

# An Application in SPSS Clementine Based on the Comparison of Association Algorithms in Data Mining

Seren Sezen Karalök<sup>1</sup>, Adnan Aktepe<sup>1</sup>, Süleyman Ersöz<sup>1</sup>

Accepted 3rd September 2016

**Abstract:** Data mining is the process of acquiring information from large data pools. In this study, association analysis method is used. The application and comparisons are found by using 3 different algorithms from SPSS Clementine which is a data mining software. In this study, the results are varied because different association methods are applied on. Therefore, new findings are obtained. Consequent to this, it will lead us to new strategies to develop for customers in Market Basket Analysis. This study is done by using a big supermarket data. Results are compared and reported for every each of 3 different algorithms.

**Keywords:** Data Mining, Association Analysis, SPSS Clementine, Market Basket Analysis.

## 1. Introduction

Currently, because of the rapid increase in amount of data in huge acceleration and becoming cheaper, many data are stored and used commonly in digital portals. Hence, it is a necessity to acquire meaningful and useful data from the existing mess of data.

Data mining is the process getting important, undiscovered, and uneasy to reach and usable data from data bases. Data mining is thought to be one of the most efficient method to get meaningful data and relations from the mess of data.

In literature, it can be easily seen that one of the commonly used method is association rules analysis. Association rule analysis enables us to see data which are often seen together.

In literature, supermarkets are places where association rules analysis is used very often. Association rules and consecutive time patterns that enable us to define purchase tendency, are used often under purchase oriented market basket analysis [1]. Customers' purchase habits are determined by using the analysis of the Market Basket method by observing items bought together often.

If we look at the literature, in supermarkets there are lots of different strategies developed depending on the association rules analysis of large data. The main property of this study what makes it unique is that the study is based on real data from a big-scale supermarket receipts and the data are analyzed according to 3 different algorithms.

If we look at the literature;

Ay, Çil (2010) In this study, a decision-making support mechanism for deciding the location of supermarket based on data gradient is provided. The study provides us methodological frame which enables us to determine location by using information discovery process at data bases. In the study, after preparing relative data bases, Apriori algorithm and Multi-

Dimensional Scaling methods are used. Experimental study is conducted on one of the leading retail sale company Migros Turk Inc[2].

Alagoz, Oge, Asilkan (2012) In this study, relationships between data mining which is a corporation intelligence system and accounting information system are examined as well with the crossing points of these 2 systems. This study consists of definition of data mining, introduction of applications areas and its methods and provides a general concept of accounting information system, data-info flow and reports [3].

Irmak, Koksall, Asilkan (2012) In this study, some major data mining techniques are applied on an in-use hospital database and thanks to applications patient intensity (density) predictions and comparisons are made[4].

Baykal (2006) Some researches are made on data mining [5].

Albayrak, Yılmaz (2009) Decision-tree method which is one of the data mining technique is applied on data from 173 companies in IMKB 100 index which are active in industrial and service sectors according to their financial indicators between 2004-2006 [6].

Iceli (2012) The main purpose of this article is transforming raw data into valuable information by using data mining on the data collected from surveys conducted on students. Relative to the purpose, a survey is conducted on students from Cumhuriyet University NureiDemirag Collage Fundamental Computer Sciences 2008-2010 class students. The credibility of survey is approved by SPSS packet program analysis [7].

Timör, Şimşek (2008) In this study, Studies made on medical data mining is explained and a brief explanation of a data mining Project which will be applied on Hacettepe Hospital is given [8].

Yıldırım, Uludağ, Görür (2008) In this study, Studies made on medical data mining is explained and a brief explanation of a data mining Project which will be applied on Hacettepe Hospital is given [9].

Liao ve Chen (2004) presented their product maps which is derived from Apriori algorithm as a new product development source and used association rules to make electronic catalog marketing and in discount management [10].

<sup>1</sup> I Kirikkale University Faculty Of Engineering Department Of Industrial Engineering, Turkey

\* Corresponding Author: Email: aaktepe@gmail.com

Note: This paper has been presented at the 3<sup>rd</sup> International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey),

Ekim (2011) In this article, association rules are applied on data from in-use system at student's registrars Office to make predictions on students future. Apriori algorithm and decision-tree algorithms are used to realize this purpose. By these rules, the factors which are affect on student's success are investigated [11].

