Araştırma Yazısı
Selçuk Tıp Derg 2014;30(3): 97-103

SELÇUK TIP DERGİSİ

Meme Kanseri Verisinde APRIORI Algoritması ile Kural Çıkarma

Rule Induction by Apriori Algorithm Using Breast Cancer Data

[1]Adnan Karaibrahimoğlu, [2]Aşır Genç

[1]NE University, Meram Faculty of Medicine Biostatistics, Konya
[2]Selçuk University, Science Faculty Statistics, Konya

Özet

Teknoloji ile birlikte yaşamın her alanında artan veri miktarı "veri ambarları" kavramını gündeme getirmiştir. Veri madenciliği, ortaya çıkan çok büyük veri kümelerinin oluşturduğu veri ambarlarının analiz edilerek yararlı bilgiler elde edilmesini sağlayan yaklaşımlar bütünüdür. Veri miktarının büyük olduğu ve her geçen gün arttığı alanlardan birisi de sağlık sektörüdür. Her gün binlerce hastaya ait gerek kişisel gerek tıbbi veriler kayıt altına alınmakta ve bu enformasyon depolanmaktadır. Ancak bu verilerin çok az bir kısmı analiz edilebilmekte ve geriye kalan kısmından faydalı olabilecek enformasyon elde edilememektedir. Özellikle hastane yönetim sistemleri, tedavi yöntemleri ve koruyucu hekimlik konusunda maliyetleri azaltıcı yöntemlerin geliştirilmesi için ambardaki verilerin analiz edilmesi gerekmektedir. Klasik istatistiksel yöntemler ile büyük veri kümelerini analiz etmek zor olduğu için, çeşitli veri madenciliği yöntemleri geliştirilmiş ve bilgisayar programcılığı yardımıyla analiz yapmak daha uygulanabilir hale gelmiştir. Birliktelik kuralı, sağlık alanında yeni kullanılan analiz yöntemlerinden birisi olup; değişkenlerin birlikte görülme olasılıkları üzerinden örüntü oluşturmak ve buna bağlı olarak destek ve güven değerlerini hesaplamak için kullanılmaktadır. Bu çalışmada, Meram Tıp Fakültesi Onkoloji Hastanesine ait retrospektif çalışma sonucu elde edilen meme kanseri verileri üzerinde APRIORI algoritması uygulanmış ve verilerdeki birliktelik örüntüleri ortaya çıkarılmaya çalışılmıştır.

Anahtar kelimeler: Veri madenciliği, Birliktelik kuralı, Apriori algoritması, Destek, Güven

Abstract

The amount of data, increasing together with the technology, has brought the concept of "data warehouse" in every field of life. Data Mining is a set of approaches analyzing these data warehouses formed by very large data sets and allows to gather useful information. One of the fields where the amount of data is large and getting larger everyday is the health sector. Many personal and medical data belonging to thousands of patients are recorded and stored. However, small part of these data can be analyzed and the remaining part may not be helpful to obtain useful information. The data in warehouses must be analyzed to improve the methods for hospital management systems, treatment and health care systems to reduce the costs. Since analyzing large data sets using classical statistical methods is difficult, various data mining methods have been developed and these methods have become more feasible with the help of certain softwares. Association rule is an important data-mining task to find hidden patterns between the variables and used recently in the field of healthcare. In this study, we have calculated the support and confidence of the associations in data set. APRIORI algorithm have been applied onto the retrospectively obtained breast cancer data belonging to Oncology Hospital of Meram Faculty of Medicine.

