International Journal of Advanced Smart Convergence Vol.2 No.2 40-42 (2013)
DOI 10.7236/IJASC2013.2.2.9

IJASC 13-2-9

Category Variable Selection Method for Efficient Clustering

Jun Heo†, Chae Yun Kim††, Yong-Gyu Jung†††

† Dept. of Information and Communication, Kyungmin University, Korea
heojun@kyungmin.ac.kr
† †,††† Dept. of Medical IT Marketing, Eulji University, Korea
Ksh_124@naver.com, ygjung@eulji.ac.kr (Corresponding Author)

Abstract

Recent medical industry is an aging society and the application of national health insurance, with state-of-the-art research and development, including the pharmaceutical market is greatly increased. The nation's health care industry through new support expansion and improve the quality of life for the research and development will be needed. In addition, systemic administration of basic medical supplies , or drugs are needed , the drug at the same time managing how systematic analysis of pharmaceutical ingredients , based on data through the purchase of new medicines and pharmaceutical ingredients automatically classified by analyzing the statistics of drug purchases and the future a system that can predict a patient is needed. In this study, the drugs to the patient according to the component analysis and predictions for future research techniques, k-means clustering and k-NN (Nearest Neighbor) Comparative studies through experiments using the techniques employ a more efficient method to study how to proceed . In this study, the effects of the drugs according to the respective components in time according to the number of pieces in accordance with the patient by analyzing the statistics by predicting future patient better medical industry can be built.

Key words : Drug component analysis, clustering, k-means, k-NN.

1. INTRODUCTION

Reliable production of pharmaceutical products for quality performance reporting system to RFID FDA drug reporting system and systematic way was through the network between each distribution subject to the efficient production management system was built. In addition, both drugs RFID barcode reader attached to the development of low-cost supply for the recognition of pharmacy insurance claims were S / W for sale in connection with the preparation work to streamline the development and dissemination of the system was conducted . In addition, by taking advantage of cloud computing information, genome assembly individuals with disease diagnostic capabilities and bioinformatics DB guchok TOOL drug candidates through the development of high-speed , low-cost development and improve public health and quality of service and efficiency and competitiveness in the field of corporate affairs for the purpose of putting . In this study, the clustering method was applied to the classification of some of the drugs which with the introduction of new medical technology, health services at the center in line with customer care through the patient's drug plan to try the service was provided . With overall and Drug Administration drug through automatic component analysis for the purpose of automating the management of patients with prescription drugs , depending on the patient's predicted future demand forecast and medicine at the same time allowing the technology to study.

2. RELATED RESEARCH

T2.1 K-MEANS Clustering

Clustering techniques to any object or target similarity or distance by objects with similar characteristics are grouped

Manuscript received: Sept. 01, 2013 / revised : Nov. 20, 2013
Corresponding Author: ygjung@euljji.ac.kr
Tel: +82-31-740-7178, Fax: +82-31-740-7195
Department of Medical IT Marketing, Eulji University, Korea

together so that the aggregate number of clusters is an analytical technique. The main purpose of clustering data into each of the clustering attribute of the group is intended to identify. In particular, the classification scheme or the number of clusters, unlike the structure of the group is not assumed, only the similarity between the objects, or by a distance to form clusters, and clusters formed by identifying the characteristics of the relationship among clusters an analysis of this type as evident by this condition is not known or can be utilized in the technique.

In this paper, we use the hierarchical clustering technique is the average connection method is used in clustering technique. Average connection method between all of the objects in the two clusters, the average squared distance between the square of the distance expressed as a cluster. Mainly using dendrogram can be expressed. Average connection method to merge and split ever Clustering is there ever present study uses the merge. Divide and Conquer populations of clustering as an individual object are merged according to their similarities to form a single cluster becomes tied. The new cluster is formed according to the similarity between clusters is gradually reduced and finally all parts of the communities they tied the lead to make one single cluster.

