Cancer Diagnosis Using K-Means Clustering

Aye Chan Mon
University of Computer, Monywa, Myanmar
ayechanmon25@gmail.com

Abstract

The proliferation, ubiquity and increasing power of computer technology has aided data collection, processing, management and storage. However, the captured data needs to be converted into information and knowledge to become useful. Data mining is the process of using computing power to apply methodologies, including new techniques for knowledge discovery, to data. Data mining identifies trends within data that go beyond simple data analysis. Through the use of sophisticated algorithms, non-statistician users have the opportunity to identify key attributes of processes and target opportunities. This paper intends to support these non-statistician users in analysis of the cancer diagnosis by implementing the k-means clustering algorithm. In this paper, the Blood Cancer diagnosis is analysis in speciality.

1. Introduction

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown. This is also the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. A primary reason for using data mining is to assist in the analysis of collections of observations of behavior. Such data are vulnerable to collinearity because of unknown interrelations. An unavoidable fact of data mining is that the (sub-) set(s) of data being analyzed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships and behaviors that exist across other parts of the domain. To address this sort of issue, the analysis may be augmented using experiment-based and other approaches. In these situations, inherent correlations can be either controlled for, or removed altogether, during the construction of the experimental design.

Data mining techniques can be classified into the following categories: classification, clustering, association rules, sequential patterns, time-series patterns, link analysis and text mining.

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values describing the objects. Often, distance measures are used. For example, Euclidian distances are used in order to specify the similarities and dissimilarities. This is an important step in any clustering which is to select a distance measure which will determine how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. For example, in a 2-dimensional space, the distance between the point (x=1, y=0) and the origin (x=0, y=0) is always 1 according to the usual norms, but the distance between the point (x=1, y=1) and the origin can be 2, or 1 if you take respectively the 1-norm, 2-norm or infinity-norm distance.

2. Related work

Clustering problems arise in many different applications, such as data mining and knowledge discovery [1], data compression and vector quantization [2], and pattern recognition and pattern classification [3]. The notion of what constitutes a good cluster depends on the application and there are many methods for finding clusters subject to various criteria, both ad hoc and systematic. These include approaches based on splitting and merging such as ISODATA [4,5], randomized approaches such as CLARA [6], CLARANS [7], methods based on neural nets [8], and methods designed to scale to large databases, including DBSCAN [9], BIRCH [10], and ScaleKM [10]. For further information on clustering and clustering algorithms, see [6], [11], [5].Clustering based on k-means is closely related to a number of other clustering and location problems. These include the Euclidean k-medians (or the multisource Weber problem) [3], [12] in which the objective is to minimize the sum of distances to the nearest center and the geometric k-center problem [1] in which the objective is to minimize the maximum distance from every point to its closest center. There are no efficient solutions known to any of these problems and some formulations are NP-hard [13]. An asymptotically efficient approximation for the k-means clustering problem has been presented by Matousek [14], but the large constant factors suggest that it is not a good candidate for practical implementation.

One of the most popular heuristics for solving the k-means problem is based on a simple iterative scheme for finding a locally minimal solution. This algorithm is often called the k-means algorithm [15]. There are a number of variants to this algorithm, so, to clarify which version we are using, we will refer to it as Lloyd's algorithm.

3. Cluster analysis

Clustering is a challenging field of research where its potential applications pose their own special requirements. This process divides a large dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another and different groups are as "far" as possible from one another, where distance is measured with respect to all available variables. Cluster analysis has been widely used in numerous applications including pattern recognition, data analysis, image processing and market research.

Interval-scaled variables are continuous measurements of a roughly linear scale. The measurement unite used can affect the clustering analysis. The dissimilarity (or similarity) between the objects described by interval-scaled variables is typically computed based on the distance between each pair of objects. The most popular distance measure is Euclidean distance, which is defined as

$$d(i,j) = \sqrt{\,|h_{i1} - x_{j1}{}^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2}$$

where i = (x_{i1}, x_{i2}, …, x_{ip}) and j =(x_{j1}, x_{j2}, …, x_{jp}) are two p-dimensional data objects. Euclidean distances satisfy the following mathematic requirements of a distance function:

(1) d (i ,j) >= 0: Distance is a nonnegative number

(2) d (i, i) = 0: The distance of an object to itself is 0.

(3) d (i, j) = d (j, i): Distance is a symmetric function.

(4) d (i, j) <= d (i, h) + d (h, j): Going directly from object i to object j in space is no more than making a detour over any other object h (triangular inequality).

Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k <= n. That is, it classifies the data into k groups, which together satisfy the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group. To achieve global optimality in partitioning-based clustering would require the exhaustive enumeration of all of the possible partitions. Instead, most applications adopt one of two popular heuristic methods: (1) the k-means algorithm, where each cluster is represented by the mean value of the objects in the cluster, and (2) the k-medoids algorithm, where each cluster is represented by one of the objects located near the center of the cluster.

K-means defines a prototype in terms of a centroid, which is usually the mean of points and is typically applied to objects in a continuous n-dimensional space. So, this technique clustering technique is called centroid-based technique. The k-means algorithm takes input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low.

Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster’s center of gravity.

The k-means algorithm proceeds as follows. First, it randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the squared-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} |p - m_i|^2$$

where E is the sum of square-error for all objects in the database, p is the point in space representing a given object, and m_i is the mean of cluster C_i (both p and m_i are multidimensional). This criterion tries to make the resulting k clusters as compact and as separate as possible.

The algorithm attempts to determine k partitions that minimize the squared-error function. It works well when the clusters are compact clouds that are rather well separated from one another. The method is relatively scalable and efficient in processing large data sets. [13]

The k-means procedure is summarized in below.

Algorithm: k-means. The k-means algorithm for partitioning based on the mean value of the objects in the cluster.

Input: The number of clusters k and a database containing n objects.

Output: A set of k clusters that minimizes the square-error criterion.

Method:

(1)Arbitrarily choose k-objects as the initial cluster centers

(2)Repeat

(3)(Re) assign each object to the cluster to which the objects is the most similar, based on the mean value of the objects in the cluster;

(4)Update the cluster means, i.e., calculate the mean value of the objects for each cluster;

(5)Until no change;

This system represents 2 for benign and 4 for malignant by analysis with k-means clustering.

4. Choosing initial centroids

Choosing the proper initial centroids is the key step of the basic k-means procedure. It is easy and efficient to choose initial centroids randomly, but the results are often poor. It is possible to perform multiple runs, each with a different set of randomly chosen initial centroids – one study advocates 30 - but this may still not work depending on the data set and the number of clusters sought. There isn't any problem as long as two initial centroids fall anywhere in a pair of clusters, since the centroids will redistribute themselves, one to each cluster, and so achieve a globally minimal error,. However, it is very probable that one pair of clusters will have only one initial centroid. In that case, because the pairs of clusters are far apart, the k-means algorithm will not redistribute the centroids between pairs of clusters, and thus, only local minima will be achieved.

This imposes the need for a technique to choose a new centroid for an empty cluster, for otherwise the squared error will certainly be larger than it would need to be. A common approach is to choose as the new centroid the point that is farthest away from any current center. If nothing else, this eliminates the point that currently contributes the most to the squared error. Updating points incrementally may introduce an order dependency problem, which can be ameliorated by randomizing the order in which the points are processed. However, this will not be really feasible unless the points are in main memory.

Updating the centroids after all points are assigned to clusters results in order independence approach. Finally, note that when centers are updated incrementally, each step of the process may require updating two centroids if a point switches clusters. However, k-means tends to converge rather quickly and so the number of points switching clusters will tend to be small after a few passes over all the points.

5. Case study

We used the cancer data set from UCI laboratory for this research by using k-means clustering methods. This data contains details of various cancer types. With this data we can try things like predicting the class of cancer from its attributes. All of 10 attributes are numeric value.

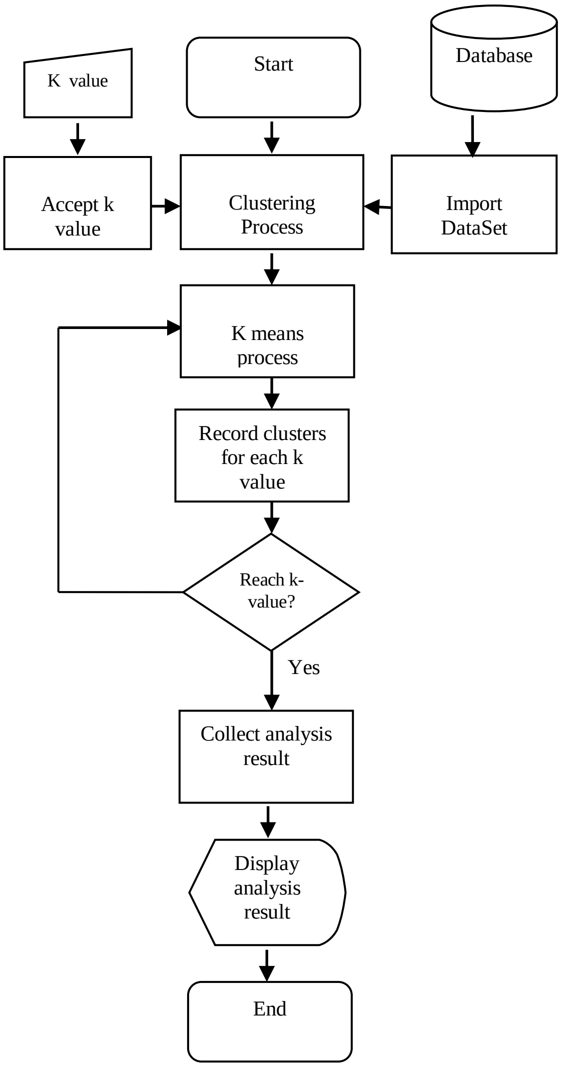
#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