Key words: Apriori Algorithm, Association Rule, Confidence, Data Mining, Support

INTRODUCTION

Two important words, "information" and "digital", have lately entered to our life. Information theory was developed by Shannon-Weaver in 1948 and the term "digital" was used firstly to refer the fast electric pulses by Stibitz (1942). Digital data gathered by technological devices is getting larger and considered as "information". This information must be processed and converted to "knowledge" to utilize. Different knowledge belonging a customer is a record and a lot of records constitute a database. If a database is arranged according to subject-oriented, belongs to a certain period and enterprise, then it is called a "data warehouse". Analyzing and querying large data in warehouse to decide certain strategies and reporting the results is called OLAP (On Line Analytical Processing). OLAP is different from usual queries. For example, querying of weekly milk sale is not OLAP whereas the probability of exceeding of ten thousand milk sale concerns OLAP. Companies, who want to be successful in their business, have a strong demand on knowledge about the production, the suppliers, the market and the customers. For instance, many companies, today, have businesses and trades all around the world. They are supplying raw material or purchasing different products that is importing and exporting. Therefore they have much information related with these affairs. Knowing and following of all these things by a single accountant staff are impossible. That's the reason why the processing of the information has to be done by computers (1). In dealing with large data sets we have faced some problems such as storing, security and analyzing of big amount of data. Therefore many computer scientists and statisticians have tried to develop new methods for these problems. At the beginning, the concept of Exploratory Data Analysis (EDA) is developed by Tukey (1977). Later on, the concept of Knowledge Discovery from Data Mining (KDD) is used to analyze large data sets. Data mining (DM) is a process to recognize unknown and usable patterns between the variables in a data set. There are many different definitions of data mining. The important concepts in data mining are that the data set must be large and the relations must be unknown. In various sectors, excursive data mining methods had been applied, but those had caused some problems with standardization. Then an approach of standardization "Cross Industry Standard Process

Yazışma Adresi: Adnan Karaibrahimoğlu, Necmettin Erbakan Üniversitesi, MeramTıpFakültesi Tebad Anabilim Dalı, Konya, Türkiye
e posta: akara@konya.edu.tr
Geliş Tarihi: 05.05.2014 Yayına Kabul Tarihi: 16.06.2014

for Data Mining (CRISP-DM)" had been developed (2). CRISP-DM cycle has six steps:

1- Business Understanding
2- Data Understanding
3- Data Preparation
4- Modeling
5- Evaluation
6- Deployment

All DM methods are based on data specifications and their presentation. The secret word in DM is "data quality". Because bad data quality means wrong decisions that cause to lose the way of success. If a data is not ready to present, then the analyzer should take measure to predispose the data. Missing, inconsistent, unnecessary or noisy data problem is a part of preparation step. It is very important to solve the data cleaning and missing value problems. Data cleaning is a difficult and time-taking process. There are several methods for dealing with data cleaning. By means of these methods, such data should be converted to a usable form. However, in some cases, a set may have complicated data and they cannot be converted to ordinal or categorical form. If a data set is prepared without a problem, then the model is set well and meets the expectations and needs of the business (3). In deployment step, the results and reports are presented also by various types of graphs. Visualization is a crucial step to use profitably the knowledge. Many tools have been introduced in literature to visualize the usual result statements. In addition to that, some types of graphs have been developed for item visualization, rules visualization and conjoint visualization. 2- or 3-D matrix representation, two-key plot, double-decker plot, parallel coordinates or factorial planes are some examples of new types of graphs (4, 5).

Data mining techniques can be successful by computerized technology since they should analyze large data. The instructions that are formed to analyze and imitate human thought mechanism are called "artificial intelligence". If a computer can supply a tangible movement, then it is called machine learning. Data mining methods are separated into two categories on machine learning base. First one is the supervised learning in which predefined and preknown variables are used to make an induction. Classification is the general form of supervised learning and is a task that occurs very frequently in everyday life. In this method, we divide up the objects so that each is assigned to one of a number of mutually exclusive categories known as classes. If a classification rule has no certainty but has a probability, then this kind of classification is called Naive-Bayes Classifiers. When all attribute values are numerical and if we want to represent all attributes in a classification that we defined in a distance sense, then this method is k-nearest neighbor classification. Therefore Manhattan, Minkowski or Euclidean distances are very important manner in classification. Decision trees, support vector machines or artificial neural networks are also methods of classification. Unsupervised learning is second type of machine learning, in which there is no predefined variable. Related algorithm tries to find the hidden relations and to extract patterns between the variables. Clustering is the general name for unsupervised learning and there are two main methods: partitioning and agglomerative. K-means, k-medoids, cohonen networks, density based partitioning and grid based partitioning are examples of clustering methods. Data mining is a long, expensive and uneasy process. Therefore it has been developing in commercial axis and the applications in public sector are rare. Data mining is widely used in many areas such as economy, finance, healthcare, communication lines, text mining, fraud detection, insurance (2).

