Usage of Kernel K-Means and DBSCAN Cluster Algorithms in Health Studies: An Application

Özge Pasin¹ and Handan Ankarali²*
¹Biostatistics Department, Istanbul University, Turkey
²Duzce University Biostatistic and Medical Informatics Department, Turkey

Abstract

With clustering methods variable and individuals which have similar characteristics may be collected in a group. Although clustering methods have many applications, there are limited studies in health researches in our country. While the purpose of this study is to introduce Kernel K-means and DBSCAN clustering algorithms and how and which cases should be correctly used. At the same time, different clustering algorithms results which can be applied on a real data set were compared. According to the evaluations, the Kappa coefficients were statistically significant and it's degree are low. In terms of data sets the most convenient and fastest algorithm is Kernel K-Means clustering algorithm. The results obtained by Kernel K-Means gave the most accurate decisions in terms of the distribution of the groups among Framingham risk groups crosstables. As a result, with taking into account the criterion of clinical information it is thought that the examination of data distribution and characteristics of data sets clustering algorithms can be used as a diagnostic tool for the planning and diagnosis of diseases in the field of health.

Introduction

Today’s science accepts the information which is based on the evidences. The substance of data come along with some calculations which is used by statistic methods and principles. Information and technology improve and increase with avoidably fast and this is known by everybody. To get beneficial and useful results from the dense information clustering, using the advanced and wide statistic methods become almost necessary. The theory of technology’s developments after the internet cause that suggested statistic method is made practicality faster. Then using these methods help us to understand complex information better and to interpret the real world better [1-3].

When the informations of so many individuals with so many characteristics exists, methods which evaluate the situations better, are collected under the general title of ‘Data Mining’.

Like in every area, in health area records are transferred to computers in other words to the data base. So, datas can be kept for a long time and it becomes useful. Also in the course of time a lot of information can be reached easily. When data base are made according to a plan, it helps that a lot of new hypothesis rise and are tested [1-3].

There are a lot of algorithms which take place under the title of data mining and are used for estimating, grouping, classifying and diagnosing which have a lot of application area in recent years. Also main objectives of a lot of hypothesis of health research are one of the algorithms or some this algorithms. Although diagnosing, classifying or estimating methods are used commonly in other areas like health areas the methods which are developed so as to grouping or clustering to variable and individuals are seen that they aren’t used commonly. This results give us an impression that the researchers have limited information about how to and where to use and how to interpret the information which is taken by the help of the clustering methods [1-3].

Also except for some clustering algorithms which exists in commonly used statistic package programme, new clustering algorithms which is developed especially in recent years, are limited in use in literatüre [1-3].

Material and methods

Clustering

Clustering is the action of collecting similar characteristics. While the objects which is in the same cluster are similar to each other in terms of characteristics which are studied. Objects which take place in different clusters aren’t similar to each other in terms of such characteristics [1-3].

In other words homogeneous groups are obtained by the help of clustering. This forms the principle of clustering analysis. This analysis is unsupervised. Because a data set in which individuals exists in a complex way and nobody are known to in which group they will enter,by the help of clustering analysis homogenous groups are emerged. So, many objects are divided into realistic groups. By the help of these methods not only objects but also characteristics can be clustered. Result of this action is so similar to factor analysis. With one side it clusters many characteristics and it helps size reduction. As in practice to cluster characteristics, methods like factor analysis and structural equation modelling is used, clustering analysis is not needed [1-3].

Unconsciously we often cluster person or people according to definite criteria in area of everyday life. For example in childhood term,
to distinguish cat and dog, plant and animals easily some characteristics are considered.

Accordingly, we develop clustering outline progressively in our subconscious. Clustering analysis are used in various different disciplines with different form and objective. For example it can be used for objectives like to identify the geographic distribution of anillness, to cluster of a car accident, to govern of hospitals stuffs, to treat the conditions of a hospital, to time of an ambulance service, to diagnose of an illness, to identify the obese people groups, to examine similar illnesses [1-3].

Clustering analysis is sometimes used as a space analysis. It is seen that lossistic regression is used with combining many statistic methods like discriment analysis, one side variance analysis. In a situation like this firstly by the help of the clustering analysis, homogenous groups are made then with the other methods groups can be analyzed in a best way in terms of which groups may be decomposed or in each homogenous groups might be made separate action [1-3].

