International Journal of Advanced Culture Technology Vol.5 No.3 40-45 (2017)
https://doi.org /10.17703/IJACT.2017.5.3.40

IJACT 17-3-5

Recommendation of Optimal Treatment Method for Heart Disease using EM Clustering Technique

Yong Gyu Jung[1], Hee Wan Kim[2]*

[1]Dept. of Medical IT, Eulji University, Korea
[2]Div. of Computer-Mechatronics, Sahmyook University, Seoul, Korea
e-mail: ygjung@eulji.ac.kr, hwkim@syu.ac.kr

Abstract

This data mining technique was used to extract useful information from percutaneous coronary intervention data obtained from the US public data homepage. The experiment was performed by extracting data on the area, frequency of operation, and the number of deaths. It led us to finding of meaningful correlations, patterns, and trends using various algorithms, pattern techniques, and statistical techniques. In this paper, information is obtained through efficient decision tree and cluster analysis in predicting the incidence of percutaneous coronary intervention and mortality. In the cluster analysis, EM algorithm was used to evaluate the suitability of the algorithm for each situation based on performance tests and verification of results. In the cluster analysis, the experimental data were classified using the EM algorithm, and we evaluated which models are more effective in comparing functions. Using data mining technique, it was identified which areas had effective treatment techniques and which areas were vulnerable, and we can predict the frequency and mortality of percutaneous coronary intervention for heart disease.

Keywords: *Clustering Analysis, EM Algorithm, Percutaneous Coronary Intervention, Heart Disease, Frequency and Mortality*

1. Introduction

With the advancement of computer science, the emergence of the Internet and the web, and the development of mobile computing, such as symptoms, tests, and prescriptions and treatments done on patients, is being automatically recorded in real time without manual intervention. For example, every time a diabetic patient measures his blood glucose level with his tester, the record is sent wirelessly to the hospital, and if it is determined that the patient is at risk, the patient may be asked to come to the hospital. Furthermore, due to the development of bioinformatics, which has attracted a great deal of attention in recent years, it has become possible to collect, analyze and mining human genetic information together with patient data.

In this paper, we analyzed the relationship between the frequency of percutaneous coronary intervention for patients with heart disease and the mortality resulted from that operation in New York area using this

Manuscript received: 17 July, 2017 / Revised: 28 July, 2017 / Accepted: 16 August, 2017
Corresponding Author: hwkim@syu.ac.kr
Tel: +02-3399-1783 Fax: +02-3399-1791
Dept. of Computer Engineering, Shamyook University, Korea

data mining technique. Cluster analysis were applied to enhance the accuracy of predicted values. EM algorithms used in cluster analysis, were compared and evaluated.

2. Related research

2.1 Clustering Algorithm

Cluster analysis is a process of grouping data into a class called a cluster so that objects in the same family have a high similarity to each other and have a high degree of similarity with objects of other clusters. Differentiability is assigned a value based on attribute values representing objects, usually distance measures are used. These cluster analysis is widely used in data mining, statistics, biology, and machine learning fields, and a wide variety of cluster algorithms have been proposed. The criteria for selecting the clustering algorithm depends on the data used and the purpose of the application. When learning clusters, the output is in the form of a diagram showing how instances belong to the cluster. The simplest case involves associating each instance with the number of clusters. The simplest case involves, for example, associating each instance with the number of clusters. In the simplest case, it involves associating each instance with the number of clusters. This is to place the instances in a two-dimensional space and visually divide each cluster into spaces. Some clustering algorithms allow one instance to belong to clusters, so the diagram is drawn by superimposing subsets representing each cluster, placing instances in two-dimensional space. Some algorithms associate instances with clusters stochastically rather than categorically. In this case, there is one probability or instance of membership in each instance that belongs to each cluster.
This special association is intended to be a probabilistic result, so the sum of the probability values for each example is 1. Other algorithms can be used to create clusters with a hierarchical structure so that the instance space from the top-level clusters can be divided into only a few clusters, each clustering pointing to its own subcluster each time it goes down. In this case, the elements combined at the lower level are clustered more closely than at the higher level.

