Mugla Journal of Science and Technology

DETERMINATION OF HEART ATTACK RISK ON PATIENTS DATA by DATA MINING APPLICATIONS

İlhan TARIMER[1], Fatih ELMAS[2]*

[1]Department of Information Systems Engineering, Technology Faculty, Muğla Sıtkı Koçman University, 48000, Muğla, Turkey
itarimer@mu.edu.tr
[2] Degree Program of Electronics and Computer Education, Graduate School of Natural and Applied Sciences, Muğla Sıtkı Koçman University, 48000, Muğla, Turkey
fatihcs@gmail.com

Accepted: 02.04.2015
*Corresponding author

Abstract

In this study, it has been investigated that feasibility of data mining which is used to extract meaningful knowledge in order to effect to decision making processes in health field. As an example to a case study, it has been tried to obtain that determining the factors which trigger heart attacks by defining common changes in blood values of patients whom have got heart attacks previously. Success of the analysis done has been measured by testing the obtained results on a group of patients. In the study, Apriori and GRI algorithms stemming from association rule algorithms have been used; success of rule sets created by these algorithms has been investigated by making several comparisons. As the result, several patterns meant to pre-signals determining heart attacks from data of the patient group which have the blood values have been put forth.
Keywords: Data mining, Apriori and GRI algorithms, heart attack.

HASTA VERİLERİ ÜZERİNDE VERİ MADENCİLİĞİ UYGULAMASI İLE KALP KRİZİ RİSKİNİN TESPİTİ

Özet

Bu çalışmada, veri tabanları içerisinde, karar verme süreçlerine etki edebilecek anlamlı bilgileri çıkarmak için kullanılan veri madenciliğinin sağlık alanında uygulanabilirliği araştırılmıştır. Alan çalışmasına örnek olarak, kalp krizi geçiren hastaların kan değerlerinde meydana gelen ortak değişimler tespit edilerek kalp krizini tetikleyen faktörlerin tespitine çalışılmış; elde edilen sonuçlar hasta grubu üzerinde test edilerek yapılan analizin başarımı ölçülmüştür. Bu çalışmada birliktelik kuralı algoritmalarından Apriori ve Gri algoritmaları kullanılmış; bunların oluşturdukları kural kümelerinin başarımı, karşılaştırmalar yapılarak incelenmiştir. Gri algoritmasının Apriori'ye göre daha az kural ürettiği halde aynı başarımı gösterdiği tespit edilmiştir. Sonuç olarak kan değerleri verilen bir hasta grubu verilerinden kalp krizi riskinin tespiti için ön sinyaller anlamına gelen çeşitli desenler ortaya konmuştur.
Anahtar Kelimeler: Veri madenciliği, Apriori ve Gri algoritması, Kalp krizi.

1 Introduction

Nowadays data storage in proportion to the increase of the amount of knowledge has been eased. All state institutions and private establishments have started to keep activities' data and transaction records electronically. In order not to struggle with any problem, data base management systems provide suitable solutions, while searching and modifying records. To obtain meaningful knowledge from data bases which increases rapidly is valuable, but for managers of data base management, it's a time consuming job. Therefore, knowledge conquering process from raw information (data mining) has been developed. Data mining is a process of to analyze data from different aspects and to obtain a beneficial knowledge [1, 2]. It is also said that data mining is all kinds of studies to obtain meaningful useful information, patterns, changes, irregularities and relations from raw data. Data mining is to occur and to examine meaningful useful rules and connections which predict future by means of software programs from huge data stores [3, 4]. Today, data mining applications are being seen in fields of engineering, health and medicine, business, shopping, banking and education [5].

Incidence of chronic diseases and its imposed financial burdens increase more and more. Therefore, preventive solutions which prevent chronic diseases are being developed [6]. Classical query methods which will be able to work on information systems do not accomplish doing all data analyze, since that hospital information systems consist of demographic information, sickness and treatments, tests conducted, billing and other chancellery regarding to patients. Using data mining as a decision support tool in management of health institutions and creation of health policies can help to health professionals when they take decisions truly [7]. Global risks within cardiovascular diseases have been dealt with a study [8]. In order to help to medical doctor, a decision support system has been designed for diagnosis of myocardial infarction among biochemical blood test results in another study. It has seen that the results developed by the system has overlapped with the decisions given by doctor [9].