There are some basic characteristics of an effective and accurate clustering algorithms. A suitable clustering algorithms should explore clustering configurations which have different shapes and sizes with scanning in one go and also it should be practicable for all kinds of data in terms of quality and quantity. An effective clustering method should be suitable for both big and small database without distinguishing the size of the database. This is also shows that whether clustering algorithms has a characteristics of scalability or not.

An effective algorithms should know what to do towards an effective, noise datas and shouldn’t be influenced. Besides of the referred criterias, an effective clustering algorithms should be easily practicable, interpretable, functional and obvious [1-3].

**Kernel K-Means Clustering Algorithm**

Kernel K-means clustering algorithm is an extension of the classical K-Means non-linear method. Firstly it developed by Girolami in 2002 iteratively. In this model entry point in multidimensional space is mapping to in 2002 iteratively. In this model entry point is produced between X and Y’s image. This point can be calculated using (X,Y) formula (55,91).

The formula (X,Y) function is a kernel function that provides K:DXD→R. In below there is a list of some of the standard kernel functions [4-6].

- **p degrees of polynomial kernel:** K (X,Y) = (X*Y + c)^p, p: positive integer and cER
- **Gaussian (RBG) kernel:** K (X,Y) = exp(-(||X-Y||^2)/(2σ^2)) σR is a Gaussian kernel parameter.
- **Sigmoidal kernel:** K (X,Y) = tanh(α(X*Y) + b), a ve beR
- **Linear kernel:** K (X,Y) = X*Y

In iterative process, for X i and X j points it is required to calculate K(X i,X j). Thus the matrix H={K ij}n*n, which is called kernel matrix, is obtained.

n is the size of the data set. Kernel matrix is calculated and recorded previously. The time and space complexity is O(n^2). This is the disadvantages of Kernel K-Means clustering method. This complexity shows that for very large databases this clustering algorithms does not give proper results. And another disadvantages of this method is the need of determine the number of clusters initially and the core points, kernel functions are affects clustering results a lot.

In order to looking operation stages of method, it is assumed that a data like D = {X1, X2, ..., XK}. K is the number of clusters that required and M(k) is starting core point, Π(k) shows starting partition result, M(k) = {m(1), m(2), ..., m(K)} are the points in the entry space [4-6].

Kernel K-Means clustering method takes D, K ve Π(k) parameters as input parameter and it gives Π(k) parameters as the output parameter [4-6].

In order to minimize criter function, following function is used [4-6].

In the above formula μi is shows the average of Ci cluster and it is calculated using the following formula

\[ J = \sum_{i=1}^{K} \sum_{x \in C_i} \| \phi(x) - \mu_i \| \]

\[ \mu_i = \frac{\sum_{x \in C_i} \phi(x)}{||C_i||} \]

In a moved space the distances between \( \phi(x) \) and \( \phi(X) \) data points is

\[ \| \phi(x) - \phi(X) \| = \phi(x) - 2\phi(x)\phi(X) + \phi(X) = K(x,X) - 2K(x,X) + K(x,X) \]

\[ \| \phi(x) - \mu_i \| \] is calculated using the following formula by knowledge of \( \phi(.) \) [4-6].

\[ \| \phi(x) - \mu_i \| = \| \phi(x) - \sum_{x \in C_i} \frac{\phi(x)}{||C_i||} \| = \phi(x)\phi(x) - F(x,C) + G(C) \]

In order to minimize criter function, following function is used [4-6].

For every Ci cluster find|C i|G(C i). For every X and Ci cluster calculate F(X,C i). Calculate \( \| \phi(x) - \mu_i \| \) equality and assigned \( \mu_i \) point to the nearest center. For j=1 to K update \( \mu_j \) using \( \mu_j = \sum_{x \in C_i} \phi(x) \) formula.

Repeat Step 1 and Step 4 until the results of clusters are close to each other.

And the results are given as \( \Pi_D = \{C_1, C_2, ..., C_k\} \).

**DBSCAN Clustering Algorithm**

DBSCAN clustering algorithm is density based algorithm. It first developed by Martin Ester and colleagues (33). DBSCAN algorithm is used to find the clusters with arbitrary shape and noisy data. With other clustering methods, it is very difficult and laborious to examine
arbitrary shapes. Also the cost of calculation is the highest [4,7-9].

In following figures it is seen that DBSCAN gives very accurate decisions about clustering [4,7-9] (Figures 1 and 2).