2.2 EM Algorithm

Cluster analysis is a process of grouping data into a class called a cluster so that objects in the same family have a high similarity to each other and have a high degree of similarity with objects of other clusters. Differentiability is assigned a value based on attribute values representing objects, usually distance measures are used. These cluster analysis is widely used in data mining, statistics, biology, and machine learning fields, and a wide variety of cluster algorithms have been proposed. The criteria for selecting the clustering algorithm depends on the data used and the purpose of the application. When learning clusters, the output is in the form of a diagram showing how instances belong to the cluster. The simplest case involves associating each instance with the number of clusters. The simplest case involves, for example, associating each instance with the number of clusters. In the simplest case, it involves associating each instance with the number of clusters. This is to place the instances in a two-dimensional space and visually divide each cluster into spaces. Some clustering algorithms allow one instance to belong to clusters, so the diagram is drawn by superimposing subsets representing each cluster, placing instances in two-dimensional space. Some algorithms associate instances with clusters stochastically rather than categorically. In this case, there is one probability or instance of membership in each instance that belongs to each cluster. This special association is intended to be a probabilistic result, so the sum of the probability values for each example is 1. Other algorithms can be used to create clusters with a hierarchical structure so that the instance space from the top-level clusters can be divided into only a few clusters, each clustering pointing to its own sub cluster each time it goes down. In this case, the elements combined at the lower level are clustered more closely than at the higher level.

The EM algorithm, on the other hand, offers more solutions than any detailed algorithm. The core of the EM algorithm is the concept of incomplete data. Incomplete data can be generated generally by omissions due to failure to fully record observations or omission from a theoretical point of view. Therefore, we

compensate for missing data in the E (Expectation) phase of the EM algorithm. When the data is sufficiently supplemented and restored to complete data, the parameter is estimated in the M (Maximization) step.

The M step can be analytically solved for the M step because it maximizes to a simpler function than the possible derivative of the observed value. The process of inserting a value for this simple function of step M is called iteration of the EM algorithm. If the estimated parameter of the reconstructed data is not equal to the parameter of the maximum likelihood of this data, this reconstructed data is not complete data. The dominance of the EM algorithm is that it is numerically stable. Therefore, it is possible to strongly avoid that the EM algorithm is less than the maximum probability or is overestimated. When calculating the EM algorithm, it is necessary to modify the parameter estimation formula slightly so as not to consider the known cluster itself for each instance but to consider the cluster probability.

The EM algorithm is not so simple. That is, the algorithm only converges near the fixed point and does not reach the exact fixed point. Given the five parameter values, we can see how well the data converges by calculating the overall likelihood that the data will be extracted from this data set. This overall likelihood value is obtained by multiplying the individual instance. The overall likelihood value is a measure of the 'appropriateness' of the cluster, and the value of the EM algorithm increases with each iteration. The EM algorithm, as well as the ideal dominance, can carefully constrain the parameters. The constraint can be solved by M steps. On the contrary, the competition of the maximization method must be well combined with the special technique of the constraint of the parameter.

3. Experiment

3.1 Experimental data

Percutaneous coronary intervention has been shown to have a positive prognostic effect in extra-cardiac surgery, leading to increased pre-eclampsia in high-risk patients with previous myocardial infarction. However, percutaneous coronary intervention (PCI) has a high risk of perioperative in-stent thrombosis in the spinal cord, but there is a high risk of hemorrhage due to platelet inhibitors used to prevent thrombosis. To assess the regional prognosis for this high risk procedure, we used Cardiac Surgery and Percutaneous Coronary Interventions data related to percutaneous coronary intervention for patients with heart disease in a hospital from the US public data website.