In the study, risk factors which triggers heart attack of patients who had undergone heart crisis have been tried to determine by studying hematology and biochemistry values together with data mining association rules. In order to determine probability risk of heart attack, several patterns from blood values of a patient group have been come forth. A system that works in digitized patient logs and also offers a method according to the physician's request has been designed. For this purpose, a decision support system which was hide behind the user

graphics interface has been developed. It is expected that the system could have reached many data instantaneously, established a relation amongst the data, analyzed all the data, had early warning capabilities, and used easily. In this study, patient data have been evaluated by Apriori and GRI algorithms, and their success have also been compared.

2 Data Mining Methods

In order to use huge sized data, to explore valuable confidential information within data, and to extract meaningful and useful patterns, methods of "clustering, classification and regression with association rules" have been preferred. In this study, Apriori and GRI algorithm technics which are the methods of extracting association rules have been used.

2.1 Apriori and GRI Algorithms

Apriori algorithm is the most used one within the association rules algorithm [10]. However, there is a problem for extracting meaningful rules from the cluster which consists of many rules. In order to make it easy, GRI algorithm was developed to find meaningful rules within the set of rules.

To find frequent item sets, database is scanned many times by means of Apriori algorithm. The frequent item sets are found in the first scanning. In the following scanning, the frequent item sets found in the first scanning are used to produce new potential frequent item sets which are called as candidate clusters.

The support values of candidate clusters are calculated during scanning, and thereby, the clusters that get minimum support from candidate clusters would be determined as frequent item sets. These frequent item sets would be a candidate cluster. This process continues until no new frequent item sets would be found [11].

In order to calculate dissimilarity of rules, GRI algorithm uses a quantitative scale, and limits possible values by such measure (Smyth and Goodman, 1992). To reach information, GRI uses a quantitative approach (J) for measuring dissimilarity within candidate association rules. GRI algorithm calculates this scale (J) by measuring each rule as possible as different from border values. So, it narrows to search field of meaningful rules.

To find rules which consist of the most significant information, it creates (J) index. It summarizes candidate association rules in term of the index created, and lists the rules by taking into account values of trust and support. As result of this, it would be selected less numbered and more meant rules by narrowing rule search field. One of the advantages of GRI algorithm is to eliminate nonsignificant rules within the meaningful ones [12].

3 Data Mining In Risk Analysis of Heart Attack

Raw data in health area changes quickly in terms of content and structure base. In this context, works on accessing to effective information at decision process have been accelerated with a strategy applied by Turkey Ministry of Health [13]. The cost brought by chronicle diseases is being gradually increased since that chronicle diseases are being seen frequently. Therefore, to develop predictions based solutions which will prevent chronicle disease, methods of data mining would be suitable. The factors that cause to chronicle diseases and their border values could be determined by using classification, regression methods or association rules.

3.1 Examination and Regulation of Data

Heart attack is a physiological situation that occurs in coronary arteries of heart after a disorder. This is an insufficiency that occurs with severe chest pain, and has a probability of death. Heart attack is defined that it is emerged as depending upon sudden decrease or cessation of blood flow as well. In such cases, there would be seemed oxygen failure within the veins which supply the heart. Cardiovascular system failures resulting heart attack(s) would be in different types. Coronary artery disease is the most common one in these types and it causes mostly to deaths.

The main symptom of coronary artery disease is to feel pain at chest, and it is the main reason of heart spasm and heart attack. Heart attack is a serious hazard for human life, and is also one of the main death reasons. In a study executed in the country, it has been determined that risks of chronicle diseases have seen as %78.8 within the older people [14]. Therefore, to take preventive measures against heart disease and heart attack is very important.