Atılğan (2011) tried to investigate one of the biggest real life problem, traffic accidents by treating data mining as extensive as much possible. In the application part of the study, the factors affecting on drivers and pedestrians are investigated. The investigation is done through decision tree method and association rules regarding the simplicity in understanding [12].

İnce, Alan (2014) wanted to make a strategic contribution to companies' product portfolio definition and creation process by using data mining methods. Association Rules method, one of the data mining method, is applied on data from a company which does. fund transactions. The investigation of whether there is a association rules when rival company buys funds and it is realized that there are 52 rules by using Tertius algorithm. A suggestion is made to companies that they can benefit from these association rules at production portfolio planning [13].

Kose (2015) determined relationships between product catalog and selling by using association rules and clustering analysis and made a sample application on national retailer [14].

Dogrulvd. (2015) made studies on extracting hidden information from enormous data of past accidents by using Apriori and GRI algorithms [15].

Kırtay, Ekmekçi, Halıcı, Ketenci, Aktaş, Kalıpsız. In order to increase performance of production recommendation system, sampling process is added to Market Basket Analysis. By this way, instead of universe but a set of lesser number of observant representing universe is used in analysis which enabling time efficiency of analysis time and efficiency on memory usage [16].

Gulce (2010) Terms related with Data Mining and especially apriori algorithm which generates association rules to be used in market basket analysis is examined in a very detailed way. Apriori algorithm is applied on a different data set rather than on market basket analysis set. In this application, data base conversion process is conducted [17].

Ulaş, Alpaydın A future promising results are obtained after applying market basket analysis on sales data obtained from Gima Turk Inc.'s various stores [18].

Durdu (2012), in his study, a structure is developed which can be a basement for activities of management of customer relationship by using data mining tools and applications. In the study customer main data and sales operations are transformed into usable and valuable data that can be used in management of customer relations. In this context, a market basket analysis is conducted and an application is developed an application which gives association rules those are determined by using apriori algorithm from market data set [19].

Ceylan (2014) in his study, in order to reveal the relationship between the drugs which tend to be purchased together on a prescription of a pharmacist apply the Cohesion Rule approach from data mining techniques. The analysis was performed by applying the Apriori algorithm from the association rule algorithms. The aim of the study is to propose a new drug-shelf scheme with the rules obtained. With the proposed new shelf facility, it was aimed to use the pharmacy area effectively, to provide services to the customers in a shorter time, and to reduce possible drug mistakes [20].

Bayram(2014), a company that conducts market research around the world, the panel in Turkey used a traditional dataset using

traditional and modern channels, and a basket analysis with the rules of association between the product groups purchased in these markets. The application was determined by the apriori algorithm in the SPSS Clementine 12.0 package program and the results were interpreted [21].

## 2. Proposed Methodology

### 2.1. Market Basket Analysis with Association Rules

Association rules are very important in Data Mining. Main purpose of the method is obtaining usable various rules from big data bases. Association rules are generally used to derive rules from mutually related, concurrently occurring data.

There are steps should be taken before using association rules;

- Finding the repetition number of repetitive numbers,
- Determining minimum support and minimum trust values,
- Tagging as frequent couplings those are repeated as much as minimum support number,
- Extracting strong association rules from very often repeated items,
- Comparing trust abed support values of extracted association rules to minimum and support values,
- Checking if these values of rules to satisfy minimum values [22]

If there are N numbers of item, there may exist  $2^N$  number of often seen set of items. Association rules are determined after finding often seen sets. Finding associtaion rules process is finding which item has effects on which other items [23].

Supermarkets are the places where association rules are used very often. A lot of applications can be seen when it 's looked it ap at literature and it is known as market basket analysis. Market Basket Analysis plays a very important role in customers' habits, sales strategy, stocks control with its help in determining the products which are tendent to be bought together (Ay veCil, 2009).

Market Basket Analysis indicates variety of distinguishing distribution by showing different perspective of customers. This distribution information has a credible proportion in making decision of planning, advertisement design, discount-promotion, facility location and product investment. (Yang ve Lai,2006)

### 2.2. Algorithms Used For Associtaion Rules

Algotiyhms which are used for obtaining information from data mess can be classified as consecutive and parallel. Consecutive algorithms contains logical statements those are forming and counting product sets. Parallel algorithms makes large product sets by creating parallelism. (Erpolat, 2012) [24].

