SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ DERGİSİ

Sakarya University Journal of Science
SAUJS

e-ISSN 2147-835X | Period Bimonthly | Founded: 1997 | Publisher Sakarya University |
http://www.saujs.sakarya.edu.tr/en/

Title: Knowledge Discovery Using Clustering Methods in Medical Database: A Case Study for Reflux Disease

Authors: Yunus DOĞAN, Fatma RIDAOUI

Recieved: 2020-12-07 00:00:00

Accepted: 2021-03-13 12:52:00.995000

Article Type: Research Article

Volume: 25
Issue: 2
Month: April
Year: 2021
Pages: 439-452

How to cite
Yunus DOĞAN, Fatma RIDAOUI; (2021), Knowledge Discovery Using Clustering Methods in Medical Database: A Case Study for Reflux Disease. Sakarya University Journal of Science, 25(2), 439-452, DOI: https://doi.org/10.16984/saufenbilder.837209
Access link
http://www.saujs.sakarya.edu.tr/en/pub/issue/60672/837209

Sakarya University Journal of Science 25(2), 439-452, 2021

Knowledge Discovery Using Clustering Methods in Medicine: A Case Study for Reflux Disease

Yunus DOĞAN[*1], Fatma RIDAOUI[2]

Abstract

Digitalization spreads day by day around the world; thus, the amount of data collected is on the rise. An increasing amount of data leads us to use the data and get the advantage of it by using methods like Data mining. Data mining is used in several industries. Especially as medical data is essential to be understood, it is crucial to work on it. Reflux disease is a painful illness spreading around the world. Reflux is more common compared to formerly known numbers of patients. Even though reflux is not as fatal as cancer, it decreases the quality of life and makes many people suffer in their daily life. So, reflux is affecting mental health directly. If we can ease the process of diagnosis of reflux, we may provide a better quality of life for people. In this study, various data mining algorithms are applied, and it is seen from results that medical care can be improved by changing. Nowadays, artificial intelligence applications in the field of gastroenterology stand out in various sources in the literature. However, a large database required that is specific for Reflux disease to implement these applications is available only at the Reflux Research Center in Ege University in Turkey. By benefiting the Short Form36 and Quadrad12 questionnaire data in this database, 3,909 patients and many artificial intelligence algorithms were used to discover the hidden associations among responses in the quality of life of these patients. The algorithms used in the tests are Apriori, Frequent Pattern Growth, Density-Based Spatial Clustering of Applications with Noise, Self-Organizing Map, and KMeans. In the tests, it was observed that the most successful algorithm in terms of the structure of the data was KMeans, and a set of remarkable 27 rules according to the optimal Sum of Square Error value was obtained.

Keywords: Clustering, data mining, medical information systems, reflux disease

1. INTRODUCTION

Databases store important and viable data but they also store irrelevant and not useful data too. Knowledge discovery can be used for every size of data; however, it is meaningful and unavoidable to use for large data repositories. Understanding and interpreting the data gained more importance than just storing the data recently.

[*] Corresponding Author: yunus@cs.deu.edu.tr
[1] Dokuz Eylül University, ORCID: https://orcid.org/0000-0002-0353-5014
[2] Gebze Technical University, E-Mail: f.balci@gtu.edu.tr. ORCID: https://orcid.org/0000-0003-1653-1466

The study aims to export the hidden knowledge in the database that is collected in Ege University Reflux Center in Turkey. In other words, this study focuses to find rules about reflux disease, personality, and the quality of life. Especially in Turkey, such studies over huge medical data have not been performed before. The Reflux Center has collected and stored patients' data for years, and the number of patients is estimated at around 7,000. However, those data are not processed for the help of obtaining hidden outcomes such as what is the probability of a person being a reflux patient, etc.

Medical data is vital and sensitive, that is why it should be protected and used carefully for the sake of patient and healthcare. The Reflux Center has a huge amount of patient information that is not processed or interpreted. There is around 7,000 patient's information stored. This study aims to examine the relationship between data and clustering with the aid of analyzability, usability, and diagnostics.

Today, computer-based solutions have become compulsory and are seen in medicine, too. Current medical studies are on Data Mining [1], Artificial Intelligence (AI) [2], Machine Learning, and Deep Learning ranging from processing medical images on radiological data to early diagnosis issues [3, 4]. Examples of intelligent solutions for malignant diseases such as heart disease, cancer, and diabetes are also encountered [5, 6]. In these studies, more than 85,000 patients were analyzed to find confidential information; therefore, these studies could only be applied via data mining techniques. If it comes to the field of gastroenterology, it can be said that artificial intelligence studies have been handled to raise awareness in this regard all over the world [7-10]. In this study, the database in the Reflux Research Center has been used to discover the hidden outcomes in the quality of life of the patients by benefiting the Short Form (SF) 36 and Quadrad12 questionnaire data. The number of the patients is 3,909 and the algorithms used and tested are Apriori, Frequent Pattern (FP) Growth, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Self-Organizing Map (SOM), and KMeans.

Association rule mining algorithms are generally Apriori, FP-Growth, and their derivatives in the usages of the data mining applications in the literature. In this study, it has been observed that the response of SF36 and Quadrad12 questionnaires are unsuitable for Apriori and FP-Growth because the logic of the association rule mining algorithms is based on 0-1 or True-False structure. In other words, it means the selections in SF36 and Quadrad12 questionnaires do not have only the true or false values; Data needs a preprocessing for fuzzy selections. Nominal values can be used correctly. Thus, Apriori and FP-Growth algorithms have been tested, but the results have not been successful. As an alternative solution, clustering algorithms have been used to obtain the rules. The centroids of the clusters show the common characteristic of the cluster; by the means of this property, clustering algorithms have been preferred.

The contributions of this study are twofold. Firstly, a rule set has been obtained for the Reflux Disease peculiar to the Turkish population. Secondly, this study declares that clustering algorithms can be used for association rule mining problems for fuzzy data. The tests show that the conventional KMeans algorithm is more successful than the conventional association rule mining algorithms at discovering a rule set for the fuzzy data by evaluating the centroids as hidden characteristics of the bag of the data in the clusters.