In this study, the laboratory data which are obtained from the patients whom are dispatched to angioma unit of a private hospital in Muğla with the diagnosis of heart attack at diagnostic phase have been used. The data have been kept in database of Microsoft Excel, and regulated by using Microsoft Visual Basic. 153 Cases of heart attack totally have been stored in database. As an application of data mining, the 103 cases of the cases were used as 'learning group', and to extract 'rule'; the 50 cases were used to test the 'rules' obtained. These data consists of complete blood count, biochemistry, CK-MB and sedimentation test results with the age and sex data. The information in data set consists of 32 category (Table 1).

Table 1. Information of 32 category data set

Hemo grams	Biochemistry	CK-MB	Sedimentation
WBC	Fasting blood sugar		
LYMP %	Cholesterol		
MONO %	HDL Cholesterol		
NEUT %	LDL Cholesterol		
GRAN %	Triglycerides		
RBC	Urea		
HGB	Creatinine		
HCT %	Sodium		
MCV	Uric Acid		
MCH	Potassium		
MCHC	SGPT		
PLT			
MPV			
PCT			
PDW			

In Table 1, complete blood count (hemograms), counts of the chemicals in blood (biochemistry), isoenzyme ratio which points tear on the heart muscles at heart attacked patients (CK-MB) and crumpling rate in blood (sedimentation) have been given.

The data has been gathered from angioma unit of a private hospital in Muğla. The data categories containing incomplete information (more than 40%) have been examined, and they have been deleted from the records since that they don't have a meaningful value.

In case of the data categories containing incomplete information (less than 40%), the average value of the category have been used instead of the incomplete information. A part of the data and graphics of these data which belongs to the patients is given in Table 2, and Fig. 1 respectively.

Table 2. A part of the laboratory data of patients'

Sedimentation	WBC	LYMP %	MONO %	NEUT %	NEUT#	LYMP #	MONO#	RBC	HGB
24,000	10,400	9,000	6,200	80,244	10,998	0,900	0,600	4,800	12,700
19,000	20,200	10,600	2,200	86,900	17,500	2,100	0,400	4,720	15,600
2,000	12,800	5,800	1,600	92,300	11,800	0,700	0,200	4,960	15,200
3,000	9,300	28,900	8,900	80,244	10,998	2,700	0,800	4,660	13,300
5,000	7,800	17,400	4,900	74,300	5,800	1,400	0,400	5,360	15,500
78,000	15,400	18,900	12,000	67,000	10,300	2,900	1,900	4,430	13,800
3,000	9,600	13,100	5,400	79,700	7,700	1,300	0,500	4,780	14,900
5,000	12,100	17,500	5,300	76,100	9,200	2,100	0,600	4,670	15,400
48,000	9,600	13,600	2,400	83,300	8,000	1,300	0,200	4,450	13,600
6,000	15,600	11,100	4,000	84,500	13,200	1,700	0,600	5,910	16,700

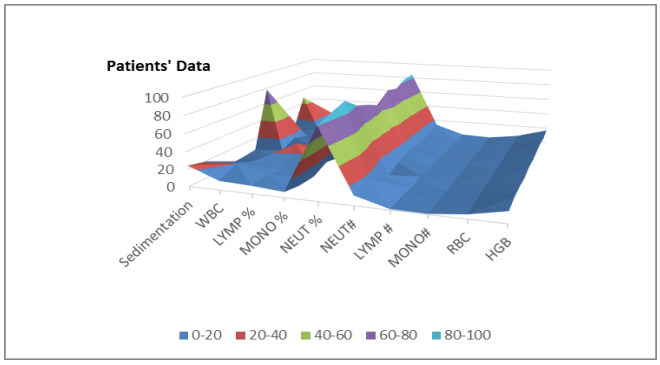


Figure 1. Graphics of the patients' data versus biochemistry and hemograms values.