DBSCAN clustering algorithm is only used one input parameter and with this parameter is identifies the clusters that have large spatial data sets by looking local density in the database [4,7-9]. In addition this method is decide what information is outlier or noise data.

For example suppose there are databases like in Figure 3, is the points really create a cluster or of is it small size cluster that consists a lot of outliers. At this point DBSCAN DBSCAN clustering algorithm is given decisions correctly in those consequences [4,7-9].

Before proceeding working method of DBSCAN clustering method, several concepts is required to define in the following [4,7-9].

**ε(Eps) Neighborhood**

It is a neighborhood of a object within ε radius.

Eps is a proximity distance to determined the neighboring of a object. In other words, it is the maximum distance between any two points in a cluster.

\[ N_{εp}(q) = \{q \in D | \text{dist}(p,q) \leq Eps \} \]

**MinPts**

If a point is intensive, then it should be have minimum number of points in the neighborhood ε .

**Core point**

If a object includes at least MinPts observation in the neighborhood ε , then this a core point.

\[ N_{εp}(q) \geq \text{MinPts} \]

**Density reachable**

When there is a chain like \( p_1, p_2, \ldots, p_n \), the equalities are \( p_i = q \) and \( p_n = p \). Under the conditions about Eps and MinPts, the density reachable is provided via \( p_i, r, p \) points [4,7-9].

This is summarized as follows. Suppose there is \( p \) and \( q \) points. In this cases, the \( p \) point is directly density reachable to \( q \) point. But \( q \) point is not directly reachable to \( p \) point. Because \( q \) point is a core point (we assumed that MinPts is three) [4,7-9].

**Directly density reachable**

Where the following conditions are provided the \( p \) point is directly density reachable from \( q \) point [4,7-9].

\[ p \in N_{εp}(q), N_{εp}(q) \geq \text{MinPts} \]

**Density connected**

Under the conditions Eps and MinPts, a \( p \) point is density connected with the help of \( q \) point and this \( q \) point is density reachable both \( p \) and \( q \) points. This situation is illustrated as follows [4,7-9] (Figures 4 and 5).

**Noise**

A set of points that not belong to any cluster in the database.

To summarize above described concepts, the following example is given. Assuming MinPts is three, in Figure 6 the points \( m, p, o, s \) are core point. Because they have three and more than three points in the ε neighborhood. \( q \) point is directly reachable to \( m \) and \( m \) is directly reachable to \( q \) point. The opposite is also true of these provisions. The point \( a \) does not belong to any cluster. So this \( a \) point is a noise point [4,7-9].

In DBSCAN clustering algorithm there is two condition that in order to include \( q \) point in \( K \) cluster. The first condition is, if \( p \) point
is belonging to K cluster, it density reachable by p point. The second condition is p point is density connected with q point. In these two cases p and q points are elements of K cluster [4,7-9].

DBSCAN algorithm randomly selects a p point and under specified conditions about MinPts and Eps it determines all density points may be accessed. If the point of p is a border point or if there is no reachable point, the process is continued with another randomly selected point. This process continues until all points in the cluster are fixed [4,7-9].

DBSCAN clustering algorithm is quite effective for arbitrary according to appropriate parameters and MinPts which are set by user. If a spatial index is used, algorithms computational complexity is $O(n\log n)$. In this formulation n is the observation number in data set. In other cases the computational complexity is $O(n^2)$ [4,7-9].

Data

In our country cardiovascular diseases are the most common ailments in recent years. Cardiovascular term is used as a generic name for the heart and vascular disorders. It leads significant deterioration of the person’s quality of life and causes of death. Therefore, to prevent the increase of cardiovascular problems and to reduce deaths, it is necessary to determine of risk factors and reduce the risk of death from cardiovascular disease by making the necessary interventions. In this study, for determining cardiovascular risk groups Framingham risk score was used. This score was frequently used in recent years. Framingham risk score was calculated on people living in the town of Framingham which is a state of Massachusetts in the United States. The American Heart Association has created the Framingham risk scale using this data. The purpose of creating this table is decreasing death rates by taking the necessary measures to identify individuals at risk of cardiovascular events.