MATERIALS AND METHODS

Association Rule Mining

Many of data miners think that association rule (AR) is an unsupervised learning method but some of them differentiate that AR is neither supervised nor unsupervised learning method since its algorithm is different from other methods. AR is one of the first methods of DM and has been firstly defined by Agrawal as an algorithm for customer data. Many other computer programmers have been developed this method. AR algorithms have been emerged as "market basket analysis". After using barcode scanners in supermarkets, many processes in trade affairs have been started to monitor electronically. Thus the businesses have huge digital data about the customers or suppliers. Then they thought that if they use the affinity information about the customers, they could enlarge the sale volume. Affinity analysis means to find the patterns from the customer data. But how do we find such patterns? These types of data usually consist of records about customers. Each record consists of items in a basket which is really purchased. This set is called the frequent itemset, the others are infrequent itemset. These itemsets are in form of n×p matrix. Such matrices are very large. The number of rows, n can be in millions and the number of columns, p can be in thousands. Therefore classical statistical methods are weak to analyze many variables simultaneously. An association rule is a simple probabilistic about the co-occurrance of certain events in a database (6, 7, 8). Its principle is very simple. We write a basic sentence using "IF …… THEN…..". For example, "IF a customer buys "bread", THEN he also buys "milk"". That is, "IF X, THEN Ywith the probablity p". Therefore we state the following premise:

$X \rightarrow Y$ with p (2.1)

In above statement, X is called antecedent and Y is called consequent of the rule. The aim of the algorithm is to find all rules satisfying the constraint that the accuracy p is greater than some minimum p. If there are a few items in dataset, then the procedure is quite simple. But if there are thousands of items, then the procedure becomes very complicated to handle. AR is formally stated as follows:

Let $D=\{T_1,T_2,\ldots,T_n\}$ be a set of n transactions and $I=\{i_1,i_2,\ldots,i_m\}$ an itemset. Each T_i is a set of items, $T_i \subseteq \mathbf{I}$. Implications of $X \rightarrow Y$; where $X, Y < \mathbf{I}$ and $X \cap Y = \Phi$ is the association rule.

Let p(X) and p(Y) be the probabilities of itemsets appearing in I and p(Y|X) be the conditional probablity of Y given X. We want to determine the rules $X \rightarrow Y$ satisfying p(Y|X). In order to decide whether a rule is interesting or not, association rule mining has two metrics: support and confidence. Support is defined as the fraction of transactions of X and Y together to all transactions in database.

$\text{support } (X \rightarrow Y) = |XUY|/|DB| = p(XUY)$ (2.2)

Confidence is defined as the fraction of transactions of X and Y together to X itemset.

$\text{confidence } (X \rightarrow Y) = (p(XUY))/(p(X)) = p(Y|X)$ (2.3)

where |DB| denotes the number of elements of all database.

Both support and confidence represent the positive correlation and take the value between 0 and 1. The nearer the value to 1, the more interesting the rule. Interestingness is the indicator of power of a rule (9, 10). An AR can be one dimensional as $X \rightarrow Y$, where X and Y represent only one item. However a rule can be multi dimensional like $x_i\, x_2 \ldots, x_j \rightarrow Y$, where $x_i < X$. In second case, the rules can be explored by OLAP (On-Line Analytical Processing) cubes. For instance, we might be interested in the rules about the customers from which branch, in which day and what have bought. Let's see some examples of one- and multi- dimensional rule tasks:

Table 1. Transactions of example dataset

TID	x_1	x_2	x_3
100	1	1	0
200	1	1	1
300	0	1	0
400	1	1	0
500	0	1	0
600	1	0	0
700	1	0	1
800	1	1	1
900	0	0	0
1000	0	0	0

a- Age (customer, "20-30") buy (customer, "LCD TV") [support= %2, confidence= %13]
b- Age (customer, "20-30")→Sex (customer, "man") buy (customer, "LCD TV")
[support = %1, confidence = %60]
c- Age (customer, "20-30")→Sex(customer, "man")→buy (customer, "LCD TV") buy (customer, "DVD Player") [support= %1, confidence= %65]
Let's give an example of binary database to understand the conditional probability. Suppose that we have three variables with ten transactions x_1 x_2 and x_3 of which x_1 is consequent and x_2, $x_3$are antecedents. Let the variables be binary (dichotomous) values and have the following entries:

