BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
http://ojs.bilpublishing.com/index.php/jcsr

REVIEW

Based On K-means Disease Diagnosis Research

Jiaqi Wu* **Qingda Zhang** **Linlin Zhao**

North China University of Science and Technology, Tangshan, 063210, China

ARTICLE INFO

Article history
Received: 14 January 2020
Accepted: 17 January 2020
Published Online: 31 March 2020

Keywords:
Disease Diagnosis
K-means
ICD

ABSTRACT

For the diagnosis of diseases, modern medicine usually searches for diseases in the disease database to find the type of disease that matches them. The diagnosis of diseases is the first step in treatment. Then the classification of diseases is the basis of disease diagnosis. Disease classification plays an extremely important role in the scientific management of medical records and the development of modern medicine, and is a bridge connecting modern medical science. Therefore, the classification of diseases is very necessary. Based on this, this article establishes a K-means model for disease diagnosis, and combines the internationally unified disease type code ICD statistics table to classify the sample data set into infectious and parasitic diseases, tumors, diabetes and circulatory diseases The training is perfect, and finally the diagnosis classification of the disease is realized.

1. Introduction

In traditional medical diagnosis and treatment, doctors understand the basic situation of the patient in advance for the diagnosis of the disease, including some basic physical data such as the patient's age, past medical history, and symptoms of onset. Then analyze the patient's condition based on previous experience and get the disease type of the patient, and then treat the patient. In this paper, the K-means model is established to improve the existing diagnosis defects, to improve the doctor's accuracy rate of the patient's illness, and to analyze the patient's condition in conjunction with scientific algorithms.

2. Research Status

For the diagnosis and classification of diseases, some scholars in modern academic research have realized its importance and started a series of studies. Xianjing Hu [1] and others used the application and method of electronic medical records in the clinical disease classification system to construct a clinical diagnosis comparison table, and designed various coding systems based on existing clinical research. Yunsi Cen [2] built a digital medical management database by comparing the medical knowledge object classification systems at home and abroad, rebuilding the medical knowledge object center, digitizing the disease type and determining its value range. Jiayi Li [3] and others analyzed the main nursing problems of 41 inpatients with chronic obstructive pulmonary disease, found out the regularity of the patient's onset symptoms, and built an Omaha classification system to promote the development of clinical nursing. Xiangju Ouyang [4] and others proposed the construction of a national-level disease classification system assisted by the Internet through the rapid development of the Internet, using computers to retrieve disease codes. Juanjuan Cheng [5], collected and analyzed the data of 586 patients undergoing full digital mammography in outpatient clinics. Professional physicians classified breast

**Corresponding Author:*
Jiaqi Wu,
North China University of Science and Technology, Tangshan, 063210, China;
Email: 1643360071@qq.com

DOI: https://doi.org/10.30564/jcsr.v2i1.1650

lesions and proposed the establishment of a BI-RADS FFDM classification system. Accuracy and sensitivity.

3. Building the Model

K-Means Clustering is a method often used to automatically divide a data set into K groups. It belongs to an unsupervised learning algorithm. It originally originated from a vector processing method of signal processing. The sum of the squares of the distances from each point to its corresponding cluster centroid is the smallest.Given a set of observations $(x_1,x_2,...,x_n)$, Each of these observations is a d-dimensional real number vector, and K-means clustering aims to divide n observations into k （k≤n） set $S=\{S_1,S_2,...,S_k\}$ To minimize the sum of squares within the cluster,Where μi is the average of the points in Si, which guarantees that the K-Means algorithm converges to a local optimum. K-means algorithm execution steps are as follows:

Step 1：Randomly select the initialized k category centers from the given data set $a_1, a_2,\ldots,a_k$;

Step 2：For each sample x_i, Mark it as distance category center $d(x_i,x_j)$ Recent categories j;

Since the K-means clustering algorithm is not suitable for processing discrete data, when calculating the distance between samples, you can choose one of Euclidean distance, Manhattan distance, or Minkovsky distance as the similarity of the algorithm according to actual needs measure.Sample data is represented as $x_i=(x_{i1},x_{i2},...,x_{id})$, $x_j=(x_{j1},x_{j2},...,x_{jd})$, Are samples x_ix_j The specific values of the corresponding d description attributes。The smaller the distance between the two rocks, the more similar they are, and the smaller the difference is.

The Euclidean distance formula is as follows:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}$$

Manhattan distance is as follows:

$$d(x_i, x_j) = \sum_{k=1}^{d} |(x_{ik} - x_{jk})|$$

Minkowski distance is as follows:

$$d(x_i, x_j) = \sqrt[p]{\sum_{k=1}^{d} |(x_{ik} - x_{jk})|^p} \ (p = 1, 2, \cdots, \infty)$$

Step 3：Update the center point of each category()Is the mean of all samples belonging to this category;

$$a_j = \frac{1}{N(c_j)} \sum_{i \in c_j} x_i$$

Step 4：Repeat steps 2 and 3 until a certain abort condition is reached.