In this paper, we aim to find the correlation between the frequency of percutaneous coronary intervention and the mortality rate by region. Therefore, we extract the detailed region, number of cases, and number of death among 14 data values and find the association through the algorithm. The attributes selected in the experimental data are as follows.

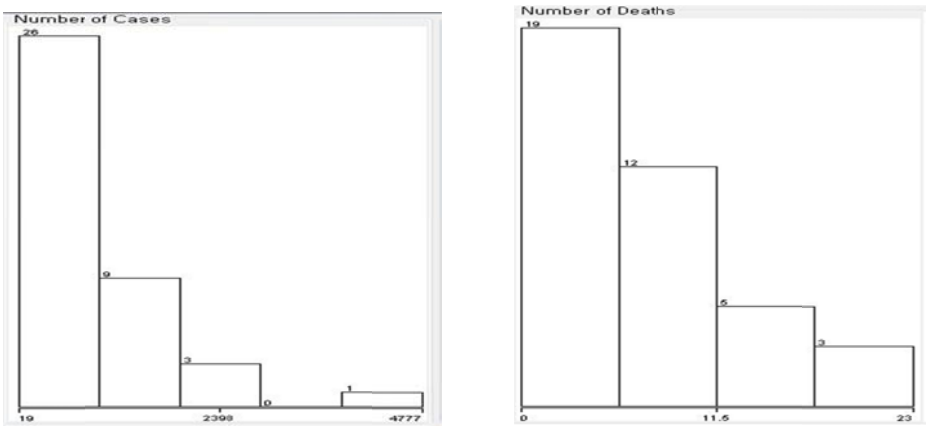

Figure 1. Number of Cases and Death Property Distribution

The minimum number of cases is 19, the maximum value is 4777, the minimum value of Number of Deaths is 0, and the maximum value is 23. The minimum value of Number of Cases is 19, the maximum value is 4777, the minimum value of Number of Deaths is 0, the maximum value is 23, and the algorithm of C4.5 and EM algorithm are executed. The minimum number of cases is 19, the maximum value is 4777, the minimum value of Number of Deaths is 0, the maximum value is 23, and C4.5 and EM algorithm is tested.

4. Experimental results and discussion

Experiments were carried out by setting the Detailed Region attribute as a dependent variable and the number of cases and number of Death as independent variables based on the data. The values were analyzed using Use training set. Figure 2 shows the EM algorithm analysis.

Attribute	Cluster 0 (0.39)	1 (0.11)	2 (0.5)
Detailed Region			
Capital District	2.1285	1.0149	4.8565
Western NY - Rochester	1.0011	2	1.9989
Manhattan	3.8181	4.2036	1.9783
Bronx	2.9935	1.0149	1.9916
Kings	4.0518	1.0033	1.9449
Western NY - Buffalo	3.2964	1.0043	2.6993
Central NY	1.0074	1	3.9926
Queens	2.8016	1.1546	3.0438
NY Metro - New Rochelle	2.0186	1.0009	1.9805
NY Metro - Long Island	2.0742	1.0059	4.9199
[total]	25.1914	14.4022	29.4064
Number of Cases			
mean	1107.992	2925.4164	373.7798
std. dev.	491.0933	1101.931	202.5003
Number of Deaths			
mean	9.8551	18.9628	3.3911
std. dev.	2.8483	4.1201	1.9187

Time taken to build model (full training data) : 0.19 seconds

Figure 2. EM algorithm analysis table

Figure 2 shows that the cluster analysis is divided into three. The dependent variable, Detailed Region, is divided into three groups, each of which corresponds to the number of independent variables, number of cases, and number of death. In the first cluster 0, 25.1914 of the total data were classified as cluster 0, representing 39%. Cluster 1 was classified into 14.4002 cluster 1 of the total data, which is equivalent to 11%. In the last cluster 2, 29.4064 of the total data were classified as cluster 2, representing 50%. EM algorithm cluster result Cluster 2 occupied the largest part, followed by cluster 0, and cluster 1. Cluster 2 is characterized by a small number of cases and a small number of deaths. Cluster 0 features many number of cases and number of deaths. The feature of cluster 1 is that the ratio of number of cases, number of death is a group having a middle value with respect to cluster 2 and 0. Figure 3 shows the visualization of the cluster analysis described above