The data used in our study was covered in a 3 year period between 2012-2014 and data were collected over people who wanted to lose weight and had periodic inspection in Duzce University Family Medicine Polyclinic. 4788 individual measurements were collected. Also Framingham risk scores were calculated by using a written macro and the individuals were categorized as low, intermediate and high risk group. For Framingham score, individuals which have less than 10% risk is belonging to low risk, with the risk between 10% and 20% intermediate risk, higher than 20% ise found high risk group. Evaluation of the first stage, descriptive statistics were given as mean and standard deviation for numerical variables, for categorical variables statistics were given as frequency and percentage. The variables age, gender, smoking, systolic blood pressure, total cholesterol, HDL and using blood pressure medication were used for calculating the risk groups with Framingham risk score.Kernel K-means and DBSCAN clustering algorithms were used as clustering algorithms. For determining which

| Table 1. The Distribution of Categorical Variables. |
|---|---|---|
| Gender | n | % |
| Female | 3998 | 83.5 |
| Male | 790 | 16.5 |
| Smoking Status | n | % |
| No Smoking | 3287 | 68.7 |
| Smoking | 1501 | 31.3 |
| Using Blood Pressure Drug Status | n | % |
| No | 4209 | 87.9 |
| Yes | 579 | 12.1 |
| Framingham Risk Score | n | % |
| Low Risk | 4017 | 83.9 |
| Intermediate Risk | 568 | 11.9 |
| High Risk | 203 | 4.2 |

| Table 2. Descriptive Statistics For Numerical Variables. |
|---|---|---|---|---|
| | Mean | Standard Deviation | Minimum | Maximum |
| Age (year) | 37.343 | 12.1024 | 18 | 80 |
| Systolic Blood Pressure (mm-Hg) | 124.849 | 17.5585 | 80 | 250 |
| Total Cholesterol (mg/dl) | 192.692 | 39.9267 | 52 | 396 |
| HDL (mg/dl) | 51.102 | 12.4480 | 21 | 108 |

| Table 3. Distribution of Individuals Per Cluster. |
|---|---|---|---|
| Cluster Methods | Clusters | n | % |
| Kernel K-ortalama | Cluster 1 (Low risk) | 1569 | 32.8 |
| | Cluster 2 (Intermediate risk) | 2279 | 47.6 |
| | Cluster 3 (High risk) | 940 | 19.6 |
| DBSCAN | Cluster 1 (Low risk) | 2907 | 65.9 |
| | Cluster 2 (Intermediate risk) | 1091 | 24.8 |
| | Cluster 3 (High risk) | 410 | 9.3 |
were very large. When evaluating clusters, DBSCAN clustering method did not use 380 people that extreme point.

Table 4 shows the results of comprasion of the average age, systolic blood pressure, total cholesterol and HDL between Kernel K-Means and DBSCAN clustering groups and p value is shown. Among the clusters, which are obtained with the application two methods, there were significant differences in terms of the numerical values. For Kernel K-Means and DBSCAN methods, high risk’s group age, total cholesterol was significantly higher than other clustering groups. But in DBSCAN method intermediate groups Systolic Blood Pressure was significantly smaller than low and high risk groups. When results were examined by HDL, Kernel K-Means and DBSCAN results were very different. In Kernel K-Means algorithm group with high risk’s HDL average was significantly higher than other clustering groups. But in DBSCAN clustering method group with low risk had significantly higher average HDL. It showed us that two methods results were not same.

Categorical variables, which were used to calculating Framingham risk score, and cluster results that obtained Kernel K-Means and DBSCAN relationships comprasion results were given in Table 5. Considering Table 5 about Kernel K-Means method there were significant differences in terms of gender, smoking status, Using Blood Pressure Drug Status. But in DBSCAN clustering method Using Blood Pressure Drug Status is not a risk factor for cardiovascular disease.

The adaptation of clustering DBSCAN and Kernel K-Means clustering methods calculating Framingham risk score.

Adaptions of each of the clustering results were investigated with Kappa statistics by the risk factors about Framingham risk score and the results were shown collectively in Table 6. According to Table 6, there was significant harmony among Kernel K-Means and DBSCAN clustering algorithms. But the value of Kappa statistics were quite small.

Framingham risk score groups and clusters that obtained from the two clustering algorithm relationships were given in Table 7. When considering Table 7, it was seen that DBSCAN clustering algorithm’s results were not significantly agreement with Framingham risk score.

But Kernel K-Means clustering algorithm results were significantly agreement with Framingham score.