3.2 Transforming Data

In order to complete the missing data, to wipe out the faulty ones, and to compare the remaining data with healthy person hemogram and biochemistry values, a transformation has been done. For this the data, if they are beneath the certain reference values, have been transformed to letter 'A'; if they are at normal level, have been transformed to letter 'N' and if they are higher than the reference values, have been transformed to letter 'Y'. Once they are transformed to letters 'A, N, and Y', the new values of the data are given in Table 3.

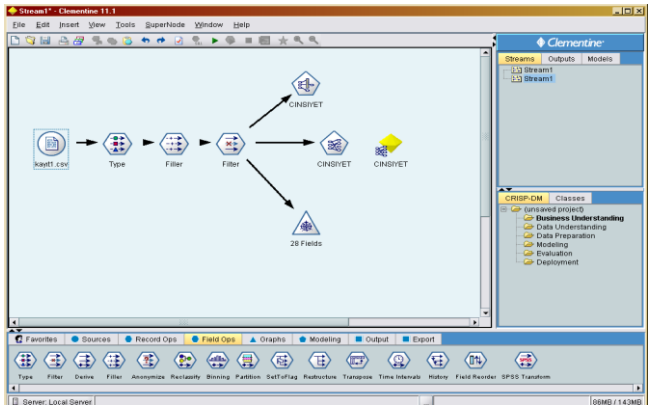
Table 3. A part of the transformed patients' laboratory data according to reference values

Sedimentation	WBC	LYMP %	MONO %	NEUT %	NEUT#	LYMP #	RBC	HGB	HCT	MCV
Y	N	A	N	Y	Y	N	N	N	N	N
N	Y	A	N	Y	Y	N	N	N	N	N
N	Y	A	A	Y	Y	N	N	N	N	N
N	N	N	N	Y	Y	N	N	N	N	N
N	N	A	N	N	N	N	N	N	N	N
Y	Y	A	Y	N	Y	N	N	N	N	N
N	N	A	N	N	N	N	N	N	N	N
N	Y	A	N	N	Y	N	N	N	N	N
Y	N	A	N	Y	N	N	N	N	N	N
N	Y	A	N	Y	Y	N	N	N	N	N

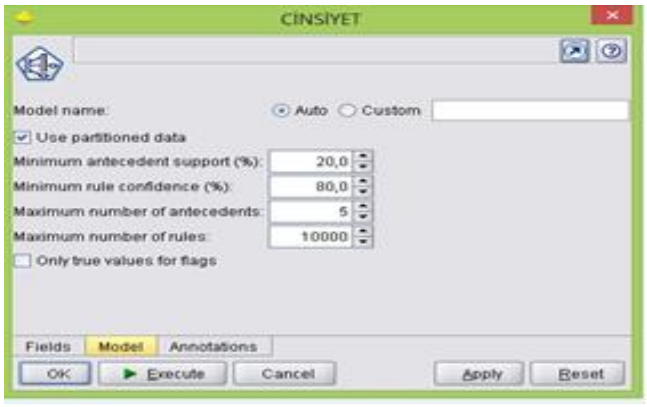
The transformed sex and age information of the patient data is seen as in Table 3. The E/K coding for the sex data has been used, interval of the patients' has been determined for the age. The age information has been grouped as the first between 30–50, the middle between 50–70, the advanced between 70–plus.

3.3 Processing Data by SPSS Clementine Program

A patients' database has been created, and then it has been imported to a visual modelling tool 'SPSS Clementine' software. Hence, it has been subjected to cleaning, consolidation and transformation processes. Afterwards, types of variables within the data set have been attended by object of 'Types'; data with specific values have been filtered by object of 'Filter'; which data would be analyzed and which data would be ignored by object of 'Filter'. GRI and Apriori objects have been used to apply GRI and Apriori algorithms to the data set. On the other hand, web object has been used to observe the structure of relations network amongst the data. In Fig. 2 (a), the application in Clementine user interface; in Fig. 2 (b), 'confidence and support' values of Apriori and GRI algorithms have been shown.



(a) Application in Clementine user interface



(b) Confidence and support values of Apriori and GRI.

Fig. 2 Views of Clementine user interface.