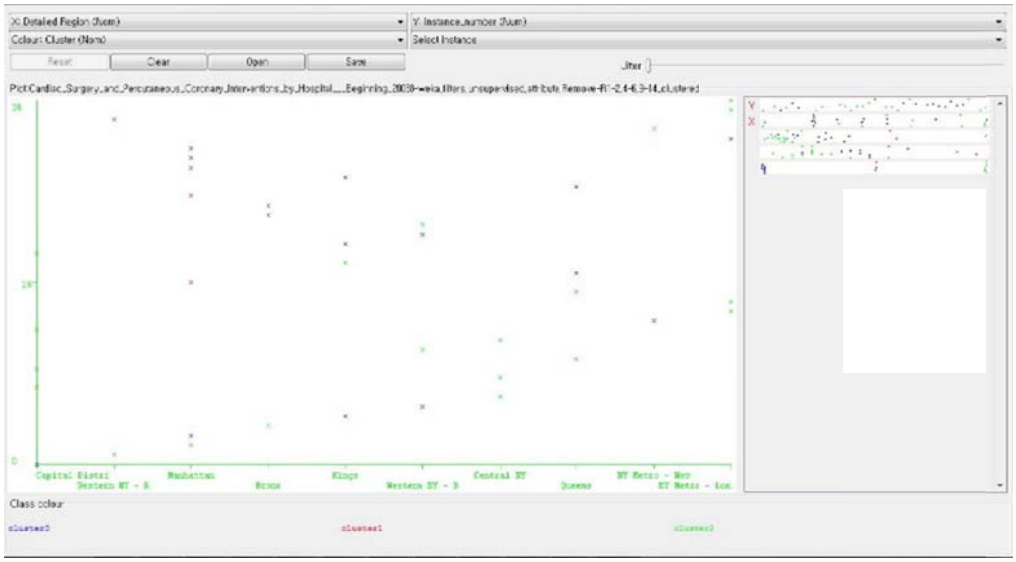

Figure 3. EM algorithm visualization

Figure 4 shows that the number of cases in Manhattan is significantly higher, of which cluster 1 is high. On the contrary, the number of cases of central NY is much lower, and the proportion of cluster 2 is high. The area with a high proportion of cluster 0 was the Kings area. Figure 5 is a visualization of EM algorithm analysis of the number of Death and Detailed Region.

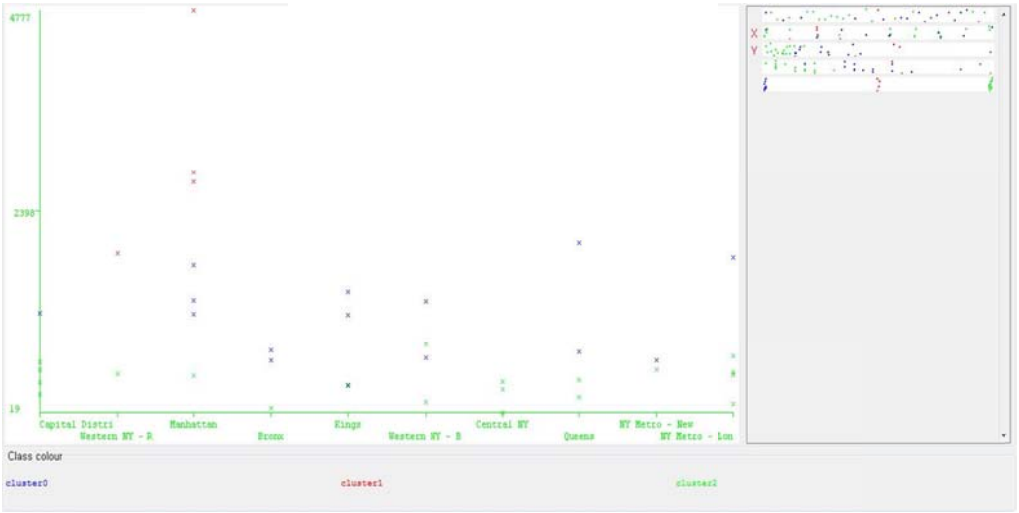

Figure 4. EM algorithm analysis visualization of number of cases and Detailed Region