Discussion and conclusion

Clustering algorithms are used in a variety of workspace. For example making text clustering on social networks like Facebook and Twitter, grouping the students according to social skills at school, the classification of a variety of diseases or symptoms of disease, developing disease groups by combined laboratory and clinical findings, identifying of epidemics, regional analysis, identifying the different morphology of the heart sound and classification of physiological status. At the same time, clustering algorithms were preferred to used for the detection of homogeneous groups before many statistics test as a pre-analysis [10-14].

Based on types of variables and presence or absence of outlier/noise observations in the data set, a variety of clustering algorithms have been developed. In particular, the number of clustering algorithms has been increased in the last 10 years. This result is an indication of increased need for clustering [10-22].

Framingham risk score, which we used in our study, has been used many international and national studies for predicting cardiovascular risk score. This score was used above 1200 articles for determining the turning point of cardiovascular disease. However, when calculating the score, limited number of risk factors is located in the model. This factors includes age, gender, total cholesterol, systolic blood pressure, smoking status and drug using status. Framingham risk score is a mathematical model that predicts the risk of cardiovascular disease over 10 years. However there are disadvantages of this method. It only predicts the cardiovascular disease risk not the other heart illness. And this model is only applied for individuals that have no heart disease before. Another disadvantage of this score is, in cohort study the young people are less represented. Therefore, it intensively emphasis on age. Framingham risk score informations have been developed from people who are living in Framigham area. However this region contains more white race and homogeneous groups. Because of this Framingham risk score may not be accurate for different ethnic and cultural structures. Framingham risk score is only able to develop a 10 year period predictive model. The model included 10 year period

Table 4. Comprasion of Numerical Variables.

| Variables            | Clusters                          | Kernel K-Means | P     | DBSCAN  | P     |
|----------------------|-----------------------------------|----------------|-------|---------|-------|
|                      |                                   | n  | Mean       | Standard Deviation |      | n  | Mean       | Standard Deviation |       |
| Age                  | Cluster 1 (Low risk)              | 1569|31.1428a  | 10.02949|<0.001|2907|37.0602a  | 12.55664|<0.001 |
|                      | Cluster 2 (Intermediate risk)     | 2279|38.2769a  | 11.51889|<0.001|1091|37.5820a  | 10.73071|<0.001 |
|                      | Cluster 3 (High risk)             | 940 |45.4319a  | 11.7966 |<0.001|410 |39.9512a  | 12.33983|<0.001 |
| Systolic Blood Pressure| Cluster 1 (Low risk)              | 1569|118.5641a | 14.59613|<0.001|2907|124.8070a | 18.22597|<0.001 |
|                      | Cluster 2 (Intermediate risk)     | 2279|126.7990a | 17.84390|<0.001|1091|123.3492a | 16.48207|<0.001 |
|                      | Cluster 3 (High risk)             | 940 |130.6128a | 18.30376|<0.001|410 |128.4268a | 17.83573|<0.001 |
| Total Cholesterol    | Cluster 1 (Low risk)              | 1569|151.6811a | 18.47103|<0.001|2907|190.4830a | 39.43385|<0.001 |
|                      | Cluster 2 (Intermediate risk)     | 2279|196.3353a | 13.85578|<0.001|1091|195.5186a | 41.24597|<0.001 |
|                      | Cluster 3 (High risk)             | 940 |252.3128a | 26.27701|<0.001|410 |199.1854a | 40.17462|<0.001 |
| HDL                  | Cluster 1 (Low risk)              | 1569|49.4989a  | 11.72261|<0.001|2907|52.9461a  | 12.09382|<0.001 |
|                      | Cluster 2 (Intermediate risk)     | 2279|51.3510a  | 12.44777|<0.001|1091|51.3844a  | 12.90315|<0.001 |
|                      | Cluster 3 (High risk)             | 940 |53.1755a  | 13.26032|<0.001|410 |43.5927a  | 10.70874|<0.001 |
Table 5. Comparison of Categorical Variables.