We compute all conditional probabilities from the table. The results are not but the confidences of the rules. Considering three variables, the interesting rules are f, g, p and r which have the confidence of 20%. The main idea in the rule induction is just that procedure. Data miners think that if the values of support and confidence are great enough, then the rules are strong. However, this proposition cannot be always correct. Because the probability of the antecedent should be smaller than the probability of the consequent in a rule, so the antecedent contributes to the consequent. This statement is described by a measure called Lift. It is third important metrics of AR and used to implicate the ratio of effect of antecedent on consequent. Let A be rule of X→Y, then Lift measure is
lift(A)= (Confidence(A)) / (Support(Y))= (p(X→Y)) / (p(X)·p(Y)) (2.4)
lift(X→Y) can be classified into three cases by Piatetsky-Shapiro's argument as follows:
1- If lift=1, then X and Y are independent,
2- If lift>1, then Y is positively dependent on X,
3- If lift<1, then Y is negatively dependent on X.
As a result of this argument, we can state the following constraints:

Table 2. Conditional probabilities of the transactions belonging to Table.1

a. $p(x_1=1)=0,6$	j. $p(x_1=0)=0,4$
b. $p(x_1=1\|x_2=1)=0,4$	k. $p(x_1=0\|x_2=0)=0,2$
c. $p(x_1=1\|x_3=1)=0,3$	l. $p(x_1=0\|x_3=0)=0,4$
d. $p(x_1=1\|x_2=0)=0,2$	m. $p(x_1=0\|x_2=1)=0,2$
e. $p(x_1=1\|x_3=0)=0,3$	n. $p(x_1=0\|x_3=1)=0,0$
f. $p(x_1=1\|x_2=1, x_3=1)=0,2$	o. $p(x_1=0\|x_2=1, x_3=1)=0,0$
g. $p(x_1=1\|x_2=1, x_3=0)=0,2$	p. $p(x_1=0\|x_2=1, x_3=0)=0,2$
h. $p(x_1=1\|x_2=0, x_3=1)=0,1$	q. $p(x_1=0\|x_2=0, x_3=1)=0,0$
i. $p(x_1=1\|x_2=0, x_3=0)=0,1$	r. $p(x_1=0\|x_2=0, x_3=0)=0,2$

Let min.sup, min.conf and min.lift >0 are given by experts for a case in a database. If
a- $p(X\rightarrow Y)\geq$min.sup
b- $p(Y \mid X)\geq$min.conf,and
c- $|p(X\rightarrow Y)-p(X)\cdot p(Y)|\geq$min.lift
then X→Y can be considered as an interesting rule. That is, if an itemset satisfies the thresholds of min.sup and min.conf, then it is called frequent item set, the others are called infrequent itemsets. In order to find patterns, the frequent itemset is scanned many times. It is proved that if an itemset is frequent then so must be all its subsets. In first scan, the support values of subsets are counted and so one can decide which sets are frequent. In later scans, only frequent itemsets are processed and candidate itemsets are determined. At the end of the pass, scanning process continues counting the support values for candidate sets by sorting and aggregating this sequential structure. Therefore the most frequent itemsets are selected from the candidate sets. Many algorithms have been developed for these processes such as AIS, SETM, APRIORI and APRIORI TID. In AIS and SETM algorithm, whole database is scanned in every pass. However Apriori type algorithms are different from AIS and SETM. Because there is no need to scan the whole database to count the candidate itemsets. The Apriori algorithms optimize the time needed to count the support value for candidate itemsets and enhance the performance. Thus they are the most famous and widely used in association rule mining. The Apriori algorithm makes multiple passes over a given database. It determines the frequent itemsets and selects the candidate itemsets from the frequent itemsets. The candidate itemsets satisfying the thresholds are our new frequent itemsets. The Apriori algorithm works simply as the following:
Let T be an itemset and t represents all nonempty subsets of T.
Generate all subsets of T
Construct a rule R: t→(T-t), where (T-t) indicates the set T without t
Generate R if R satisfies the min.conf
Do this for every subset of T

