International Journal of Advanced Smart ConvergenceVol.5 No2 53- 58(2016)
http://dx.doi.org/10.7236/IJASC.2016.5.2.53

IJASC 16-1-2

Similarity Analysis of Hospitalization using Crowding Distance

Yong Gyu Jung[*] , Young Jin Choi[**], Byeong Heon Cha[***]

[*]*Dept. of Medical IT Marketing, Eulji University, Korea*
[**] *Dept. of Health Management, Eulji University, Korea*
[***]*Dept. of Biomedical Laboratory Science, Eulji University, Korea*
e-mail: {ygjung, yuzin, jobogy}@eulji.ac.kr

Abstract

With the growing use of big data and data mining, it serves to understand how such techniques can be used to understand various relationships in the healthcare field. This study uses hierarchical methods of data analysis to explore similarities in hospitalization across several New York state counties. The study utilized methods of measuring crowding distance of data for age-specific hospitalization period. Crowding distance is defined as the longest distance, or least similarity, between urban cities. It is expected that the city of Clinton have the greatest distance, while Albany the other cities are closer because they are connected by the shortest distance to each step. Similarities were stronger across hospital stays categorized by age. Hierarchical clustering can be applied to predict the similarity of data across the 10 cities of hospitalization with the measurement of crowding distance.

In order to enhance the performance of hierarchical clustering, comparison can be made across congestion distance when crowding distance is applied first through the application of converting text to an attribute vector. Measurements of similarity between two objects are dependent on the measurement method used in clustering but is distinguished from the similarity of the distance; where the smaller the distance value the more similar two things are to one other. By applying this specific technique, it is found that the distance between crowding is reduced consistently in relationship to similarity between the data increases to enhance the performance of the experiments through the application of special techniques. Furthermore, through the similarity by city hospitalization period, when the construction of hospital wards in cities, by referring to results of experiments, or predict possible will land to the extent of the size of the hospital facilities hospital stay is expected to be useful in efficiently managing the patient in a similar area.

Keywords: *big data, clustering, hierarchical analysis, attribute vector*

1. Introduction

In recent years, industry of the market to take advantage of the open and big data of government 3.0 becomes huge, the importance of big data has emerged. Big Data for the diverse and large-scale data, it can be utilized as an important resource that affects the superiority of future competitive. In Fig. 1, a variety of big data industry of the market is over 2011 in 2017, shows how the size or growth to some extent. The work of extracting the information of value to have these large big data is called data mining, using data mining, in addition to medical, agricultural, distribution, economic forecasts in various industrial fields such as has been used for the purpose of extraction of valuable information.

Manuscript Received: Mar. 16, 2016 / Revised: Mar. 25, 2016 / Accepted: April 15, 2016
+ Corresponding Author: jabogy@eulji.ac.kr
Tel:031-740-7223
Dept. of Biomedical Laboratory Science, Eulji University, Korea

It will be examined the data mining of the medical field. In the medical field, data mining using the relevant rules can be found the relationship between the data by modeling. The potential customers to provide an indication of the objective medical to prevent diseases caused future and contribute to health of the people. A medical institution, it is possible to increase customer satisfaction by utilizing comprehensive medical system predictable data.