|                  | Kernel K-means |       |       |       |
|------------------|----------------|-------|-------|-------|
|                  | Low Risk       | Intermediate Risk | High Risk | P     |
| Gender           |                |       |       |       |
| Female           | 1348           | 1896  | 754   | 0.001 |
| Row%             | 33.7           | 47.4  | 18.9  |       |
| Column%          | 85.9           | 83.2  | 80.2  |       |
| Male             | 221            | 383   | 186   |       |
| Row%             | 28.0           | 48.5  | 23.5  |       |
| Column%          | 14.1           | 16.8  | 19.8  |       |
| Smoking Status   |                |       |       |       |
| Non Smoking      | 1143           | 1531  | 613   | <0.001|
| Row%             | 34.8           | 46.6  | 18.6  |       |
| Column%          | 72.8           | 67.2  | 65.2  |       |
| Smoking          | 426            | 748   | 327   |       |
| Row%             | 28.4           | 49.8  | 21.8  |       |
| Column%          | 27.2           | 32.8  | 34.8  |       |
| Using Blood Pressure Drug Status |                |       |       |       |
| No               | 1488           | 1995  | 726   | <0.001|
| Row%             | 35.4           | 47.4  | 17.2  |       |
| Column%          | 94.8           | 87.5  | 77.2  |       |
| Yes              | 81             | 284   | 214   |       |
| Row%             | 14.0           | 49.1  | 37.0  |       |
| Column%          | 5.2            | 12.5  | 22.8  |       |
| DBSCAN           |                |       |       |       |
| Gender           |                |       |       |       |
| Female           | 2907           | 1091  | 0     | <0.001|
| Row%             | 72.7           | 27.3  | .0    |       |
| Column%          | 100.0          | 100.0 | .0    |       |
| Male             | 0              | 0     | 410   | <0.001|
| Row%             | .0             | .0    | 100.0 |       |
| Column%          | .0             | .0    | 100.0 |       |
| Smoking Status   |                |       |       |       |
| Non Smoking      | 2907           | 0     | 0     | <0.001|
| Row%             | 100.0          | .0    | .0    |       |
| Column%          | 100.0          | .0    | .0    |       |
| Smoking          | 0              | 1091  | 410   |       |
| Row%             | .0             | 72.7  | 27.3  |       |
| Column%          | .0             | 100.0 | 100.0 |       |
| Using Blood Pressure Drug Status |                |       |       |       |
| No               | 2533           | 962   | 357   | 0.663 |
| Row%             | 65.8           | 25.0  | 9.3   |       |
| Column%          | 87.1           | 88.2  | 87.1  |       |
| Yes              | 374            | 129   | 53    |       |
| Row%             | 67.3           | 23.2  | 9.5   |       |
| Column%          | 12.9           | 11.8  | 12.9  |       |

In this direction, we investigated the risk factors of cardiovascular disease that caused the most of deaths among all disease. By using the risk factors we clustered people into three groups including low risk, intermediate risk and high risk. Then we compared the clusters that risk factors effects may be stronger depending on the time. But this score is neglect this effect. For this reasons Framingham risk score should not be used as a gold standard test for the estimating risk score of cardiovascular disease in our country. However, it will be benefited for a preliminary assessment for studies [23].
Table 6. The Relationships Between DBSCAN and Kernel K-Means Results.

| Kappa=0.039 | Kernel K-Means | \( p \) |
|-------------|----------------|-------|
|             | Low Risk       | Intermediate Risk | High Risk |
| DBSCAN      | Count          | 1023  | 1135 | 529   |
|             | Row%           | 35.2  | 46.6 | 18.2  |
|             | Column%        | 27.0  | 64.4 | 61.8  |
| Intermediate Risk | Count | 325   | 541  | 225   |
|             | Row%           | 29.8  | 49.6 | 20.6  |
|             | Column%        | 22.4  | 25.7 | 26.3  |
| High Risk   | Count          | 101   | 207  | 102   |
|             | Row%           | 24.6  | 50.5 | 24.9  |
|             | Column%        | 7.0   | 9.8  | 11.9  |

Table 7. Framingham Risk Score’s Kappa Statistic About DBSCAN and Kernel K-Means.

| Framingham Risk Score | DBSCAN | Kernel K-Means |
|-----------------------|--------|----------------|
| KAPPA                 | -0.001 | 0.06           |
| \( p \)               | 0.964  | <0.001         |

obtained from Framingham risk score and DBSCAN, Kernel K-means results.

Kernel K-Means clustering algorithm has been observed more consistent results compare with DBSCAN method due to the small amount of categorical variables and large size of own data set. However, Kernel K-means method gives correct results when applied clustering process with accurate parameters, despite the lack of a convenient method of clustering large data sets. At the same time, Kernel K-Means method produce faster results used for data set. DBSCAN clustering algorithm is more successful to find low-risk individuals findings while Kernel K-Means clustering method has a tendency who are at high risk.