Procedure AprioriAlg( )
begin
L_1 = {frequent 1-itemsets};
for (k=2; $L_{k-1} \neq \emptyset$; k++) //k, itemsets count **do** {
C_k = apriori-gen(L_{k-1}); // New candidates
for all transactions $t \in D$ **do** {
for all candidates $c \in C_k$ contained in t **do**
c.count++;
}
$L_k = \{c \in C_k \mid \text{c.count} \geq \text{min.sup}\}$
}
Answer =
$\bigcup_k L_k$
end

Figure 1. The Apriori algorithm

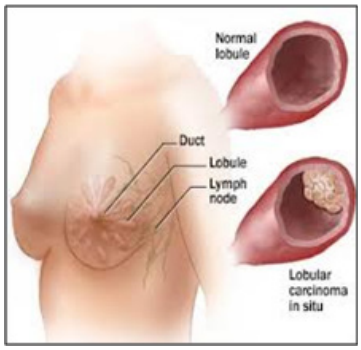

Figure 2. Illustration of lobular breast cancer (29)

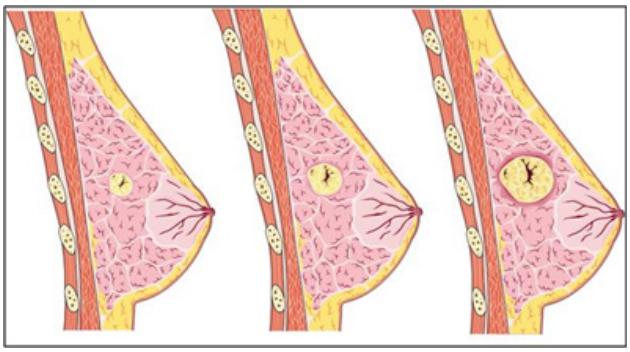

Figure 3. Illustration of tumor size of breast cancer (30)

The Apriori algorithm generating all frequent itemsets in a database D is given in Figure 1 (11-14)

AR has been developed as "market basket analysis" and then enhanced to banking and finance to determine the risk of lending. But in recent years, AR is applied to hospital management systems, healthcare and medical imaging to diagnose or for preventive medicine. Since healthcare applications and treatments need high costs, and the consumption of pharmaceuticals is sharply increased, reducing the costs is the main goal of health sector. Therefore the large data obtained in this area must be carefully analyzed (15-19)

Cancer and Breast Cancer

Cancer is the general name for approximately 100 diseases and emerges in different parts of the body. There are millions of living cells in a body. These cells grow and die normally in an orderly fashion in humans, but if they grow out of control and do not die then cancer starts in related organ or tissue. Unregulated growth of cancer cells forms a malign or benign tumor. The cancer may spread to distant parts of the body by means of bloodstream or lymph vessels and invade the tissues. This spreading process is called metastasis (20).

Breast cancer is one of the important types of cancer originating from breast tissue. A malign tumor develops usually in the milk ducts or in the lobules (Figure 2). By the time, cancer cells may invade the lymph nodes in underarm or spread into the bones. Tumor stages are classified according to tumor size, metastasis grade and spreading to

Table 3. List of variables used in analysis

Order	Variable	Values	Order	Variable	Values
1	Name	--	31	Operation Date	--
2	File ID	--	32	Right-Left	Right, Left, Bilateral
3	Height	Numerical	33	Multifocality	Yes, No
4	Weight	Numerical	34	LVI	Yes, No
5	BMI		35	Grade	1, 2, 3
6	Blood Group	+/- A, B, AB, 0	36	ER	Positive, Negative
7	Province	--	37	PR	Positive, Negative
8	Telephone	--	38	CERB2	Positive, Negative
9	Sex	M, F	39	Triple	Positive, Negative
10	Birth Date	--	40	Fish	Positive, Negative
11	Age in Diag.	Numerical	41	TM Size Stage	T1, T2, T3, T4, None
12	Menopause	Pre, Post, Man	42	Tumor Size	Numerical
13	Oral C.	Yes, No, Man	43	PLN	Numerical
14	HRT	Yes, No, Man	44	NLN	Numerical
15	Co-Morbid	Yes, No, KOAH, CAD, ...Etc	45	Metastasis	Yes, No
16	DM	Yes, No	46	Stage of Mets	M0, M1
17	Tyroid	Yes, No	47	Date of RT	--
19	Hyper/Tyroid	Yes, No	48	ADJ-NEOADJ	ADJ., NEOADJ.
20	Alcohol	Yes, No	49	Herceptin	Yes, No
21	Smoking	Yes, No	50	Zolodex	Yes, No
22	Secondary CA.	Yes, No	51	KT	Yes, No
23	Family	Mother, Father, Brother, Sister	52	RT	Yes, No
24	Family CA	Yes, No	53	Micro Calc.	Yes, No
25	FMD	Numerical	54	Bone Mets	Yes, No
			55	Pathologic Calc.	Yes, No