While GRI and Aprori algorithms were being applied, confidence and support values which are function parameters, have been determined as 80% and 20% after many trials. The 'confidence' value states validity of the rule; namely, when situation A is occurred, frequency of occurring situation B is 80%. The 'support' value shows the least number of patients' value which provides status 'A and B' within the all patients' records that the ratio is 20% in the study. It has been wanted that more than 5 premises should produce rules, and the results have been obtained.

In the study, the data taken at biochemistry and hemogram tests have been filtered separately, since that it has been wanted to determine common abnormalities which are occurred in blood values of the heart attacked patients'. Nevertheless, common abnormalities have been committed (processed) many times both Apriori and GRI algorithms by 'confidence and support' values; hence, it has been reached to the most healthy rule clusters. In the analysis, the association rate and the incidence of rules sought are high; therefore, the rules have been determined by testing many probabilities.

3.4 Determination Rule Abnormalities and Testing

In the study, SPSS Clementine has been determined many patterns both in Apriori algorithm and GRI algorithm, it has produced many association rule as shown in Fig. 3. The most important reason of this is to use many data belong to each patient in the study. When the obtained rules are examined, the rules that happen from togetherness of abnormalities seen in the patient's blood values and togetherness of many normal values, draw the attention. The meaningful rules regarding to the study within the rules examined are the ones that trigger heart attack or the changes that cause abnormalities to come up with. In this point of view, the patterns regarding to determination of meaningful rules within the many rules have been obtained. Then they have been filtered by the developed program codes, and the contradictory situations have solely been determined. In order to determine the risk of heart attack within the patients' data, 38 rules by GRI algorithm and 119 rules by Apriori algorithm have been extracted. The factors extracted from these are given in Table 3.

Table 4 Results of the data by Apriori and GRI algorithm

Variables	Status by Reference Value
HDL Cholesterol	Low
NEUT#	High
Sodium	Low
Cholesterol	High
Triglycerides	High
WBC	High
Fasting Blood Sugar	High
LYMP%	Low
CKMB	High
NEUT%	High

In this study, the analysis made by Apriori algorithm has compared to the analysis made by GRI algorithm. It has seen that the obtained results of Apriori algorithm from the analyze in the comparison have covered the results taken by GRI algorithm as shown in Table 5. This result shows that the success of these two algorithm are closed to each. Reliability and availability of the results at determination of sickness as given in Table 5, have been tested by a written program. By means of that, several rules have been obtained. All the rules have been questioned for every patient located in the patient database separately; validity of each rule in which is for patients, has been determined. The obtained data has been shown in Table 5 and 6. The valid Apriori rules and their validity ratios for the patients' data are given in Figure 3.

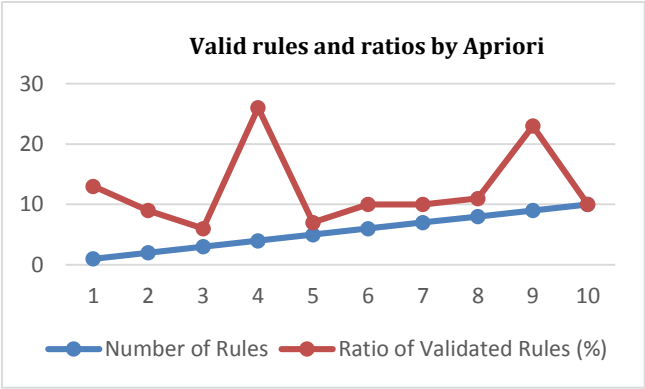


Fig. 3 Valid rules and ratios in some patients by Apriori.