In this paper, Section 2 mentions the related works; algorithms and methods are given in detail in Section 3; finally, results and their discussions are in Section 4, and Section 5 mentions the conclusion respectively.

2. RELATED WORKS

Data mining is a combination of a multidisciplinary area containing statistics subjects, database management systems, machine learning algorithms, visualization approaches, and artificial intelligence [11]. Data mining is used to provide preliminary information required for knowledge-driven and data-driven expert systems for the topics of study by exploring the

information in large datasets [12]. It is likely to discover knowledge based on data analysis from various perspectives with data mining; text mining, web mining, medical decision support systems, financial predictions, forecasting studies at all engineering fields, etc. can find application in a wide range of fields including subject titles [13]. In addition, data mining is part of the overall process for discovering information in databases, as the science and technology of data discovery to discover previously unknown patterns. In up-to-date information systems, large-scale knowledge in data warehouses also includes patterns to explore. The availability and abundance of this information make data mining very important and necessary [14]. In data mining, there are many techniques such as optimization algorithms, Association Rule Mining (Market Basket Analysis), supervised learning, and unsupervised learning algorithms. These techniques including Apriori, FP-Growth, Genetic Algorithms, Decision Trees, Deep Learning, Support Vector Machines, Clustering Algorithms, etc. are implemented to discover hidden knowledge that can guide the decisions at a sector and for processing knowledge received from various fields to provide information and for decision-making [15].

If medical studies are focused on; the aim of a study in 2009 [16] is to highlight a patient profile that lies with the association rules mining, underlining the groups of parts that often occur together in the 1919 group process. Moreover, the association rules are the determination of relationships and associations between the special values of the variables in large data sets. This approach enables to reveal patterns hidden in the big data sets of analysts and researchers [17]. Its usefulness, easy understanding, and revealing of all possible patterns are the strengths of the technique. But it has also a weakness that reveals all possibilities because decision-makers must overcome a large amount of knowledge to evaluate all these possibilities, which is difficult and time-consuming [18]. Nowadays, data collection has been accelerated with the information technologies used. This means more information and raises a challenge for physicians to analyze.

With the Association Rule Mining, evaluations are made mostly for the analysis of the consumption habits of the customers. Also, it supplies the identification of the products or product groups that tend to occur together in the purchasing processes [16]. In the Association Rules technique, algorithms such as Apriori and FP-Growth are used. These algorithms find rules based on validation values and use first-order logical representations. These values include different options such as class index, class verification threshold value, verification value, frequency threshold value, missing values and data, noise threshold value, number patterns, repeat patterns, ROC analysis, and output values [19]. Apriori is an inductive logic programming algorithm that looks for the highest values of the validation evaluation function. It explores the simplest different validation measures weighted with relative accuracy. A verification criterion shows the part of a rule's examples of the unexpected and expected counter. In the algorithm, two expected and observed probability values are calculated. Apriori extracts first-order rules and rules for association with other programs are used in mining tasks [20]. The Apriori algorithm has become a standard approach in mining association rules. It was first introduced by Agrawal and Srikant [21].

It can be seen in recent studies that new approaches have been developed for analysis in medical text data due to the inadequacy of the Apriori algorithm in terms of semantic relationship [22, 23]. In another current study, a novel association rule mining algorithm has been developed by combining a MapReduce distributed computing model and association rule [24].

3. METHODS AND THEIR COMPARISONS

Prior to data collection research, Non-Interventional Ethics Committee permission numbered 17-5.1/49 was granted and Ege University Clinical Researches Ethics Committee approval dated 08/09/2017 and numbered 70198063-050.06.04 were obtained.

In literature, a statistical study about SF36 and QoLRAD questionnaires, which is implemented by Hancerlioglu, et al., has been encountered. This study has detected important outcomes; however, these outcomes contain statistical results, and deep relations could not be discovered, because there has not been any data mining approach [25]. Conversely, in our study, data mining algorithms have been focused on.

At the beginning of the study, the data is not useful to explore the quality of life of the patients; therefore, the required steps have been done to make data ready to be mined. The steps can be summarized as the database preprocesses and the data mining process. The study aims to find hidden relationships and knowledge in the database with the help of data mining methods. Association rule mining, clustering algorithms are applied to data. Apriori and FP-Growth algorithms did not give expected results for the target dataset because of the existence of high density for the "No" response. The presence of no responses leads to the extraction of useless and meaningless relationships because the targeted output is to detect the hidden links between the "Yes" responses. Therefore, clustering algorithms have been implemented and examined. In Figure 1, 4 clusters are represented, and it means that 4 separate centroids as rules. This approach has been applied to clustering algorithms tested.

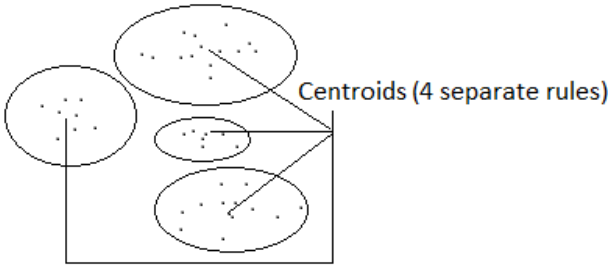


Figure 1 The match of centroids and rules

Quadrad12 and SF36 questionnaires have been stored separately and irregularly in the database. The database preprocessing has been basically done to prepare the dataset in the required structure. Before preprocessing the database, two questionnaires results were not a dataset structure where each column symbolizes a possible response and each row does a patient.