Table 4. Tumor Size Stage frequencies

Stage	Frequency	Percent
T1	225	16,4
T2	785	57,3
T3	238	17,4
T4	119	8,7
None	4	,3
Total	1371	100,0

the lymph nodes (Figure 3). There are many reasons to be a risk factor for breast cancer. However, like in all cancer, nobody knows the exact reason affecting the genes responsible for regulating the growth of cells. Breast cancer usually occurs in women, but uncommonly in men. The prevalence of this disease occurred in women is 135 times greater than men. Untreated breast cancer causes a severe illness or die. Although there are some invasive or non-invasive treatment methods, the crucial point is to stay healthy before disease. Eating balanced diet, no smoking and alcohol, exercising and sleeping regularly are some examples of preventive way of life (21).

Data Specifications

This is a retrospective cohort study whose data is related with the breast cancer patients of 1371 cases. The records are collected from the patients during ten years from 2002 to 2012. The dataset is provided by Oncology Hospital of Meram Faculty of Medicine in Konya, Turkey. There are 76 variables for each patient. According to CRISP-DM process, the dataset has been cleaned from noisy data. The misspelled values have been detected and corrected. The missing values for some variables have been replaced with their mean values. However the missing values of some categorical variables such as blood group, diabetes mellitus or co-morbidity could not be filled with any values. Some string type variables which are considered as of no contribute to modeling have been omitted. Thus the itemset has 42 variables. Although this size of 1371×42 for a dataset is considered very small in DM sense, it can be taken large enough as a medical data. It is very difficult to collect data from patients or hospital information systems due to ethics rules or irregular records. In dealing with medical data, we face a lot of inconsistent information in files or folders. The list of the descriptions of the variables is given in Table 3.

RESULTS

Firstly, we calculated the descriptive statistics and frequencies of the variables. Besides, the chi-squared tests specify that there is a significant relation between tumor size stage and bone metastasis, pathologic calcification, lymph vascular invasion (LVI) and Multifocality ($p<0.05$). We only give here, as a convenience, the frequencies of tumor size stages in Table 4 and we omit the rest of the results since we have many variables.

According to above attributes, we build the association rule steps by SPSS CLEMENTINE 12.0 (Powered by IBM).

Step.1 Replacing the excel file to the canvas
Step.2 Reading the variables and determining the types
Step.3 Detecting the missing and anomalous values
Step.4 Completing and correcting
Step.5 Connecting the values with related analysis nodes
Step.6 Obtaining the output files

We determined the attribute “tumor size stage” as consequent and other attributes as antecedents considering the min.sup=10% and min. conf=50%. After compilation, we obtained the results of approximately 125,000 rules. The five most interesting rules according to their lift values are shown in Table 5. According to the results, the tumor size stage T2 is important and seen frequently in most of the rules. Rule.1 means that stage T2 is associated with {(positive LVI (lymph vascular invasion)) ∩ (high progesterone receptor) ∩ (high estrogen receptor) ∩ (negative bone metastasis) ∩ (no information about breast feeding)} with 10.56% of support, 68.27% of confidence and 1.933 of lift value. Other rules can be interpreted in the same manner. Since the frequency of stage T2 in the data is greater than the others, we couldn’t get any interesting rule about