#### 2.2.1. GRI Algorithm

Generalized Rules Induction (GRI) discovers node association rules at data. GRI picks over the highest information content by considering generality (support) and confidence rules. GRI can work with numerical and categorical in outs but the objective should be categorical. (Clementine Users Guide, 2007) [25].

#### 2.2.2. Apriori Algorithm

Being developed Agrawal and his friends Apriori Algorithms provide great benefits to be achieved during the development of

the association rules of data mining. Because of this reason, Apriori Algorithms became the most popular algorithm in application of Association Rules. Name of the algorithm "Apriori" derived from "prior" because the algorithms system continue to work based on prior step [26].

Apriori node extracts rules which contain large information then, picks over set of rules. Apriori provides 5 different methods to pick over rules and uses sophisticated induction schematic to process efficient large data sets. Apriori is faster than Gri for bigger problems. Apriori requires input and output zones to be categorical because it is optimized fort his type of data. (Clementine Users Guide,2007) [25].

In Market Basket Analysis problems to identify the relation between products on sale there 2 criteria used which are "support" and "confidence/trust". "Rules Support Criterion" identifies in one relation what is the proportion of repetition to every shopping. "Rules Trust Criterion" identifies what is the probability of someone buying B product who bought A product already [24].

### 2.2.3. CARMA Algorithm

CARMA makes the calculation of small-scale farms online (Hidb, 1999). CARMA displays the existing association rules online to the user and allows the user to change the minimal support and minimum trust parameters in any operation of the first scan of the database. CARMA constitutes a set of objects as they pass through the movements. After reading each movement, it first increases the numbers of the objections of the sub-clusters of the movement.

Then, if all of the existing subclasses of the object instance provide the minimum support value, and if they are larger than the read-out of the database, the object instances from the movement are created. An upper bound on the number of objects is computed so that the probability that a object is likely to be extreme can be precisely predicted. This is the sum of the current number and the estimate of its occurrence before the object instance is created. Estimating the probability of occurrence (maximum leaks) is calculated when the node is first created [24].

The CARMA model extracts a set of rules from the data without having to guess and target fields. In contrast to "Apriori" and "GRI", CARMA node provides only preliminary support instead of structure settings for support rule (support for the premise and consequent). At this point, these rules can be used for a variety of applications in a wider area. (Clementine Users Guide, 2007) [25].

## 3. Aplication

The study was conducted using one of the data mining software SPSS Clementine.

### 3.1. Data Processing

The data received from the company has been passed through certain steps in order to implement the data mining application. These steps are;

**Data collecting:** Selection of appropriate data sets to be used in the data mining study.

**Merge and cleaning:** In this step, the noisy data that caused the wrong results are cleared and the differences are eliminated.

**Data conversion:** In this step, the data is transformed into a data mining application. For GRI, Apriori and CARMA algorithms to

be used in data analysis, the display format and assignments of data have been changed.

### 3.2. Algorithm Selection

After the data processing step has been performed and the data has been adjusted, the GRI, Apriori and CARMA algorithms in the data mining software SPSS Clementine have been selected for the study.

### 3.3. Modelling

In the study, GRI, Apriori and CARMA algorithms, which are the data mining software, 3 algorithms of SPSS Clementine software, are used. The prepared model is solved separately for 3 algorithms. The association rules of the model created in the program shown in Figure 1. For modeling; 8 product groups were selected from the supermarket to be implemented. Figure 2 shows the assignment of 8 product groups to the model. In order to work correctly with SPSS Clementine software, 100 pieces from the supermarket were used. The basic principle of using 3 different algorithms in the study is to achieve more accurate results. Thus, the right strategies for the firm will be realized.