Class distribution:

Benign: 458 (65.5%)

Malignant: 241 (34.5%)

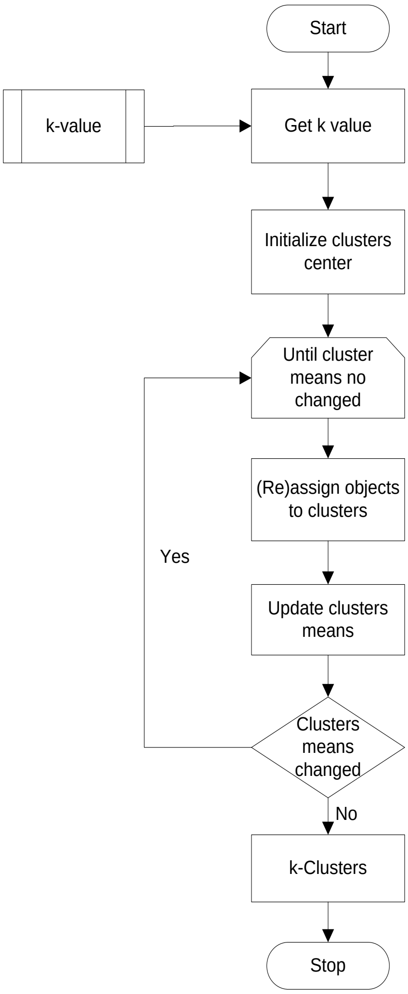
This system consists of 10 attributers. They are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses and Class. Each attributes has 1 to 10 id number but class attribute has two main pasts-2 for benign and 4 for malignant. In class distribution, benign has 458(65.5%) and 241(34.5%) in malignant of 699 ,total instances.



“**Figure 1. System flow of the k-means clustering**”

Import Dataset into the system. And then, check and modify Dataset. Before starting the analysis process, accept maximum k value from manual. By using three partitioning method, the user can view the analysis result of clustering as original import dataset.

Accept k value from the user. Define initialize clusters center. The algorithm can process until cluster means no changed. Reassign objects to clusters based on the distance between the objects and the clusters means. All the points are assigned, recalculate the cluster’s means

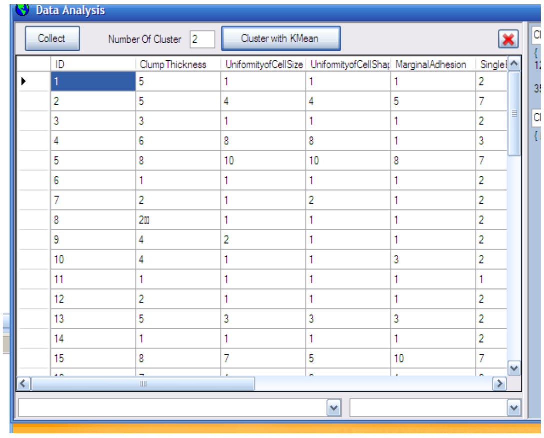

Start
k-value
Get k value
Initialize clusters center
Until cluster means no changed
(Re)assign objects to clusters
Yes
Update clusters means
Clusters means changed
No
k-Clusters
Stop

“Figure 2. K-means process flow chart”

6. Result

The figure 3 presents the analysis of k-means clustering. Firstly, we import data from database. After that we input k value, k=2 for the case study: cancer diagnosis. When we click the cluster with k-means button, this system will be presented k groups. We can see the two groups of records in figure 4 .

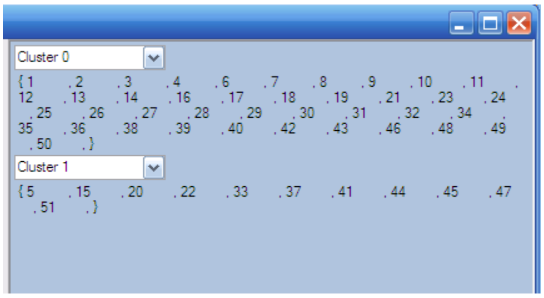
The experimental results demonstrated that the overall accuracy is over 80% from each implementation. There is no changes in accuracy whenever the user apply the various data set for this system.