3.1. Quadrad12 and SF36 questionnaires

The questions of SF36 used in the study are given below:

- SF36_1 What can you tell about your health in general?
- SF36_2 If you compare your health with last year, how do you evaluate your health now, in general?
- SF36_3 The following items are related to your activities throughout the day. Does your health now restrict these activities? How much if it does?
- SF36_4 During the last 4 weeks, have you experienced any of the following problems with your work or other daily activities as a result of your physical health?
- SF36_5 During the last 4 weeks, have you experienced the following problems with your work or other daily activities as a result of your emotional problems (such as depression or anxiety)?
- SF36_6 During the last 4 weeks, how much did your physical health or emotional problems affect your social activities with your family, friends, or neighbors?
- SF36_7 How many pains have you had in the last 4 weeks?
- SF36_8 During the last 4 weeks, how much did your pain affect your normal work (both your housework and your out-of-home work)?
- SF36_9 The following questions are about how you feel during the last 4 weeks. For each question, please choose the best response to your feelings by considering the frequency of the last 4 weeks.
- SF36_10 How often did your physical health or emotional problems affect your social activities (such as visiting friends or relatives) during the last 4 weeks?
- SF36_11 How true or wrong are each statement below? Check the most appropriate for each statement.

As can be seen, while SF36 focuses on the health status of patients, Quadrad12 consists of questions having the aim of evaluating how much the complaints affect a patient's life. The

questions of Quadrad12 questionnaires are given below:

- Q12_1 How often did you feel tired or exhausted because of the pain/sour water coming from behind the chest during the last week or coming into your mouth / esophagus?
- Q12_2 During the past week, how often did you avoid leaning over the back of the breastbone or because of pain/sour water coming into your mouth/esophagus?
- Q12_3 During the last week how often you had to eat less than you used to eat, because of the blazing chest behind the breastbone or sucking water from your mouth/esophagus?
- Q12_4 During the past week, how often did the pain bruise from the back of the chest and/or bitter water into your mouth/esophagus prevent you from doing something with your family or friends?
- Q12_5 During the past week, how often did you feel tired or exhausted due to insomnia that is caused by the burning of the chest of the chest or bitter/sour water in your mouth/esophagus?
- Q12_6 During the last week, how did you wake up from the back of the chest and sneak from your mouth/esophagus/snuff/water?
- Q12_7 During the past week, how often did you feel very frustrated or impatient because of the burning of sour water on your back/esophagus during the last week?
- Q12_8 During the past week, how often did you find yourself awake and rested in the morning because of the painful or sour water coming from your chest/esophagus?
- Q12_9 During the past week, how often have you been worried or feared about your health because of complaining of bruising behind your breastbone or suffering/bitter water in your mouth/esophagus?
- Q12_10 During the past week, how often did you have to avoid certain foods and beverages because of the burning of the bones behind the breastbone or the bitter/sour water in your mouth/esophagus?
- Q12_11 During the past week, how often have you not been able to fulfill your daily activities (including jobs at home and outside the house) because of the burning of the chest behind the chest, or bitter/sour water in your mouth/esophagus?
- Q12_12 During the past week, how often have you not been able to fulfill your normal physical activities (including sports, full-time activities, and going out of the house) to the left because of the painful or sour water coming from your chest/esophagus?

3.2. Sub-operations in Methods

The data file formatted includes patients' information in each row and columns including the related titles of the responses. For example, in the first column and the first row, the first patient's response to the question about social security is represented. Each row for the same question includes the different kinds of social security such as "Social Security Institution", "Pension Fund", etc. Therefore, the format obtained was not able to use for data mining. The discretization method is used to turn them into the nominal form in the WEKA tool [26]. Thus, the "nominal to binary" method has been implemented in WEKA to obtain the dummy attributes. After all, the required binary matrix has been obtained. The final form of the dataset has contained 3,909 instances with 323 attributes as the binary matrix structure. Containing only binary values, the matrix has become useful for mining.

Figure 3 shows that firstly, patient responses are obtained in a view by joining the tables related to the questionnaire in the database. Secondly, the row-based database view is transformed into the column-based dataset by the "pivot" database queries; finally, nominal contents are converted into the binary attribute structure. Also, to discover hidden relationships, answer attributes with only one "1" in the data set were deleted and focused on attributes containing at least 2 "1". With this feature selection process, the number of attributes has been reduced by 12.

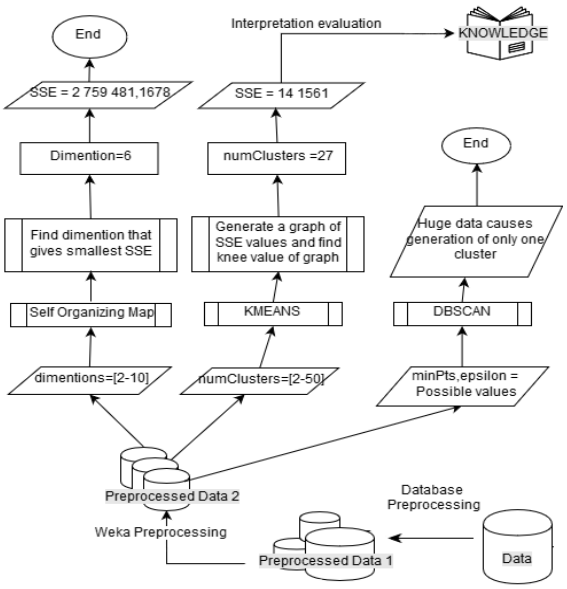


Figure 2 The entire process of the study

The next phase is to compare multiple algorithms to determine which algorithm has been giving better results. SOM is suitable as a clustering algorithm for the subject to determine rules and relationships among the instances.

Also, KMeans and DBSCAN algorithms are used to cluster data, too. In Figure 2, the entire process of the study has been figured as a flow chart. The comparisons have been accomplished with the help of the sum of squared error (SSE) for each applied algorithm. This error has been calculated with the formula in Eq. 1.