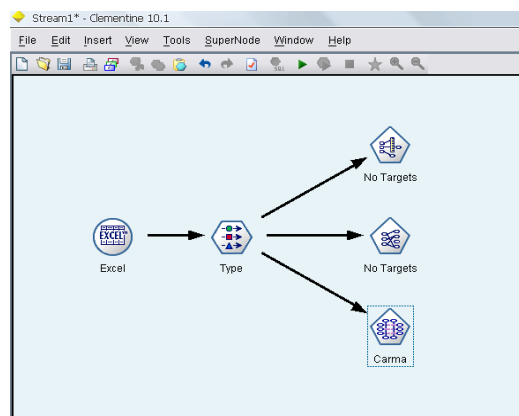


Fig.1.Association Module

Field	Type	Values	Missing	Check	Direction
bakery	Flag	1.0/0.0	None		Both
newspaper	Flag	1.0/0.0	None		Both
cheese	Flag	1.0/0.0	None		Both
egg	Flag	1.0/0.0	None		Both
greengrocer	Flag	1.0/0.0	None		Both
detergent	Flag	1.0/0.0	None		Both
delicatessen	Flag	1.0/0.0	None		Both
dairy products	Flag	1.0/0.0	None		Both

Fig.2. Configuration in SPSS Clementine 1

## 4. Result and Discussion

For the sales strategy of a large-scale supermarket, a study of market basket analysis was conducted using data mining rules. In the SPSS Clementine software, 100 previously received chips were evaluated separately for 3 algorithms. The results of the association analysis for GRI, Apriori and CARMA algorithms are given in Table 1, Table 2, and Table 3 below.

For GRI algorithm which is the first among the algorithms, the results are shown below.

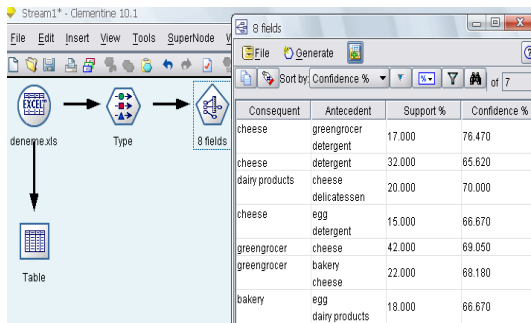


Fig.3.Gri Model

Table 1.RESULTS FOR GRI ALGORITHM

Consequent	Antecedent	Support	Confidence
Cheese	Greengrocer Detergent	17,000	76,470
Cheese	Detergent	32,000	65,620
Dairy products	Cheese Delicatessen	20,000	70,000
Cheese	Egg Detergent	15,000	66,670
Greengrocer	Cheese	42,000	69,050
Bakery	Egg Dairy products	18,000	66,670
Greengrocer	Bakery Cheese	22,000	68,180

GRI Algorithm based results are in Table I;

- a customer who buys product greengrocer and detergent, has 76.47% probability for buying product cheese. The possibility of coexistence of these products in shopping vouchers is 17 %.

The results for Apriori algorithm are shown below.

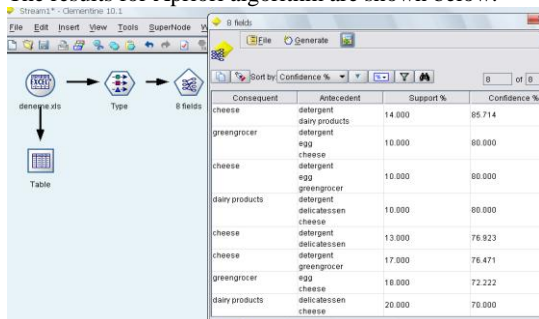


Fig.2.Apriori Model

Table 2.RESULTS FOR APRIORI ALGORITHM

Consequent	Antecedent	Support	Confidence
Cheese	Detergent Dairy products	14,000	85,714
Greengrocer	Egg Cheese	32,000	65,620
Dairy products	Delicatessen Detergent	10,000	80,000
Cheese	Detergent Delicatessen	13,000	76,923
Cheese	Detergent Greengrocer	17,000	76,471
Greengrocer	Egg Cheese	18,000	72,222
Dairy products	Delicatessen Cheese	20,000	70,000

Apriori Algorithm based results are in Table II;

- a customer who buys product dairy products and detergent, has 85,714% probability for buying product cheese. The possibility of coexistence of these products in shopping vouchers is 14 %.

The results for CARMA algorithm are shown below.