The techniques used in this study of clustering are hierarchical clustering method is most suitable for obtaining information on the number of clusters it is possible to apply the present study, the pharmaceutical composition analysis of drugs with the same or similar components merge with each other in a manner that proceeds. Finally merged to constitute clusters in accordance with the components of pharmaceutical products to the pharmaceutical use according to the components it is possible to grasp.

2.2 k-NN (Nearest Neighbor)

k-NN algorithm based on the characteristics of the observations in the training sample is the closest way to classify observations. Machine learning methods are classified into one of the most simple way to be. k-NN method is based on non-parametric probability density estimation of the probability density function is a nonparametric data number (k) is fixed and a volume (V) is set as a variable x as a function of laying considered. Thus, by Bayes' theorem nonparametric probability density estimation equation is as follows.

$$p(x \mid C(k)) = (1 / V(x)) * (k/N)$$

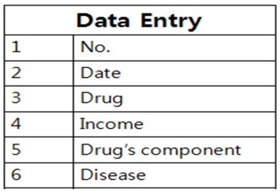
N = number of data
k = the number of fixed data
V = volume

k is 1, the value x for a given data of each class by calculating the distance to the closest data to the distance that the smallest value assigned to the class classification results of the method with the closest value can be extracted. The caveat is the selection of the appropriate value of k is the x-value should be less susceptible to noise.

In this study, the drugs used in the associated data set according to the composition analysis and the efficacy of each product is classified into. First, the training data of each data item value is a String value, int, transform items. Then the data is classified as k-nn algorithm.

3. EXPERIMENT

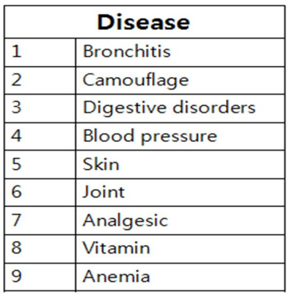
Experimental data of the Korea Pharmaceutical Traders Association and the Korea Health Authority Drug input amount statistical data were used. Through the pre-treatment experiment was to allow easily proceeded, drugs listed in the wearing date of the drug in order to facilitate the experiments were study. Used in the study data and the k-means clustering technique as k-nn Comparative Test Method by the term pharmaceutical composition analysis of a patient's disease over the next specific period of the system to learn in advance the prediction. Nine minutes of this data is in the process of drug dealing and drug supplies and other medical equipment, but with the exception of item validity and cost of drugs after removing the item attribute that you plan to use for research. This data is an address data for use in the pretreatment process proceeds. Drug names by default according to the analysis component according to the disease and the drug entry is added.



Data Entry	
1	No.
2	Date
3	Drug
4	Income
5	Drug's component
6	Disease

Figure 2. Data Entry

Through the pre-processing, 9 items are categorized as the result. It is also discretized as numbers, which are used as primary key separated.



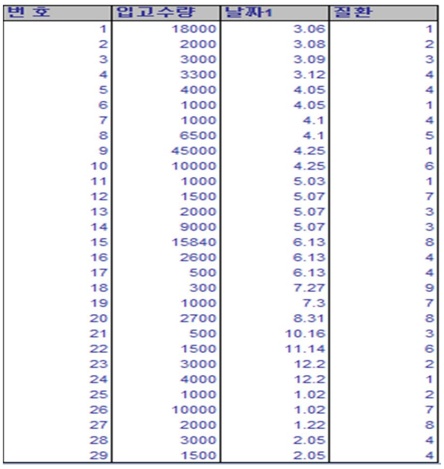
Disease	
1	Bronchitis
2	Camouflage
3	Digestive disorders
4	Blood pressure
5	Skin
6	Joint
7	Analgesic
8	Vitamin
9	Anemia

Figure 3. Items of the disease through the Preprocessing

4. EXPERIMENTAL RESULT

Quantity received and date of input data, the drug, which is the item number of the numerical result. It was carried through the analysis controlled by varying the variable k.