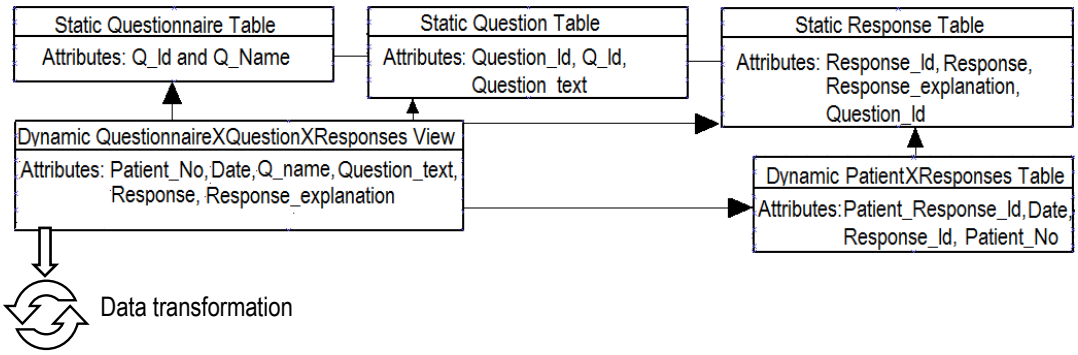
$$\mathrm{SSE} = \sum_{k=1}^{K} \sum_{i=1}^{I} (x_i - c_k)^2, \quad (1)$$

where K is the total number of clusters, l is the total number of instances and c is the current centroid.

KMeans is one of the popular unsupervised learning algorithms. The algorithm collects data that has many similarities around center points. After setting the number of clusters, each center allocates nearby data points. The algorithm is represented in Figure 5 as a pseudo-code. SSE values help the determination of the proper cluster number for the related dataset. For the study, after each cluster number between 2 and 50 observed, the pattern having 27 clusters has been decided to be used according to the SSE values given in Table 1.

The tests of KMeans clustering have been applied as KMeans++ version in WEKA 3.9 tool because of its accuracy advantage [27]. In tests, the significant parameter, "numClusters" has been changed between 2 and 50 to find which number of clusters makes the best pattern.

SSE values lead to compare the results of each number of clusters applied. Figure 4 shows the distribution of the SSE values to evaluate the optimum K number.



Patient	q1r1	q1r2	q2r1	q2r2	q2r3	q2r4	q3r1	q3r2	q4r1	q4r2	q4r3	q5r1	….
1	0	0	0	1	0	0	0	1	0	1	0	1	….
2	1	1	0	1	1	1	1	1	0	1	0	0	….
….	….	….	…..	….	….	….	….	….	….	….			….
3,909	1	0	1	0	0	0	1	1	1	0	1	0	….

Figure 3 Data transformation phases (q: question, r: response)

Table 1
SSE values for each cluster number

# of clusters	SSE	# of clusters	SSE
2	163003	27	141561
3	159846	28	141699
4	156112	29	141388
5	157383	30	141268
6	153617	31	140737
7	151553	32	140585
8	151044	33	140662
9	148529	34	140070
10	148672	35	139711
11	147512	36	139166
12	148082	37	139030
13	146555	38	139105
14	146379	39	138803
15	146145	40	138788
16	146018	41	138399
17	144917	42	138579
18	144671	43	138182
19	144535	44	137808
21	143390	45	138154
22	143067	46	137586
23	142674	47	137482
24	142588	48	137545
25	142042	49	137482
26	141842	50	137199

As can be seen in Figure 4, as cluster numbers increase, the SSE value is decreasing. It can be determined that increasing cluster numbers cause more accumulation and it decreases SSE value. Therefore, that elbow point [28] of the graph is significant to us which is 27 in the study. The SSE value obtained with 27 clusters has been compared to other algorithms' SSE values to choose the proper one. To detect 27 rules from this study, this pattern has been used.

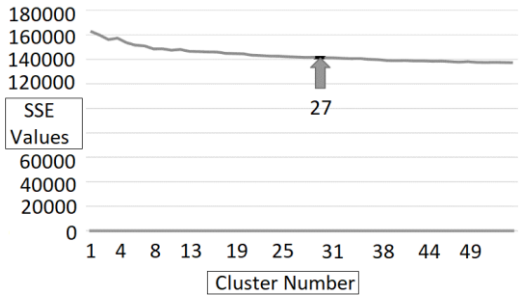


Figure 4 Graph representation of SSE values

In Figure 6, each column represents a cluster and each cluster includes attributes that are represented by 1. Thus, a data structure that can keep each cluster has been preferred and it has been possible that each instance is reachable from another instance easily. Also, another property of this structure is that the first element will be a string while the others are integers.

Function K-MEANS ($\{\overrightarrow{x_1}, \dots, \overrightarrow{x_N}\}$, K)
1: $(\overrightarrow{s_1}, \overrightarrow{s_2}, \dots, \overrightarrow{s_k}) \leftarrow$RandomCentroidSelector($\{\overrightarrow{x_1}, \dots, \overrightarrow{x_N}\}$, K)
2: **for** $k \leftarrow 1$ **to** K
3: **do** $\overrightarrow{\mu_k} \leftarrow \overrightarrow{s_k}$
4: **while** ending criteria has not met
5: **do for** $k \leftarrow 1$ **to** K
6: **do** $\overrightarrow{\propto_k} \leftarrow \{\}$
7: **for** $n \leftarrow 1$ **to** N
8: **do** $j \leftarrow argmin_{j'} \mid \overrightarrow{\mu_{j'}} - \overrightarrow{x_n} \mid$
9: $\propto_j \leftarrow \propto_j \cup \{\overrightarrow{x_N}\}$
10: **for** $k \leftarrow 1$ **to** K
11: **do** $\overrightarrow{\mu_k} \leftarrow \frac{1}{|\propto_k|} \sum_{\vec{x} \in \propto_k} \vec{x}$
12: **return** $\{\overrightarrow{\mu_1}, \dots, \overrightarrow{\mu_k}\}$

Figure 5 The pseudo-code of the KMeans algorithm

Two arrays used in the algorithm are given in Figure 6. In the next step, the data needs to be prepared as a text file. As it is expressed in the figure, there are two types of arrays. To split them into two arrays, the comma mark has been used in the data file and space characters have been cleared to transform into the substrings zeros and ones. Responses are kept in a string array as can be seen in Figure 6.