In this research, high risk of cardiovascular disease has more valuable to find clinically so upscale ratio is desired in cross table. Although, for some studies, they can be approved success such as finding a low-risk individuals. In this case, the results should be analyzed this assessment extent. In these data sets, researchers make mistake while thinking of this algorithm gives definitive results. For these reasons, clustering algorithms should be reviewed whole clinical informations evaluated with criteria, assumptions, algorithm conditions, disadvantages and advantages. Statistical methods should be supported that the clinical findings therefore the application stage continues easily and correctly in backstage [10-14].

As a result we propose to increase the clustering algorithms applications in the health field. If the correct method is used in the development of health policies, it is necessary to taking the necessary precautions for determined the risk of disease. Thus it will lead to improved quality of life for our people and the average life expectancy will increase. Therefore, even a simple clustering algorithm perhaps will cause a lot of different developments in the field of health and medicine [10-14].

References
1. Köktürk F, Ankarali H, Sümbüloğlu V (2009) Veri madenciliği yöntemlerine genel bakış. Türkiye Klinikleri Journal of Biostatistics 1: 20-5.
2. Fraley C, Raftery AE (2005) How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal 41: 578-588.
3. Ferligoj A, Batagelj V (1983) Some types of clustering with relational constraint. Psychometrika 48: 541-552.
4. AL-Zand Har (2013) Bölümleyici Kümelerle Algorıtmaların Farklı Veri Yeğenliklerinde Karsılıstırılması. 2013, Gazi üniversitesi, Bilşim Ekitimi, Master thesis, 107 Page, Ankara.
5. Samir TH, Viswanath P, Reddy BE (2013) Single pass kernel k-means clustering method. J dhan 38: 407-419.
6. Zhang R, Rudnicky A (2012) A large scale clustering scheme for kernel k-means. ICPR 289-292.
7. Bäcklund H, Hedblom A, Neijman N (2015) DBSCAN: A density-based spatial clustering of application with noise. Linköpings Universitet.
8. Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings 2nd International Conference on Knowledge Discovery and Data Mining. Oregon 226-231.
9. Han J, Kamber M (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc. USA, 418-419.
10. Fraley C, Raftery AE (2005) How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal 41: 578-588.
11. Syed AA (2004) Performance Analysis of K-means Algorithm and Kohonen Networks, Florida Atlantic University, Master of Science, Master thesis,127 Page, America, (Prof. Dr. Abhijit Pandya).
12. Han J, Kamber M (2006) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc. USA, 5-10.
13. Koh HC, Tan G (2005) Data mining applications in healthcare. J Healthc Inf Manag 19: 64-72. [Crossref]
14. Ferligoj A, Batagelj V (1983) Some types of clustering with relational constraint. Psychometrika 48: 541-552.
15. Godara S, Yadav R (2013) Performance analysis of clustering algorithms for character recognition using Weka tool. International Journal of Advanced of Computer and Mathematical Sciences 4: 119-123.
16. Singh B, Dubey G (2013) A comparative analysis of different data mining using WEKA. International Journal of Innovative Research & Studies 2: 380-391.
17. Revathi S, Nalini T (2013) Performance comparison of various clustering algorithm. International Journal of Advanced Research in Computer Science and Software Engineering 3: 66-72.
18. Balabantaray RC, Samra C, Jha M (2015) Document clustering using K-means and K-medoids. International Journal of Knowledge Based Computer Systems 1.
19. Kakkar P, Parashar A (2014) Comparison of different clustering algorithms using WEKA tool. International Journal of Advanced Research in Technology, Engineering and Science 1: 20-2.
20. Jung YG, Kang MS1, Heo J2 (2014) Clustering performance comparison using K-means and expectation maximization algorithms. Biotechnol Biotechnol Equip 28: S44-S458. [Crossref]
21. Goyal VK (2014) An experimental analysis of clustering algorithms in data mining using Weka tool. International Journal of Innovative Research in Science & Engineering 2: 171-6.
22. Ramya D, Rao D (2014)Performans evaluation of learning by example techniques over different datasets. International Journal of Innovative Research in Science, Engineering and Technology 3: 15960-5.
23. Hemann BA, Bimson WF, Taylor AJ (2007) The Framingham Risk Score: an appraisal of its benefits and limitations. Am Heart Hosp J 5: 91-96. [Crossref]