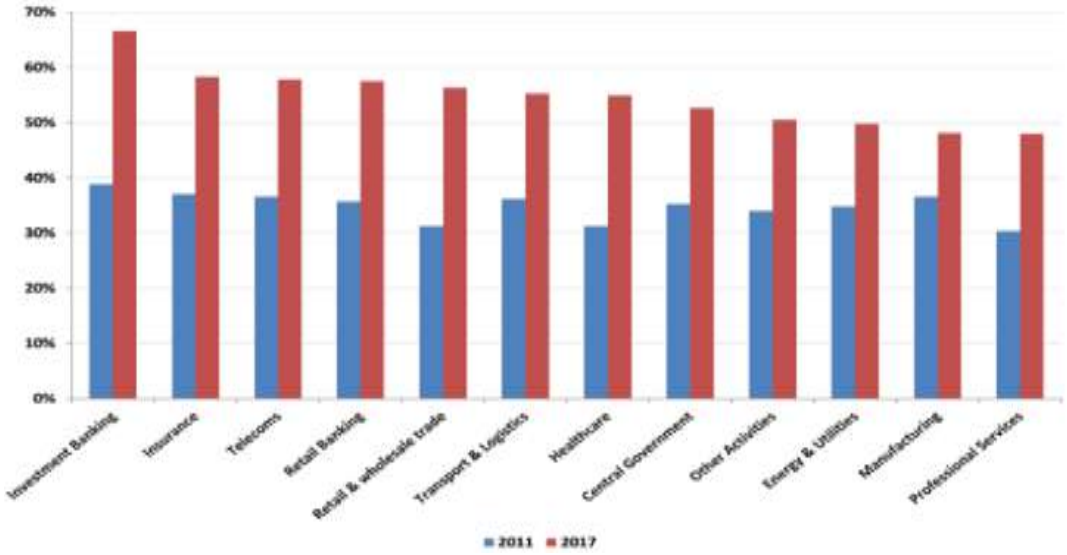

Figure 1. Big Data industry market

In this paper, length of hospital stay in the hospital discharge data of patients, It is tried to analyze whether there is some degree of difference in age and region. The data of discharge and hospitalization in the medical fields, has been utilized as a tool to be able to grasp the health statistics at the national level of health and medical care utilization form. For this purpose, it is necessary to construct a production system for sustained and stable health statistics. Especially as to if development is our society, because its importance has become increasingly large, are the production of more reliable quality is high variety of statistics is required. Mainly been obtained and to clearly understand the phenomenon by using statistics, the useful information through the results detailed analysis based on, by utilizing this, providing a highly efficient policies that This is because it is. In this paper, in the field of data mining, by utilizing the k-means congestion algorithm Clustering techniques to extract the information through statistical analysis and modeling of the data, and converts the text to attribute vector special techniques can be applied in this experiment, it is desired to analyze the hospital patient data.

2. Related Research

2.1 Hierarchical cluster analysis

Cluster Analysis, also called data segmentation, has a variety of goals that relate to grouping or segmenting a collection of objects (i.e., observations, individuals, cases, or data rows) into subsets or clusters, such that those within each cluster are more closely related to one another than objects assigned to different clusters. Central to all of the goals of cluster analysis is the notion of degree of similarity or dissimilarity between the individual objects being clustered. There are two major methods of clustering: hierarchical clustering and k-means clustering. For information on k-means clustering, refer to the k-Means Clustering section. In hierarchical clustering, the data is not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run

from a single cluster containing all objects to n clusters that each contains a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by a series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer groupings. Agglomerative techniques are more commonly used, and this is the method implemented in XLMINER. Hierarchical clustering may be represented by a two-dimensional diagram known as dendrogram, which illustrates the fusions or divisions made at each successive stage of analysis.

2.2 The Distance Formula

Very often, especially when measuring the distance in the plane, we use the formula for the Euclidean distance. According to the Euclidean distance formula, the distance between two points in the plane with coordinates (x, y) and (a, b) is given by

$$\text{dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

The source of this formula is in the Pythagorean theorem. Look at the diagram. In the plane - since the Earth is round, this means within relatively small areas of Earth's surface - it is pretty good, provided the distance is exactly estimated. While moving at a given speed, the Euclidean formula may not be very useful providing the answer. It is often impossible to move from one point straight to another. There are buildings and streets with traffic fences to be accounted.

$$\text{dist}((x, y), (a, b)) = |x - a| + |y - b|$$

It is more useful. In mathematics, the Euclidean distance is most fundamental. As one of the mechanical proofs of the Pythagorean theorem shows, the same is also true in physics, although in either science it's not the only distance formula used.

3. Experiment and Results

Tools that are used for the experiments using the S-link is a statistical package developed in Japan. Experimental data is a Hospital Inpatient Discharges by County of Residence (SPARCS) Beginning2009 that has been collected through the research cooperation system of the US states. This data is data that summarizes the patients discharged on hospital including outpatient surgery or emergency room visit SPARCS. The data is fixed by the residence of the patient. SPARCS data, two different data sets of the equipment and the hospital discharge Residence patients to maintain the confidentiality of personally identifiable information is divided into the hospital discharge plasma.

Table 1. Experimental data attributes

Attribute	Type	Value
Patient County of Residence	nominal	Albany, Allegany, Bronx, Broome, Cattaraugus, Cayuga, ...
Patient Age Group	nominal	1-12, 13-17, 18-40, 41-64, 65-74, 75+
Average Length of Stay	numeric	continuous from 1 to 52

This data set does not include a facility. The attributes of the data to be used in the experiment, Patient County of Residence, Patient Age Group, is composed of such as Average Length of Stay, Patient County of Residence is, Albany, Allegany, Bronx, Broome, Cattaraugu, Cayuga, Chautauqua , Chemung, Chenango. It was selected a total of 10 of the city of Clinton. Of the experimental results of Example data, statistics are shown in Figure 2.