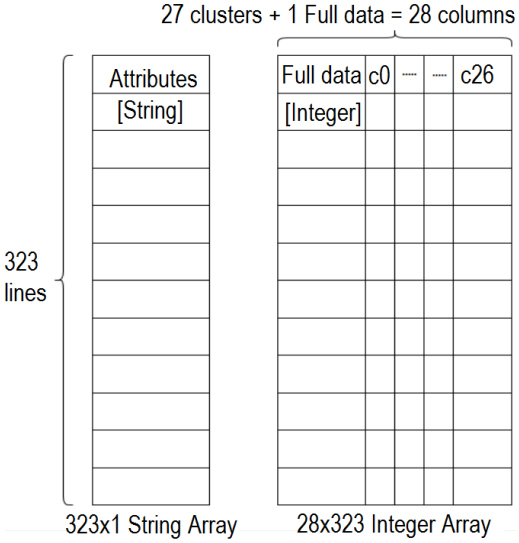


Figure 6 The data structure of the study

Each response corresponds to each cluster's attribute row. Therefore, the algorithm is developed to scan each column and its responses then write them in a text file. The algorithm finds 27 rules in a short time. The algorithm is summarized in Figure 7.

```
Function RuleFinder for K-Means ()
1:    file = readFile(example.txt)
2:    parsed = file.Split(',')
3:    foreach (string line in file)
4:        Answers[] = Parsed[0]
5:        for i←1 to 28
6:            integerArray[] = Parsed.Substring()
7:        end
8:    end
9:    iterateOver ( integerArray )
10:       if ( integerArray.element ==1 )
11:           takeCorresponding ( answers.element )
12:           StreamWrite(rules.txt)
13:       end
14:   end
```

Figure 7 The pseudo-code of the "Rule Finder" method

DBSCAN is another clustering algorithm that calculates each point's distance with its neighbors with a threshold value given and clusters the closest ones. To implement this algorithm two parameters are necessary to be checked: epsilon and minPts. Epsilon adjusts the radius of the area that will be observed and minPts designates the minimum number that will be included inside the area. DBSCAN is used to determine arbitrarily shaped data. DBSCAN is explained in detail in Figure 8 [29].

```
Function DBSCAN(K, minpts, ε)
1: for i ← 1 to K
2:    visit each point i
3:    E = ε_neighborhood of i
4:    if (E < minpts)
5:        noise += i
6:    else
7:        new cluster += i
8:        foreach ( point in E )
9:            E' = ε_neighborhood of point
10:           if size(E') >= minpts
11:               E += points
12:           end
13:       end
14:   end
15: end
```

Figure 8 The pseudo-code of DBSCAN algorithm

Table 2
Data distribution among lengths after implementing SOM algorithm

	3x3	4x4	5x5	6x6	7x7
(0,0)	10	2	1	1	2
(0,1)	8	5	9	1	0
(1,0)	3816	2	0	0	3
(1,1)	4	2	0	3	5
(2,0)	1	4	0	0	0
(2,1)	8	3792	1	0	8
(0,2)	18	2	9	9	7
(1,2)	35	10	1	2	0
(2,2)	9	2	13	1	2
(3,0)		29	0	2	3
(3,1)		2	0	0	3
(3,2)		2	5	12	0
(0,3)		13	1	5	0
(1,3)		1	1	3	0
(2,3)		13	11	1	0
(3,3)		28	0	26	0
(4,0)			5	0	0
(4,1)			15	0	4
(4,2)			0	0	0
(4,3)			1	3	2
(0,4)			25	25	0
(1,4)			3	2	1
(2,4)			3794	6	0
(3,4)			10	0	0
(4,4)			4	3779	0
(5,0)				0	0
(5,1)				2	0
(5,2)				0	0
(5,3)				3	0
(5,4)				0	8
(0,5)				5	0
(1,5)				0	0
(2,5)				1	0
(3,5)				0	25
(4,5)				13	8
(5,5)				4	3782
(6,0)					2
(6,1)					3
(6,2)					0
(6,3)					15
(6,4)					12
(6,5)					1
(0,6)					0
(1,6)					0
(2,6)					1
(3,6)					1
(4,6)					1
(5,6)					6
(6,6)					4

Results have been examined by changing parameters and DBSCAN has not fit the huge data. It has placed 3,909 data into the same

cluster, or all data have been determined as noise data, and any pattern having clusters more than one has not occurred.

The number of clusters is not asked in DBSCAN as opposed to KMeans, as it is introduced in some studies [30]. However, it does not suit datasets that have large density differences among data. So that minPts and ε cannot be determined appropriately. Additionally, SOM is a model based on a neural network that is trained by unsupervised. Usage of the SOM algorithm gives the advantage by the means of visualizing the multidimensional data with two-dimensional views. The algorithm shows similarities between the neighbors of the neurons with the distribution of neurons on the map. The steps of the algorithm are given as a pseudo-code in Figure 9 [31]. After training, each patient has been allocated by a nearby center point that is called the winner.

Algorithm SOM()

1: Initialize each node's weight
2: **for** $i \leftarrow 1$ **to** N
3: Load data from given file
3: Choose a vector from the set of training data
4: Find **winner** node (Most likely to input vector)
5: Calculate neighborhood of winner node
6: **end**

Figure 9 The pseudo-code of the SOM algorithm

In an effort to compare algorithms to choose the one that gives better results, SSE values have been calculated for the SOM algorithm for each length between 3x3 and 10x10 as square matrix maps. SSE values for the lengths of 6 and 7 are given in Table 2. The distribution of nodes between the lengths of 3 and 7 is given in Table 2.

Table 3
SSE values for two different lengths

SSE value for 6x6	SSE value for 7x7
2 739 954,39069696	2 725 156,4785282

In Figure 10, SOM maps for 3x3, 4x4, 5x5, 6x6, and 7x7 are visualized, respectively. Especially the black clusters in the middle parts of the maps represent the clusters of the pattern with high error, which have higher SSE values than the general SSE average of the patterns.