Table 5 A part of test rules obtained by Apriori algorithm

Rule Number	Sex	Values		Values		Values			
1	M	Triglycerides	Y	CKMB	Y	LYMP%	A		
2	M	Cholesterol	Y	WBC	Y	NEUT#	Y		
3	M	Cholesterol	Y	CKMB	Y	LYMP%	A		
4	F	MCV	A	Fasting Blood Sugar	Y				
5	F	MCV	A	Fasting Blood Sugar	Y	Sodium	A		
6	M	Triglycerides	Y	WBC	Y	CKMB	Y	LYMP%	A
7	M	Triglycerides	Y	NEUT#	Y	CKMB	Y	LYMP%	A
8	M	Cholesterol	Y	WBC	Y				
9	M	Triglycerides	Y	CKMB	Y	LYMP%	A	Sodium	A
10	M	Triglycerides	Y	WBC	Y	CKMB	Y	Sodium	A

Table 6 Rules valid in some patients and validity view of Apriori rules

Number of Patients	Number of Validated Rules	Rule Number	Validity Ratio (%)
16	6	1	21
16	13	2	26
16	2	3	26
17	6	4	22
17	13	5	21
17	2	6	24
18	3	7	26
18	6	8	24
18	7	9	22
18	8	10	24
18	9		
18	10		
18	11		
18	12		
18	13		
18	9		

The data given in Table 6 and the graphics seen in Fig. 3 present the rules obtained in the study, which rule would be valid in each patient, and validity ratios of every rule. In terms of Apriori algorithm's results, rules have the highest success have been come up with as 26%. According to these rules, it is inferred that risk in repetition of heart attack would be 26%. Validity of the patterns obtained by using Apriori algorithm has been tested to determinate the risky patients as separate rules; the risk percentage of heart attacked patients has been determined as 70.59%. It is given that how many rules are valid in every patient regarding to Apriori algorithm in Table 7.

In the study, the testing process for comparing the patients' data has repeated in terms of GRI algorithm as well. Validity rules determined by GRI algorithm in the patients and ratio of valid rules at every patient have been given in Table 8. Table 9 gives the results obtained by GRI algorithm.

Table 7 Valid rules and their ratios by Apriori algorithm

Number of Rules	Ratios of Validated Rules (%)
1	12
2	7
3	3
4	22
5	2
6	4
7	3
8	3
9	14
10	0

Table 8 Valid rules and their ratios by GRI algorithm

Number of rules	Ratio of Patient (%)	Number of valid rules	Ratio of patient (%)
1	30	1	25
2	28	2	4
3	28	3	10
4	28	4	0
5	26	5	0
6	26	6	2
7	26	7	4
8	26	8	8
9	30	9	0
10	24	10	6
		11	0
		12	10
		13	0
		Total	70

From Table 8, rules have the highest success have been come up with as 30%. It is inferred that repetition risk of heart attack would be occurred as 30%. It is understood that these validity ratios are higher than Apriori.

Table 9 A part of test rules obtained by GRI algorithm

Rule Number	Sex	Values		Values		Values		Values	
1	M	Fasting Blood Sugar	Y	HDL	A	Sodium	A		
2	M	LYMP%	A	PDW	Y				
3	M	NEUT#	Y	Triglycerides	Y	CKMB	Y	Age	<5.5
4	M	LYMP%	A	Triglycerides	Y	Sodium	A	CKMB	Y
5	M	WBC	Y	Cholesterol	Y				
6	F	MCV	A	Fasting Blood Sugar	Y	Sodium	A		
7	F	MCV	A	Fasting Blood Sugar	Y				
8	M	LYMP%	A	Triglycerides	Y	CKMB	Y		
9	M	WBC	Y	LYMP%	A	Triglycerides	Y	CKMB	Y
10	M	WBC	Y	NEUT#	Y	Cholesterol	Y		

The ratio of the rules having highest success obtained by using GRI algorithm is 70.59%. When both two algorithms were tested on patient's cluster, it has seen that validity of the rules obtained by using GRI algorithm were higher than validity of the rules obtained by Apriori algorithm. It has been understood that validity values of the rules in GRI algorithm were to be closer to each other from Table 9.