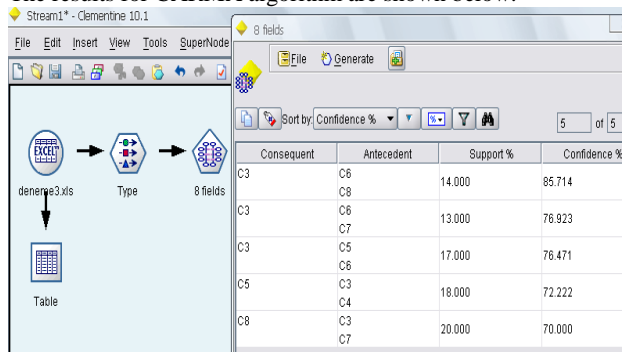


Fig.3. CARMA model

Table 3.RESULTS FOR CARMA ALGORITHM

Consequent	Antecedent	Support	Confidence
Cheese	Detergent Dairy products	14,000	85,714
Cheese	Detergent Delicatessen	13,000	76,923
Cheese	Greengrocer Detergent	17,000	76,471
Greengrocer	Cheese Egg	18,000	72,222
Dairy products	Cheese Delicatessen	20,000	70,000

CARMA Algorithm based results are in Table II;

- a customer who buys product detergent and dairy products, has 85,714% probability for buying product cheese. The possibility of coexistence of these products in shopping vouchers is 14 %.

We can list the associations for 3 algorithms as follows:

Table 4.RESULTS OF INTERSECTION

GRI	APRIORI	CARMA
Greengrocer Detergent → Cheese	Detergent Dairy products → Cheese	Detergent Dairy products → Cheese
Detergent → Cheese	Detergent Egg Cheese → Greengrocer	Detergent Delicatessen → Cheese
Cheese Delicatessen → Dairy products	Cheese Delicatessen Detergent → Dairy products	Egg Cheese → Greengrocer
Egg Detergent → Cheese	Detergent Delicatessen → Cheese	Cheese Delicatessen → Dairy products
Cheese → Greengrocer	Greengrocer Detergent → Cheese	
Egg Dairy products → Bakery	Egg Cheese → Greengrocer	
Bakery Cheese → Greengrocer	Cheese Delicatessen → Dairy products	

If we will reach common rules from the results of the

relationship;

- Customers who purchase greengrocer and detergent product groups are buying cheese product groups the min confidence value above 70,0% probability. The min support value above 17,0% probability, products are available in shopping vouchers.

- Customers who purchase cheese and delicatessen product groups are buying dairy products the min confidence value above 70,0% probability. The min support value above 20,0% probability, products are available in shopping vouchers.

In this study, GRI, Apriori and CARMA algorithms, which are the 3-way rule algorithm of SPSS Clementine software, are used. As a result of the implementation, each algorithm gave different results. Alliances have been taken together.

At the next stages of the work, new plant layout arrangements will be made in line with the cooperation provided by the software.