Number Of Cluster 2

ID	ClumpThickness	UniformityofCellSize	UniformityofCellShap	MarginalAdhesion	SingleE
1	5	1	1	1	2
2	5	4	4	5	7
3	3	1	1	1	2
4	6	8	8	1	3
5	8	10	10	8	7
6	1	1	1	1	2
7	2	1	2	1	2
8	2	1	1	1	2
9	4	2	1	1	2
10	4	1	1	3	2
11	1	1	1	1	1
12	2	1	1	1	2
13	5	3	3	3	2
14	1	1	1	1	2
15	8	7	5	10	7

“Figure 3. Data analysis of case syudy”



```
Cluster 0
{1     ,2     ,3     ,4     ,6     ,7     ,8     ,9     ,10    ,11    ,
12    ,13    ,14    ,16    ,17    ,18    ,19    ,21    ,23    ,24
 ,25    ,26    ,27    ,28    ,29    ,30    ,31    ,32    ,34    ,
35    ,36    ,38    ,39    ,40    ,42    ,43    ,46    ,48    ,49
 ,50    ,}
Cluster 1
{5     ,15    ,20    ,22    ,33    ,37    ,41    ,44    ,45    ,47
 ,51    ,}
```

“Figure 4. The result of case syudy”

7. Conclusion

Clustering analysis is a useful (and interesting) field. Many people use cluster analysis for a wide variety of useful tasks.

By the theory, k-means method is sensitive to outliers since an object with an extremely large value may substantially distort the distribution of data.

By the experimental results, the system is calculated this method best base upon using various data sets. We conclude that the formation of clustering is essentially depends on the choosing of initial centroids.

8. References

[1] C. P. Wei, Y. H. Lee and C. M. Hsu,“*Empirical Comparison of Fast Clustering Algorithms for Large Data Sets*”, National Sun Yat-Sen University. 3rd Hawaii International Conference on System Sciences-2000.

[2] J. Han & M. Kamber “*Data Mining Concepts and Techniques*”, ISBN 1-55860-489-8, Morgan Kaufmann Publishers.

[3] M. Ester, H. P. Kriegel, X. Xu “A Database Interface for Clustering in Large Spatial Databases”, Institute for Computer Science, University of Munich. 1st International Conference on knowledge Discovery and Data Mining (KDD-95) pp. 94-99, 1995.

[4] M.Stiefel, R. J. Oberg “*Application Development Using C# and .NET*" ISBN 0-13-093383-X Prentice HallPRT.

[5] P. N. Tan, M. Steinbach, V. Kuma, “*Introduction to Data Mining*”.

[6] Z. Huang “*A fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining*”, CSIRO Mathematical and Information Sciences, GPO Box 664, Canberra2601, AUSTRALIA.

[7] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, “Advances in Knowledge Discovery and Data Mining”, AAAI/MIT Press, 1996.

[8] A. Gersho and R.M. Gray, “Vector Quantization and Signal Compression", Boston: Kluwer Academic, 1992.

[9] R.O. Duda and P.E. Hart, “Pattern Classification and Scene Analysis”, New York: John Wiley & Sons, 1973.

[10] G.H. Ball and D.J. Hall, “Some Fundamental Concepts and Synthesis Procedures for Pattern Recognition Preprocessors”, Proc. Int'l Conf. Microwaves, Circuit Theory, and Information Theory, Sept. 1964.

[11] A. K. Jain and R.C. “*Dubes, Algorithms for Clustering Data. Englewood Cliffs*”, N.J.: Prentice Hall, 1988.

[12] L. Kaufman and P.J. Rousseeuw, “*Finding Groups in Data: An Introduction to Cluster Analysis*”, New York: John Wiley & Sons, 1990.

[13] R.T. Ng and J. Han, “*Efficient and Effective Clustering Methods for Spatial Data Mining, Proc. 20th Int'l Conf. Very Large Databases*”, pp. 144-155, Sept. 1994.

[14] T. Kohonen, “*Self-Organization and Associative Memory*”, third ed. New York: Springer-Verlag, 1989.

[15] T. Zhang, R. Ramakrishnan, and M. Livny, BIRCH: A New Data Clustering Algorithm and Its Applications, Data Mining and Knowledge Discovery, *vol.* 1, no. 2, pp. 141-182, 1997.