The region with the highest number of Death was Manhattan area. The number was same as number of cases. Among them, cluster 2 occupied a large portion of the population. Western NY-Rochester was the region with the largest number of deaths in the same area. The area with the lowest number of deaths was central NY. The number of deaths in the cluster 2 was the largest.

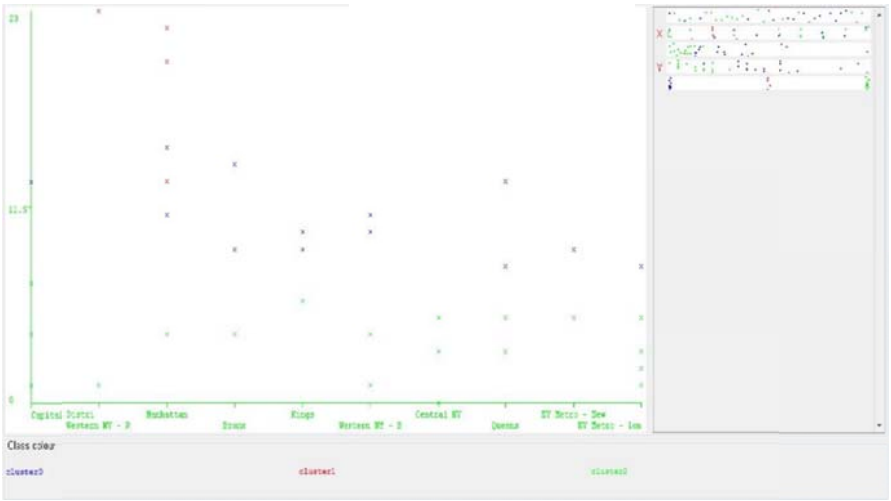

Figure 5. EM algorithm analysis visualization of number of Death and Detailed Region

5. Conclusion

Analysis of the results of this study showed that central NY and NY metro-Long Island had the lowest mortality rate compared to frequency, through decision tree and cluster analysis, and showed excellent survival rate of percutaneous coronary intervention. Manhattan has the highest mortality compared to other regions, indicating the lowest survival rate. This will lead to a better quality of medical treatment by analyzing the trend of medical treatment in regions where there is a high level of survival rate.

References

[1] Brindle, Joanne T., et al. "Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics." Nature medicine 8.12 (2002): 1439.
[2] Pepine, Carl J., et al. "A calcium antagonist vs a non–calcium antagonist hypertension treatment strategy for patients with coronary artery disease: the International Verapamil-Trandolapril Study (INVEST): a randomized controlled trial." JAMA 290.21 (2003): 2805-2816.
[3] Young-Moon Lee, Jun-Koo Kwag and Young-Seob Hwang, A Feature Analysis of Industrial Accidents Using C4.5 Algorithm, Journal of the KOSOS, Vol.20 No.4, pp.130-137, 2005.
[4] Risk of coronary artery disease patients undergoing extra-cardiac surgery, http://blog.naver.com/didvk4444?Redirect=Log&logNo=90105192330
[5] Yancy, Clyde W., et al. "Race and the response to adrenergic blockade with carvedilol in patients with chronic heart failure." New England Journal of Medicine 344.18 (2001): 1358-1365.
[6] Stukel, Thérèse A., et al. "Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods." JAMA 297.3 (2007): 278-285.