'Similarity Matrix

	1	2	3	4	5
2	5.4863				
3	3.6069	7.8019			
4	4.1593	3.1780	5.9724		
5	3.5581	4.0522	4.5000	3.8756	
6	2.0125	3.8249	4.8724	3.2604	2.6965
7	2.6944	5.1595	3.0282	4.2237	1.5684
8	3.0430	3.0627	5.6312	2.0688	3.4583
9	5.1546	3.5707	7.6981	3.5595	4.9447
10	6.2193	11.3754	4.9950	10.0409	8.0461

	6	7	8	9
7	2.8125			
8	1.9365	3.5972		
9	4.2615	5.4562	2.7911	
10	7.8543	6.7587	9.1652	10.8862

Figure 2. Statistics of data Example

Here, Cattaraugus and Chautauqua cities are connected by a first stage, hospitalization period of two cities it is believed that there is a similarity. In the second step, it can be seen that it is connected by Albany City and the shortest distance to the final stage from the third stage is connected to the Cayuga Chemung. Congestion distance Clinton cities, from being displayed in the longest 4.9950 in the last stage, it can be seen that out of 10 cities is less similarity of hospital stay.

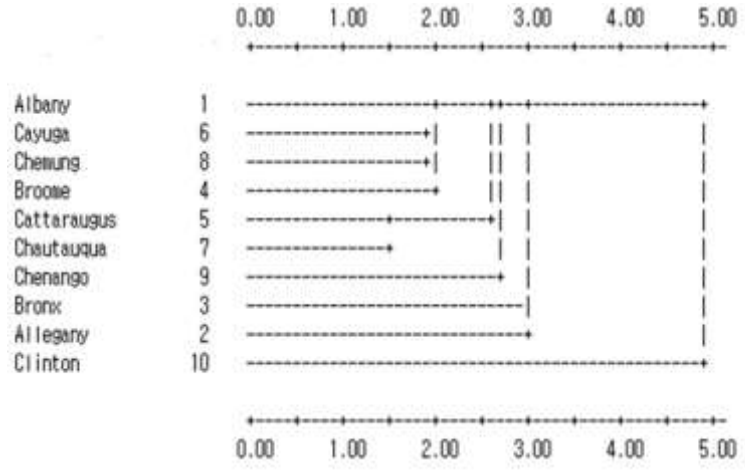

Figure 3. Dendrogram of the data Example

Much crowding distance than the previous data congestion distance Example data employing the special technique is applied it can be expected to close. This is a set of variables, S in the existing six variables, M, and by three conversion H, since easily convert a number of each data, it is expected that such performance is shown in Figure 2 and Figure 3.

Table 2. Distance Changes as per Preprocess

Before Preprocess				After Preprocess			
Step	Cluster1	Cluster2	Distance	Step	Cluster1	Cluster 2	Distance
1	5	7	1.5684	1	5	7	0.5385
2	6	8	1.9365	2	6	8	1.0677
3	1	6	2.0125	3	2	6	1.3153
4	1	4	2.0688	4	2	4	1.5297
5	1	5	2.6944	5	2	5	1.578
6	1	9	2.7911	6	1	2	1.6553
7	1	3	3.0282	7	1	9	1.7
8	1	2	3.0627	8	1	3	1.9875
9	1	10	4.995	9	1	10	3.9013

4. Conclusion

In the experiment of similar through the crowding distance of data of age-specific hospitalization period, crowding distance is the longest, most little similarity to other urban cities, it is expected in the city of Clinton, Albany City the crowding distance of the other cities are close, because they are connected at the shortest distance to each step, It was found to be high similarity of hospital stay according to each age. Was also measured congestion distance to predict similarity of data 10 hospitalization in the city through a hierarchy of Clustering technique in this experiment. In order to enhance the performance of the hierarchical Clustering techniques in the experiment were compared congestion distance after applying the congestion distance before applying by applying a special technique for converting the text to an attribute vector. Similarity measure of the two objects in Clustering, depending on the measurement method, but is distinguished from the similarity of the distance, where as the distance value is smaller, it means that both objects are similar to each other. By applying special technique it is found that the distance between the congestion is reduced reliably, in accordance with similarity between the data increases through this, to enhance the performance of the experiments through the application of special techniques. Furthermore, through the similarity by hospitalization period, when the construction of hospital wards in cities, by referring to results of experiments, or predict possible will land to the extent of the size of the hospital facilities hospital stay is expected to be useful in efficiently managing the patient in a similar area. Further, in this experiment, we expect that the various data mining techniques using a number of special techniques to improve the performance of the experiment.

References

[1] Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition, Morgan Kaufmann Publishers, 2011

[2] Jun-ho Lim, medical data mining using association rules , School of Computer & Information Technology Korea University, 2010

[3] Disease Control Division, Korea Research Society, the study of specimens correction and weight calculation of discharge patient survey, 2007. 12

[4] Disease Control Division, Korea Research Society, Hospital patient survey sampling and weighting correction calculation study 2007. 12

[5] Ltifi, Hela, et al. "A human-centred design approach for developing dynamic decision support system based on knowledge discovery in databases." Journal of Decision Systems 22.2 (2013): 69-96.

[6] Barnes, Sean, Bruce Golden, and Stuart Price. "Applications of agent-based modeling and simulation to healthcare operations management." Handbook of Healthcare Operations Management. Springer New York, 2013. 45-74.