] Ö. Salor, S. Buhan, Ö. Ünsar, et al., “Mobile Monitoring System to Take Nationwide PQ Measurements on Electricity Transmission Systems”, the Measurement Journal of Elsevier, vol. 42, pp. 501-515, 2009.

[15] S. P. Lloyd, “Least Square Quantization in PCM”, IEEE Trans. Information Theory, vol. 28, 129-137, 1982.

[16] IEEE Std 1159-1995, Recommended Practice for Monitoring Electric Power Quality.

[17] Cheeseman, P. and J. Stutz, “Bayesian classification (AutoClass): Theory and results”. Advances in Knowledge Discovery and Data Mining. In U. Fayyad, G. Piatesky-Shapiro, P. Smyth Uthurusamy. AAAI Press (153-180), 1995.

[18] http://www.esri.com/software/mapobjects/index.html. 5 September 2009.

[19] http://xceed.com/?gclid=CMykoM_1mJwCFQ8TzAodEkJghQ. 5 September 2009.

[20] J. C. Dunn, A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well Separated Clusters, J. Cyber., 3, 32-57, 1974.

[21] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, NY, 1981.

[22] Liyan Zhang “Comparison of Fuzzy c-means Algorithm and New Fuzzy Clustering and Fuzzy Merging Algorithm”, Computer Science Department University of Nevada, Reno, 2001.

[23] F. Kolen and T. Hutcheson, Reducing the Time Complexity of the Fuzzy C-Means Algorithm, IEEE Transactions on Fuzzy Systems. V. 10, pp. 263–267, 2002.

[24] http://www.datamining.monash.edu.au/software/snob/. 5 September 2009.

[25] http://ti.arc.nasa.gov/project/autoclass/. 5 September 2009.

[26] http://www.the-data-mine.com/bin/view/Software/ClementineSoftware. 5 September 2009.

[27] http://www.postgresql.org/. 5 September 2009.

[28] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2003

[29] Barakbah and A. Helen, Optimized K-Means: An Algorithm of Initial Centroids Optimization for K-means. In: Seminar “Soft Computing, Intelligent Systems and Information Technology”, SIIT, 2005.

[30] Arthur, D. and Vassilvitskii, S. "K-means++: the advantages of careful seeding", Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027—1035, 2007

[31] M. Güder, N. K. Çiçekli, Ö. Salor, I. Çadırcı, “Clustering of Power Quality Event Data Collected via Monitoring Systems Installed on the Electricity Network, KDD, 2009.

[32] Paul E. Black, "Manhattan distance", in Dictionary of Algorithms and Data Structures [online], Paul E. Black, ed., U.S. National Institute of Standards and Technology. 5 September 2009.