## References

- [1] Döşlü, A. (2008). Veri Madenciliğinde Market Sepet Analizi ve Birliktelik Kurallarının Belirlenmesi. Yayınlanmamış Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, İstanbul.
- [2] Ay, D., ve Çil, İ., The Use Of Association Rules in Store Layout Planning at Migros Türk A.Ş.. Endüstri Mühendisliği Dergisi, 21(2), 14-29, 2008.
- [3] ALAGÖZ, Ali, Serdar ÖGE, and Metehan ORTAKARPUZ (2014) " The Relationship of Data Mining, as a Business Intelligence Technology, with the Accounting Information System." *Selçuk University Social Sciences Institute Journal* 32.
- [4] Irmak, Sezgin, Can Deniz Köksal, and Özcan Asilkan (2012). " Predicting Future Patient Volumes of The Hospitals By Using Data Mining Methods." *International Journal of Alanya Faculty of Business*, 4.1.
- [5] BAYKAL, Abdullah. (2006), " Application Fields of Data Mining." *DÜ Ziya Gökalp Eğitim Fakültesi Dergisi* 7 : 95-107.
- [6] Albayrak A., Yılmaz Ş. (2009), "DATA MINING: DECISION TREE ALGORITHMS AND AN APPLICATION ON ISE DATA" ,Suleyman Demirel University The Journal of Faculty of Economics and Administrative Sciences, C.14, S.1 s.31-52.
- [7] İçeli N. (2012), Veri Madenciliği Yöntemi ile Divriği Nuri Demirdağ Meslek Yüksekokulu Öğrencilerinin Temel Bilgisayar Dersine Ait Başarı Analizi Uygulaması" *MBD*, 1(1): 18 – 37
- [8] Timor, M., and U. T. Şimşek. (2008), "Veri Madenciliğinde Sepet Analizi ile Tüketici Davranış Modellemesi." *Yönetim*, 19 (59): 3-10.
- [9] YILDIRIM, Pınar, Mahmut ULUDAĞ, and Abdülkadir GÖRÜR (2008), "Hastane Bilgi Sistemlerinde Veri Madenciliği." *Çanakkale On Sekiz Mart Üniversitesi Akademik Bilişim*.
- [10] Liao, Shu-Hsien, and Yin-Ju Chen. (2004), "Mining customer knowledge for electronic catalog marketing." *Expert Systems with Applications* 27.4: 521-532.
- [11] Ekim U. (2011), "EDUCING OF ASSOCIATION RULES FROM STUDENTS DATABASE USING DATA MINING ALGORITHMS" , Selçuk Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı.
- [12] ATILGAN, E. (2011), "Karayollarında Meydana Gelen Trafik Kazalarının Karar Ağaçları ve Birliktelik Analizi İle İncelenmesi." *Hacettepe Üniversitesi İstatistik Anabilim Dalı Yayınlanmamış Yüksek Lisans Tezi*
- [13] İnce, Ali Rıza, and Mehmet Ali Alan. (2014), " A STUDY OF UTILIZING DATA MINING ON PRODUCT PORTFOLIO PLANNING " *LAÜ Sosyal Bilimler Dergisi* 5.2.
- [14] Doğrul G., Akay D., Kurt M., (2015). " ANALYSIS OF TRAFFIC ACCIDENTS BY RULES OF ASSOCIATION", syf. 265-284
- [15] Köse, Y. (2015), "Değerli Müşterilerde Ürün Kategorileri Arasındaki Satış İlişkilerinin Veri Madenciliği Yöntemlerinden Birliktelik Kuralları ve Kümeleme Analizi İle Belirlenmesi ve Ulusal Bir Perakendecide Örnek Uygulama", Yüksek Lisans Tezi, Selçuk Üniversitesi Sosyal Bilimler Enstitüsü, İşletme Ana Bilim Dalı.
- [16] Kırtay, S. H., Ekmekçi, N., Halıcı, T., Ketenci, U., Aktas, M. S., & Kalıpsız, O., "Pazar Sepeti Analizi için Örneklem Olusturulması ve Birliktelik Kurallarının Çıkarılması", *Bilgisayar Mühendisliği Bölümü, Elektrik-Elektronik Fakültesi Yıldız Teknik Üniversitesi, İstanbul AR-GE Merkezi, Cybersoft, İstanbul*
- [17] Gülce A. (2010), "Veri Madenciliğinde Apriori Algoritması ve Apriori Algoritmasının Farklı Veri Kümelerinde Uygulanması" , *Bilgisayar Mühendisliği Ana Bilim Dalı, Trakya Üniversitesi Fen Bilimleri Enstitüsü.*
- [18] ULAŞ M.A., ALPAYDIN E., SÖNMEZ N., ve KALKAN, A. (2001), " Market Basket Analysis for Data Mining", *Bilişim Zirvesi 2001, TBD 18. Bilişim Kurultayı, 4-7 Eylül, İstanbul.*
- [19] DURDU, M. (2012). Application of data mining in customer relationship management market basket analysis in a retailer store (Doctoral dissertation, DEÜ Fen Bilimleri Enstitüsü).
- [20] Ceylan Z. (2014), "Veri Madenciliği Teknikleri ile Raf Düzenleme ve Bir Vaka Çalışması.", Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, Endüstri Mühendisliği Ana Bilim Dalı.
- [21] Bayram, O. (2014), "Birliktelik Analizi ve Bir Uygulaması", *Mimar Sinan Güzel Sanatlar Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı.*
- [22] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.. Morgan Kaufmann Publishers, San Francisco.
- [23] Shekhar, S., Vatsavai, R. R., & Celik, M. (2008). *Spatial and spatiotemporal data mining: Recent advances. Data Mining: Next Generation Challenges and Future Directions.*
- [24] Erpolat, S. (2012). Comparison of Apriori and FP-Growth Algorithms on Determination of Association Rules in Authorized Automobile Service Centres, *Cilt/Vol.: 12 - Sayı/No: 1 (151-166).*
- [25] Clementine® 12.0 User's Guide (2007)