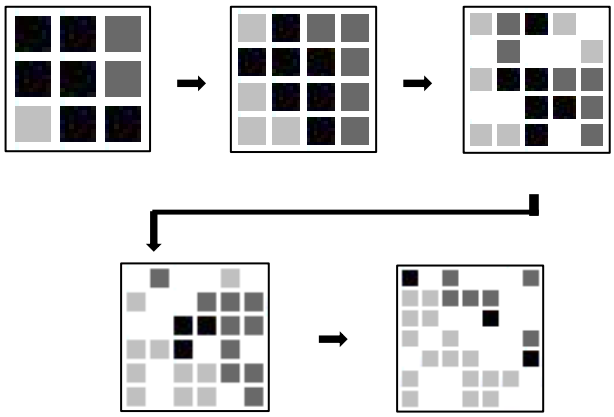


Figure 10 3x3, 4x4, 5x5, 6x6, and 7x7 SOM maps tested, respectively.

The number of these clusters increases as the map size gets smaller. Likewise, as the size of the map increases, the number of empty clusters also increases. The number of empty clusters is tolerable, but as can be seen from Table 3, high SSE values of these patterns have been obtained. Therefore, the SOM approach has been not preferred.

4. EXPERIMENTAL RESULTS AND OUTCOMES

Algorithms having different approaches have been applied for the dataset and results have been examined. DBSCAN algorithm has created only one cluster from 3,909 data. The data was excessive for efficient clustering in DBSCAN. FP-Growth and Apriori algorithms have been found incompatible for the dataset as mentioned in the previous section. Exclusion of some data like these may help extract more meaningful rules. Similarly, the SOM algorithm has given higher SSEs values than the KMeans algorithm. That is why the KMeans algorithm has been preferred to maintain the study and obtain the rules. In Figure 11, the chart of the patient number in the 27 clusters obtained as the most optimum pattern shows that the numbers among the clusters do not have any clear local density. The centroids of these 27 clusters are listed below. In other words, the outcomes discovered are given.

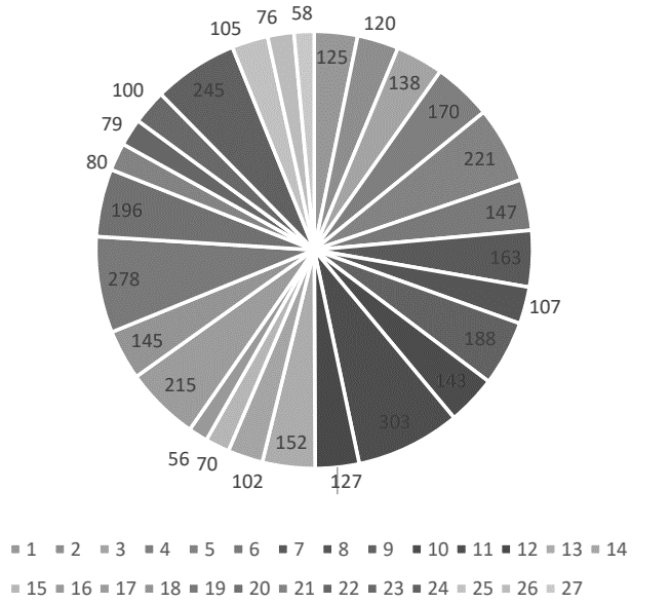


Figure 11 The chart of the patient number in the 27 clusters

- The first cluster includes men from İzmir. Those 125 patients as independent from age and level of education are limited in their daily life and work-life by complaints. They are affected by the problems given. We examined that they gave average answers to all questions. Their health status is good in general.
- The second cluster is independent of age, gender, and city, comprised of secondary and high school graduates. And they have the same complaints for one or more than one year.
- Patients' health in the third cluster deteriorates. We conclude that with the help of the high answering ratio of SF36_2 is "Almost the same". As we know, those patients come with complaints of suffering from sour water in their mouths. As they tell their health is almost the same as last year it means they keep suffering from similar problems for one year. However, other questions answered as never shows that they do not get affected badly in their life.
- The fourth cluster includes university graduates that are independent of the city, age, gender. Their answers for SF36_9 and SF36_11 display that they do not feel disturbed much from their complaints. They do not consider their health is in danger even if it remains the same for one year, this group is aware of complaints and bears it. They claim they are not badly affected by complaints in daily life.
- The fifth cluster that includes 221 people involved in the research is affected by complaints in their daily life. This group is independent of age, gender, and city, and education status. But we understand complaints spring currently because patients did not claim their health is the same as last year.
- The sixth cluster includes 147 people who are university graduates and ages between (36- 54], who are generally affected badly by the complaints. They assert that their health is worse than last year. So, it can be understood they came to the doctor after one year suffers. Complaints restrict their work and daily life for one or more than one year.
- We have observed that the health status of 163 women that are university graduates has deteriorated in the seventh cluster. Complaints restrict their daily and work life. They also stated their health is worse than last year which means they suffer for one or more than one year.
- The eighth cluster includes secondary and high school graduates independently from age, gender, and city generating a group of 107. They suffer from complaints currently and less than one year. We observe education level is affecting patients' psychology and health.
- The ninth cluster is a group of 188 that consists of primary school graduates independently from age, gender, and city. Their health is getting worse and they suffer one or more than one year with the complaints. Those patients complain because they suffer in their daily life and work life.
- In the tenth cluster it is observed that male patients who are secondary or high school graduates aged 36 to 54 years in İzmir are suffering from the restricted capability of activities such as jogging, carrying the heavy thing, etc.
- The eleventh cluster is the biggest with 303 people, is independent of age, gender, education status and city suffer from complaints currently. The group is restricted in their daily life and career. It is claimed that they cannot focus on their work.
- The twelfth cluster of 127 people that are between ages 36 and 54, is feeling tired and