Data analysis proceeds as a result changing the value of the variable optimum category type was attempted. This experiment, the optimal k value is 1, the result was that the unknown. The analysis of this data to proceed with less data when seen as a result of the experiment, this value is also changed according to the change of the data amount can be learned.



번 호	입고수량	날짜1	질환
1	18000	3.06	1
2	2000	3.08	2
3	3000	3.09	3
4	3300	3.12	4
5	4000	4.05	4
6	1000	4.05	1
7	1000	4.1	4
8	6500	4.1	5
9	45000	4.25	1
10	10000	4.25	6
11	1000	5.03	1
12	1500	5.07	7
13	2000	5.07	3
14	9000	5.07	3
15	15840	6.13	8
16	2600	6.13	4
17	500	6.13	4
18	300	7.27	9
19	1000	7.3	7
20	2700	8.31	8
21	500	10.16	3
22	1500	11.14	6
23	3000	12.2	2
24	4000	12.2	1
25	1000	1.02	2
26	10000	1.02	7
27	2000	1.22	8
28	3000	2.05	4
29	1500	2.05	4

Figure 4. Analysis of the value of the best k

5. CONCLUSION

The data used in this study, the pharmaceutical composition according to the result of analysis in the future find any disease in a patient comes , any drugs which can be sold a lot of time depending on whether you can predict future patient basic data analysis was proceeding . According to this study, according to the date associated with the patients' symptoms are judged . However, in this study using the patient data through the whole future predictions too little data. Thus the overall data provide a comprehensive and more accurate analysis is needed. In fact , this can be achieved through the study of k-means clustering is a clustering analysis to proceed as a result of a small amount of data , and you can proceed to set the initial value data , but analysis of the k-NN effective only in case of less data to predict future patient for data analysis could not be obtained.

REFERENCES

[1] O'Connor AM, Rostom A, Fiset V, et al. Decision aids for patients facing health treatment or screening decisions: systematic review. BMJ 1999;319:731–734.

[2] Lucas PJ, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artif Intell Med 2004;30(3):201–214. 202 Chapter 11

[3] Garg AX, Adhikari NK, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005;293(10):1223–1238.

[4] MandlKD, Szolovits P,Kohane IS. Public standards and patients' control: howto keep electronic medical records accessible but private. BMJ 2001;322(7281):283–287.

[5] Yong-Gyu Jung, Seung-Ho Lee and Ho Joong Sung, Effective Diagnostic Method Of Breast Cancer Data Using Decision Tree, Journal of IWIT (2010), Vol.10 No. 5 pp.57-62

[6] Yong-Gyu Jung , Jun Heo, Kyu Ho kim, Using Discretization of Numeric Attributes to Compare the Changes in Performance of C4.5 and CART algorithms, International Conference of the Korea Distribution Science Association, ISSN2287-478X Vol.4 pp353-358 2013.7.11,

Jun Heo received his B.S., M.S. degrees and Ph.D. Candidate in computer science and engineering from Sogang University, Seoul, Korea in 2000, 2002, and current respectively. And he received Ph.D. degree in information and control engineering, Kwangwoon University, Seoul, Korea in 2013. In 2013, he joined the faculty of the Department of Information and Communication, Kyungmin College, Uijeongbu, Kyunggi, Korea, as an assistant professor. His research interests include wired and wireless network protocol, military communication network, ubiquitous computing, energy management system and ad hoc network.

Chae-Yun Kim is a student of Department of Medical IT Marketing, Eulji University, Korea. Her interest area is information analysis and in medical and system implementation in hospital business

Yong Gyu Jung received the B.S. in physics Education from Seoul National University in 1981. And then he got the M.S. and ph.D. degree of Computer Science from Yonsei and Kyonggi University in 1994 and 2003 respectively. Since 1999, he has been a Faculty of Medical IT marketing Dept. in Eulji University. His research interests are in the areas of medical information analysis and international standards including e-Business. He has been a leading member of the UN/ECE/CEFACT and ISO/TC154 for international standardiztion