Table 5. The results of Apriori algorithm of breast cancer data

Consequent	Rule	Support	Confidence	Lift
TM size stage = T2	LVI = positive and PR percent = positive and ER percent =positive and bone metastasis = negative and feeding = Nil	10.5685	68.2759	1.1933
TM size stage = T2	micro-calcification = negative and comorbid disease = negative and bone metastasis = negative and hypertiroid/hypotiroid = negative	10.7872	68.2432	1.1927
TM size stage = T2	micro-calcification = negative and comorbid disease = negative and HT = negative and bone metastasis = negative and hypertiroid/hypotiroid = negative	10.7872	68.2432	1.1927
TM size stage = T2	micro-calcification = negative and comorbid disease = negative and bone metastasis = negative and DM = negative and Hypertiroid/hypotiroid = negative	10.7872	68.2432	1.1927
TM size stage = T2	LVI = positive and ER percent = positive and bone metastasis = negative and DM = negative and HRT = Nil	11.0058	68.2119	1.1922

Table 6. The GRI results of breast cancer data

Consequent	Rule	Support	Confidence	Lift
TM size stage = T4	Menopause = POST and Oral Contraceptive = No and Multifocality = No and metastasis = Yes and micro-calcification = Yes	2.480	50.000	5.765
TM size stage = T1	Menopause = POST and comorbid disease = Yes and Family cancer story = No and Multifocality = No and LVI = Positive and metastasis = Yes and micro-calcification = Yes	0.360	100.000	6.098
TM size stage = T3	Family cancer story = No and Multifocality = Yes and LVI = Positive and micro-calcification = Nil	0.360	100.000	5.765

other stages. But there is a different method for rule induction. It is called the generalized rule induction (GRI) or J-measure proposed by Smyth-Goodman (1992). By means of this procedure, we got the following rules with higher confidence and interest measure. As seen in the Table.6, the confidence values of T1 and T3 are maximum and the lift values are very high whereas the support values are very small. This case shows that the rules for T1 and T3 are rare, but the rule for T4 is strong with 2.4% of support and 50% of confidence, which means that stage T4 is frequently diagnosed after menopausal period with metastasis and calcification, without using any oral contraceptive pills. In stage T1, the patient has generally a comorbid disease, lymph vascular invasion, metastasis and calcification; in post menopausal period and no cancer story in her family (22-28).

CONCLUSION

In this study, we have applied the association rule mining to breast cancer data. We have used the Apriori algorithm which is very useful for analyzing large datasets. Classical statistical methods are weak to be applicable for such datasets and tests of hypothesis should be repeated many times to determine the relations between the variables. On the other hand, we may save time and find many relations by the Apriori algorithm. Therefore, the extraction of hidden rules is an important procedure, especially for medical data which is increasing dramatically in all healthcare areas. Health management and drug costs are threatening the nations and they may cause a gap in their national incomes. Especially, expenditures of cancer researches create a burden on the budget. Therefore we conclude that huge medical data should be analyzed carefully to make successful decisions. The quality of data is also important in DM process since we always have a large dataset. The missing values or anomalous data should be cleaned carefully to reach a valuable result.

As a future work, we are planning to apply the different types of Apriori algorithm, such as negative association or rare association rule mining which are more suitable for medical data, to the myocardial infarction diseases.

Acknowledgment

We would like to thank to Prof. Melih Cem BÖRÜBAN, the head of Medical Oncology Department, for breast cancer data (by the decision of Noninvasive Clinical Researches Committee with Date: 04/24/2014 and No:2013/05).