angry for one month. They suffer less than one year.
- The thirteenth cluster comprises 152 women in İzmir; feel their health situation is bad in general. They are restricted in daily and business life by the complaints. They seem to be anxious and exhausted during the last month.
- 102 women from İzmir, are restricted much because of complaints in the fourteenth cluster. They claim that it is hard to walk 1-2 km. That group believes their health is bad. They seem pessimistic about their health. It is observed that they are limited in their work and social life because of complaints.
- The fifteenth cluster of patients' health, between ages 36 and 54, is observed to deteriorate. They state that their health is worse than last year. Even low effort requiring activities are restricted.
- The sixteenth cluster contains university graduates between ages 18 and 36, who are observed to be stronger than primary school graduates mentally. It is understood after analyzing their answers. Even if they suffer, they claim their health is not bad.
- The seventeenth cluster of patients between ages 18 and 36, which is independent of the city, gender, and education are mentally and socially limited by the complaints. This group consists of 215 patients.
- The eighteenth cluster comprises 145 patients that independent from the city, education status, and gender are not specified are not affected by complaints in their life. But they claim their complaints remain one or more than one year.
- The nineteenth cluster contains primary school graduates that suffer from complaints currently and they visited the doctor directly. They are not restricted in their daily or business life. City, gender, and age are independent.
- The twentieth cluster of 196 women in İzmir came with similar complaints. They claim they suffer from sour water or heartburn. But after the survey, they are found healthy. They have current effects of reflux but their daily or business life is not affected.
- 80 patients are healthy from cluster 21. They are independent of age, city, gender, and education status. They came with complaints that are similar to reflux, but they have not affected not restricted in their life.
- 79 patients that are secondary or high school graduates, answers Q12_10 as "always" in cluster 22. So, they limit themselves from eating some foods to avoid pain. They visited the doctor because their health is worse than last year. It can be understood they suffer from complaints one or more than one year and their health deteriorates. They are restricted from their daily activities. Also, their personal life is affected by complaints.
- The twenty-third cluster of 100 primary school graduates that are independent of age, gender, and city, have a huge list of complaints. Their psychology is not well because they claim they have never felt bouncy in their life and they always feel exhausted, tired, and upset. They believe they catch disease easier than others. Their daily activities, life, and business life are affected badly. They state they suffer from complaints for one or more than one year. Their answer for Q12_12 takes attention; they tell that also moving around the home is hard for them.
- The twenty-fourth cluster of 245 university graduates filled questionnaires, and their complaints are currently disturbing them, but they do not complain or feel restricted yet.
- The twenty-fifth cluster of 105 patients between ages 36 and 54, claims they have stress in their life last one month (SF36_5). Those patients are middle-aged, and the restriction of their daily activity may be related to their age. They feel very restricted by simple activities such as 1-2 km of walk.
- The twenty-sixth cluster of primary school graduates that are women and between ages 54 and 72 are restricted in their daily life. They had emotional problems during the last month.
- The twenty-seventh cluster of 58 patients that are primary school graduates are restricted and complaining from their problems in their health currently. They stated their business is affected badly. Also, their health affects their daily activities.

Dataset consists of patients from the Aegean region intensely. Among 3909 patients, there are 1338 women, 1026 men, and 1545 not specified. It shows intensities between the ages of 36 and 54.

In conclusion, we observed that education affects reflux patients' situations. It has been observed that 1114 primary school graduates participated in the questionnaires.

5. CONCLUSION

Knowledge discovery in the medical database is completed as aimed in the study. For future work, the diagnosis will be analyzed as well. The research is concluded with twenty-seven clusters. 3,909 data are clustered into 27 clusters with the KMeans algorithm that is found effective after comparing among other algorithms with their SSE values. After analyzing clusters, we end up with some results, our study will help patients to understand their pain reason if they are suffering from reflux or psychological reasons involved. As the process of the reflux center is examined, it is understood that process is painful and hard to overcome. So, the study aims to target only real patients instead of every patient suffering from unnecessary tests. The process is painful as we expressed before, so healthy patients are affected psychologically after tests and they feel weaker mentally. Already some patients cannot handle tests 24 hours and leave it in 4 hours or less. So, our aim is to decrease the number of patients who will take the test, successfully. We believe that computers and algorithms may be more efficient than doctors before examining the health status of patients. As we know it is impossible for doctors to know and remember similar cases with the relevant patient; however, for computers, it takes minutes to calculate and give a result according to previous similar occurrences. So those data will be used to determine new patients' health status without applying several tests.

Acknowledgments

This study is supported by Ege University Scientific Research Projects Coordination Unit. Project Number: 2.101.2015.0078. The authors would like to thank Prof. Dr. Serhat Bor in Ege University Faculty of Medicine Reflux Research Center. In addition, they would like to thank the reviewers for all useful and instructive comments on our manuscript.

Funding

The authors received no specific funding for this study.

The Declaration of Conflict of Interest/ Common Interest

No conflict of interest or common interest has been declared by the authors.

Authors' Contribution

All authors have contributed in experimental study and writing of the manuscript equally.

The Declaration of Ethics Committee Approval

The authors declare that this document has the ethics committee approval as Ege University Medical Faculty Ethical Committee Confirmation Number: 2017-5.1/49.

The Declaration of Research and Publication Ethics

The authors of the paper declare that they comply with the scientific, ethical and quotation rules of SAUJS in all processes of the paper and that they do not make any falsification on the data collected. In addition, they declare that Sakarya University Journal of Science and its editorial board have no responsibility for any ethical violations that may be encountered, and that this study has not been evaluated in any academic publication environment other than Sakarya University Journal of Science.

REFERENCES

[1] N. A. Farooqui, and R. Mehra, "Design of a data warehouse for medical information system using data mining techniques," Fifth International Conference on Parallel, Distributed and Grid Computing, pp. 199-203, 2018.

[2] S. Mishra, and M. Panda, "Artificial intelligence in medical science," In

Intelligent Systems for Healthcare Management and Delivery, pp. 306–330, 2019.

[3] B. Allen Jr, S. E. Seltzer, C. P. Langlotz, K. P. Dreyer, R. M. Summers, N. Petrick, D. Marinac-Dabic, M. Cruz, T. K. Alkasab, R. J. Hanisch, W. J. Nilsen, J. Burleson, K. Lyman, and K. Kandarpa, "A road map for translational research on artificial intelligence in medical imaging," Journal of the American College of Radiology, vol. 16, no. 9, pp. 1179–1189, 2019.

[4] G. Currie, K. E. Hawk, E. Rohren, A. Vial, and R. Klein, "Machine learning and deep learning in medical imaging: intelligent imaging," Journal of Medical Imaging and Radiation Sciences, vol. 50, no. 4, pp. 477–487, 2019.