4 Results and Suggestions

In this study, blood values of the patients who have had a heart attack have been taken firstly, then Apriori and GRI algorithms have been applied to these blood values for determining the changes occurred on the patients. More than five thousand rules have been created by Apriori association rules algorithm. On the other hand, number of the rules has been two thousand by GRI association rules algorithm. Because Apriori algorithm consists of four elements from 1 to 4; they are sub-cluster of each, and it consists of mostly repetitive but less rules. In Apriori algorithm application, single variable changes which are seen the most; then, the second variables with the single variable are being determined. On the subsequent steps, new rules together with the rules extracted in previous step are being produced. In GRI algorithm, the extracted rules are being generalized; less number but more meaningful rules are being produced.

In this study, when the valid and meaningful patterns obtained are tested on the patients cluster, it has been seen that probability of heart attack ratio would become 70.59% in terms of results taken by Apriori and GRI algorithms. Results of this study could make the doctor to determine heart attack risk of the patients recorded in database early. In case of adding patients' stories to the created patient information database, it could become possible to determine the sicknesses which trigger heart attack and others earlier. If the precursor signals obtained are added as online to an information processing system, an early warning could be provided regarding to illnesses which are likely to occur and could be benefited to doctor at disease diagnosis.

5 References

[1] Şentürk, Z.K., The Diagnosis of Cancer with Data Mining, *Master Thesis*, Düzce University, Düzce, p.58, 2011.

[2] Çakır, F., *Akgöbek, Ö, Designing An Expert System in Data Mining*, *Academic Information 2009 Conference*, Harran University, Şanlıurfa, Proceedings Book, p.801–806, 2009.

[3] Savaş, S. Topaloğlu, N. ve Yılmaz, M., Data Mining and Practices in Turkey, İstanbul Commerce University, *Journal of Natural and Applied Sciences*, Year:11, No: 21, p. 1-23, 2012

[4] Han, J. ve Kamber, M., Data Mining Concepts and Techniques, *Morgan Kauffmann Publishers Inc*., 1-35., 2006.

[5] Azimli, M., Data Mining Applications in Medicine, *Master Thesis*, Gazi University, Ankara,. p.63, 2011.

[6] Obenshain, K., *Application of data mining techniques to healthcare data*, Data Infect Control Hosp Epidemiol, 25: 690-695, 2004.

[7] Koyuncugil, A.S. ve Özgülbaş, N. (2009) Data Mining: Use and Applications in Medicine and Health Services, Journal of Information Technologies, Vol.2, No:2, p.21-32, 2009.

[8] Güleç, S., Global Risk of Cardiovascular Disease and Objectives, *Turkish Society of Cardiology*, Volume:37, No:2, p.1–10, 2009.

[9] Doğan, Ş., Türkoğlu, İ., Yavuzkır, M., Heart Attack Detection From Cardiac Enzymes By Using Decision Trees, *e-Journal of New World Sciences Academy*, http://www.newwsa.com/download/sayi_icerikleri/TO63OBO.pdf , ISSN:1306-3111, Vol.: 2, Nu.: 3, p. 39–50, 2007.

[10] Alataş, B., and Arslan, A., Mining of Fuzzy Association Rules with Genetic Algorithms, Journal of Polytechnic, Vol: 7, No: 4 pp. 269-276, 2004.

[11] Agrawal, R., Imielinski, T. ve Swami, A., Data Mining Association Rules Between Sets of Items in Large Databases, *Proceedings of The Acm Sigmod International Conference On Management of Data*, Washington USA,. 1993.

[12] Aggelis, V. ve Christodoulakis, D., E-Trans Association Rules for e-Banking Transactions, *IV. International Conference on Decision Support for Telecommunications and Information*, 2004, Warsaw Poland.

[13] Akdağ, R., Aydın, S., Buzgan, T., Demirel, H. ve Gündüz, F., *Cycling programs at health and basic health services in Turkey*, T.C. Health Ministry Publications, Ankara, p.175, 2008.

[14] Özdemir, L., Koçoğlu, G., Sümer, H., Nur, N., Polat, H., Aker, A., Bakıcı, Z., Frequency of Some Chronic Diseases and Risk Factors Among the Elderly People in Sivas*, C. U. Journal of Medicine Faculty*, 27 (3), p.89 – 94, 2005.