REFERENCES

1. Çınar H, Arslan G. Veri Madenciliği ve CRISP-DM Yaklaşımı, 17. İstatistik Araştırma Sempozyumu Bildiriler Kitabı 2008; 304-14.
2. Larose DT. Discovering Knowledge in Data- An Introduction to Data Mining. USA: John Wiley & Sons Inc., 2005: 30-6.
3. Ayad AM. A New Algorithm for Incremental Mining of Constrained Association Rules, MS Thesis, Alexandria University (unpublished) 2000; 25-37.
4. Bayardo R, Agrawal R. Mining the Most Interesting Rules, Proceedings of SIGMOD Int'l Conference on Knowledge Discovery and Data Mining 1999; 145-54.
5. Benoit G. Data Mining, Annual Review of Information and Technology 2002; 36: 265-310.
6. Agrawal R, Imielinski T and Swami A, Mining Association Rules between Sets of Items in Large Databases. SIGMOD Report of Association for Computing Machinery 1993; 207-16.
7. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules, Proceedings of the 20th VLDB Conference 1994; 1-13.
8. Bruzzese D, Davino C. Visual Mining of Association Rules, Visual Data Mining 2008; 4404: 103-22.
9. Berardi M, Appice A, Loglisci C, Leo P. Supporting Visual Exploration of Discovered Association Rules through Multi-Dimensional Scaling, Lecture Notes in Computer Sciences 2006; 4203: 369-78.
10. Ramaswamy S, Mahajan S, Silberschatz A. On the Discovery of Interesting Patterns in Association Rules, Proceedings of the 24th Very Large Data Bases Conference USA, Morgan Kaufmann Publishers Inc., 1998: 368-79.
11. Savasere A, Omiecinski E, Navathe S. An Efficient Algorithm for Mining Association Rules in Large Databases, Proceedings of the 21st Int'l Conference on Very Large Databases 1995: 432-44.
12. Srikant R, Agrawal R, Mining Generalized Association Rules, Proceedings of the 21st VLDB Conference 1995:1-13.
13. Srinivas K, Rao GR, Govardhan A, Mining Association Rules from Large Datasets towards Disease Prediction, Proceedings of Int'l Conf. On Information and Computer Networks 2012; 27: 22-6.
14. Tsay Y, Chiang J, CBAR: an efficient method for mining association rules, Knowledge-Based Systems 2004; 18: 99-105.
15. Kotsiantis S, Kanellopoulos D. Association Rules Mining: A Recent Overwiev, GESTS Int'l Transactions on Computer Science and Engineering 2006; 32: 71-82.
16. Hahsler M, Grün B, Hornik K. A Computational Environment for Mining Association Rules and Frequent Item Sets, Journal of Statistical Software 2005; 14: 1-25.
17. Doğan Ş, Türkoğlu İ. Diagnosing Hyperlipidemia using Association Rules, Mathematical and Computational Applications 2008; 13: 193-202.
18. Güllüoğlu SS. Tıp ve Sağlık Hizmetlerinde Veri Madenciliği Çalışmaları:

Kanser Teşhisine Yönelik Bir Ön Çalışma, Online Academic Journal of Information Technology, Online Academic Journal of Information Technology 2011; 2(5): 1-7.
19. Hu R. Medical Data Mining based on Asociation Rules, Computer and Information Science 2010; 3: 104-8.
20. Imberman SP, Domanski B, Thompson HW. Using Dependency/Association Rules to Find Indications for Computed Tomographing a Head Trauma Dataset, Artificial Intelligence in Medicine 2002; 26: 55-68.
21. Jabbar MA, Chandra P, Deekshatulu BL. Cluster Based Association Rule Mining for Heart Attack Prediction, Journal of Theoritical and Applied Information Technology 2011; 32: 196-201.
22. Kwasnicka H, Switalski K. Discovery of Association Rules from Medical Data- Classical Evolutionary Approaches, Proceedings of 21st Autumn Meeting of Polish Information Processing Society 2005;163-77.
23. Nahar J, Tickle K, Shawkat A, Chen YP. Significant cancer Prevention Factor Extarction: An Association Rule Discovery Approach, J Med Syst 2009; 35: 353-67.
24. Obenshain MK. Application of Data Mining techniques to Healthcare Data, Infection Control and Hospital Epidemiology 2004; 25(8): 690-5.
25. Ordonez C, Santana C, Braal L. Discovering Interesting Association Rules in Medical Data, Proceedings ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery 2000: 1-8.
26. Yıldırım P, Uludağ M, Görür A, Hastane Bilgi Sistemlerinde Veri Madenciliği, Akademik Bilişim- Onsekiz Mart Üniversitesi Çanakkale 2008; 11(21): 429-34.
27. http://en.wikipedia.org/wiki/Cancer (visit date: 12/06/2014)
28. http://en.wikipedia.org/wiki/Breast_cancer (visit date: 12/06/2014)
29. http://www.healthtap.com (visit date: 12/06/2014)
30. http://www.ladycarehealth.com (visit date: 12/06/2014)