[5] S. Agrawal, B. Singh, R. Kumar, and N. Dey, "Machine learning for medical diagnosis: A neural network classifier optimized via the directed bee colony optimization algorithm," U-Healthcare Monitoring Systems, vol. 1, pp. 197–215, 2019.

[6] V. Levshinskii, M. Polyakov, A. Losev, and A. V. Khoperskov, "Verification and validation of computer models for diagnosing breast cancer based on machine learning for medical data analysis," Creativity in Intelligent Technologies and Data Science, pp. 447–460, 2019.

[7] C. Le Berre, W. J. Sandborn, S. Aridhi, M. D. Devignes, L. Fournier, M. Smaïl-Tabbone, S. Danese, and L. Peyrin-Biroulet, "Application of artificial intelligence to gastroenterology and hepatology," Gastroenterology, vol. 158, no. 1, pp. 76–94, 2019.

[8] J. K. Ruffle, A. D. Farmer, and Q. Aziz, "Artificial intelligence in gastroenterology," Precision Medicine for Investigators, Practitioners and Providers, pp. 343–350, 2020.

[9] L. Q. Zhou, J. Y. Wang, S. Y. Yu, G. G. Wu, Q. Wei, Y. B. Deng, X. L. Wu, X. W. Cui, and C. F. Dietrich, "Artificial intelligence in medical imaging of the liver," World Journal of Gastroenterology, vol. 25, no. 6, pp. 672–682, 2019.

[10] M. McDonnell, R. Harris, T. Mills, L. Downey, S. Dharmasiri, R. Felwick, F. Borca, H. Phan, F. Cummings, and M. Gwiggner, "P384 High incidence of hyperglycaemia in steroid treated hospitalised inflammatory bowel disease (IBD) patients and its risk factors identified by machine learning methods," Journal of Crohn's and Colitis, vol. 13, no. 1, pp. 299–300, 2019.

[11] J. Han, and K. Micheline K. "Data Mining, Southeast Asia Edition: Concepts and Techniques," Morgan Kaufmann, pp. 30-39, 2006.

[12] U. Fayyad, G. Piatetsky-Shapiro, and S. Padhraic, "From data mining to knowledge discovery in databases," AI magazine, vol. 17, no. 3, pp. 37-54, 1996.

[13] Y. K. Jain, V. K. Yadav, and G. S. Panday, "An efficient association rule hiding algorithm for privacy preserving data mining," International Journal on Computer Science and Engineering, vol. 3, no. 7, pp. 2792-2798, 2011.

[14] O. Maimon, and R. Lior, "Data mining with decision trees: Theory and applications," World Scientific New Jersey, pp. 42-51, 2008.

[15] T. Wu, and L. Xiangyang, "Data Storage and Management, The Handbook of Data Mining," Ed. Nong Ye, New Jersey: Lawrence Erlbaum Associates, Inc., pp. 393-407, 2003.

[16] P. Giudici, and S. Figini, "Applied Data Mining for Business and Industry,Second Edition," Wiley Publicition, West Sussex, pp. 23-29, 2009.

[17] R. Nisbet, J. Elder, and G. Miner, "Handbook of Statistical Analysis and Data Mining Applications," Elsevier Inc, Burlington, pp. 18-26, 2009.

[18] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms," John Wiley & Sons J. B. Speed Scientific School, University of Louisville IEEE Computer Society, Sponser, pp. 52-66, 2003.

[19] J. Arora, N. Bhalla, and R. Sanjeev, "A review on association rule mining algorithms," International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, no. 5, pp. 1246-1251, 2013.

[20] J. Nahar, T. Imam, K. S. Tickle, and Y. P. Chen, "association rule mining to detect factors which contribute to heart disease in males and females," Expert Systems with Applications, vol. 40, no. 4 pp. 1086-1093, 2013.

[21] R. Agrawal, and R. Srikant, "Fast algorithms for mining association rules," In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile. Citeseer, pp. 487–499, 1994.

[22] M. Thamer, S. El-Sappagh, and T. El-Shishtawy, "A Semantic Approach for Extracting Medical Association Rules," International Journal of Intelligent Engineering Systems, vol. 13, no. 3, pp. 280-293, 2020.

[23] L. Wang, J. Li, T. H. Zhou, and W. Q. Liu, "Association Rules Extraction Method for Semantic Query Processing Over Medical Big Data," In Proceeding of Asian Conference on Intelligent Information and Database Systems, Singapore, Springer. pp. 109-120, 2020.

[24] S. Zhou, J. He, H. Yang, D. Chen, and R. Zhang, "Big Data-Driven Abnormal Behavior Detection in Healthcare Based on Association Rules," IEEE Access, vol. 8, pp. 129002-129011, 2020.

[25] S. Hançerlioğlu, Y. Yıldırım, and S. Bor, "Validity and reliability of the Quality of Life in Reflux and Dyspepsia (QoLRAD) questionnaire in patients with gastroesophageal reflux disease for the Turkish population," The Turkish Journal of Gastroenterology, pp. 511-516, 2019.

[26] S. Kodati, R. Vivekanandam, and G. Ravi, "Comparative analysis of clustering algorithms with heart disease datasets using data mining Weka tool," In Soft Computing and Signal Processing, Springer, Singapore, pp. 111-117, 2019.

[27] D. Arthur, and S. Vassilvitskii, "KMeans++: The advantages of careful seeding,". Stanford, pp. 1-11, 2006.

[28] P. Bholowalia, "EBK-Means : A clustering technique based on elbow method and KMeans in WSN," International Journal of Computer Applications, vol. 105, no. 9, pp. 17–24, 2014.

[29] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN," ACM Transactions on Database Systems (TODS), vol. 42, no. 3, pp. 1-21, 2017.

[30] T. Bilgin, and Y. Çamurcu, "Applied Comparison of DBSCAN, OPTICS and KMeans Clustering Algorithms," Journal of Polytechnic, vol. 8 no. 2, pp. 139–145, 2005.

[31] T. Kohonen, "Essentials of the self-organizing map," Neural networks, vol. 37, pp. 52-65, 2013.