In the K-means clustering algorithm, the Euclidean distance between different points is an index used to measure the similarity between different quantities. In the K-means clustering algorithm, when the distance from a different point to a certain point is closer, the point will be classified into a class only with the point closest to it. The traditional K-means clustering algorithm is first classified into a class based on the centroid, and so on to achieve classification [5]. Before performing the K-means clustering algorithm, it is of great significance to determine the number k of category centers.Common methods for determining k are: Silhouette Coefficient Calinski-Harabasz Index. Calinski-Harabasz Index The relative calculation is relatively simple and the value of k is more practical. Therefore, the Calinski-Harabasz Index was selected as the evaluation standard for this experiment.

Calinski-Harabasz score s is calculated as:

$$s(k) = \frac{tr(B_k)}{tr(W_k)} \frac{m-k}{k-1}$$

Where m is the number of training samples and k is the number of categories; B_k Covariance matrix; W_k Covariance matrix of the data within the category; *tr* Is the trace of the matrix.That is, the smaller the covariance of the data within the category, the better, and the larger the covariance between the categories, the better, so that the Calinski-Harabasz score will be high.

In this paper, the sample data set is obtained by consulting the relevant hospital data and collation, and the data is preprocessed. The discrete values of the sample data set are removed, the missing values of the sample data set are filled, and the data set used for model testing is finally collated. The results are used as clustering attributes. The K-means clustering method is used to cluster the data. The number of clusters ranges from 2 to 24 through continuous loop iterations to obtain the corresponding Calinski-Harabasz score, as shown in the figure

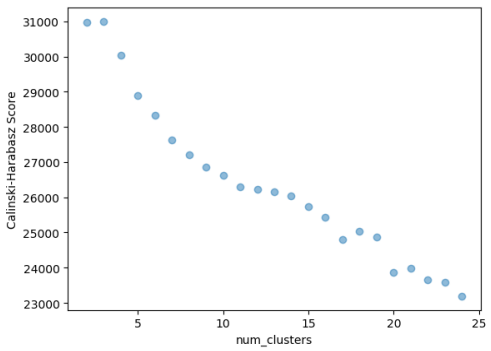

Figure 1. K-value training results

 DOI: https://doi.org/10.30564/jcsr.v2i1.1650

As can be seen from the above figure, when the number of clusters is 3, the Calinski-Harabasz score is the highest and the best clustering effect is obtained. As the number of clusters increased, the Calinski-Harabasz score gradually decreased, and the clusters became worse and worse. When the number of clusters k = 3 is set, K-means clustering is performed on the data.

The International Classification of Diseases (ICD) is a powerful tool for achieving domestic and international health information exchange. It is the basis for grouping medical-related diseases and an important way to extract hospital management information [6]. The diagnosis value in the data set is the ICD code. The main diagnosis value is folded into 10 disease categories according to the international disease classification. The specific disease categories are shown in the table below.

Table 1. ICD-coded disease classification

Category	Diagnostic value range	Disease type	Number	Percentage
0	[1,140]	Infectious and parasitic diseases	2699	2.7%
1	[140,240)	Tumor	3353	3.4%
2	250	diabetes	8568	8.6%
3	[390,460)or785	Diseases of the circulatory system	29753	30.0%
4	[460,520)or786	Breathe	14109	14.1%
5	[520,580)or787	digestion	9297	9.3%
6	[580,630)or788	Urogenital	5026	5.1%
7	[710,740)	Musculoskeletal	4826	4.9%
8	[800,1000)	damage	6815	6.8%
9	V or E or else	other	15045	15.1%

It can be seen from the table that the percentage of infectious and parasitic diseases is 2.7%; the percentage of tumor diseases is 3.4%; the percentage of diabetes is 8.6%; and the percentage of circulatory diseases is 30.0%; Respiratory diseases accounted for 14.1%; digestive diseases accounted for 9.3%; urogenital diseases accounted for 5.1%; musculoskeletal and connective tissue diseases accounted for 4.9%; injuries And poisoning accounted for 6.8%; other types of diseases accounted for 15.1%.

Test and summarize the established K-means algorithm model. The above operation process uses the training set of processing data set. The divided test set data is now input into the model, and the results are shown in the figure below.

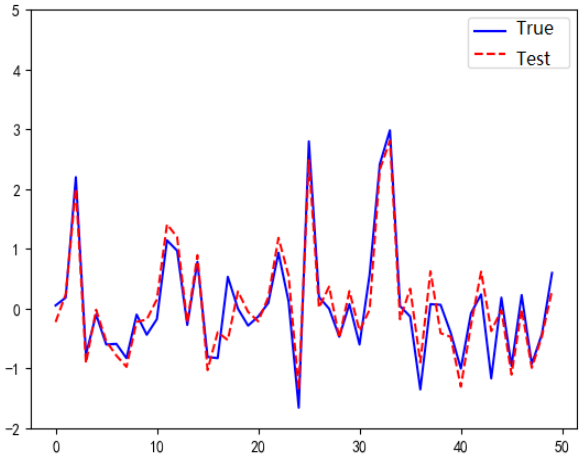

Figure 2. Model training results

As can be seen from the above figure, the K-means algorithm model established in this paper has a good training result on data processing. It clusters the data into three major categories and combines it with actual disease classification problems. Divided into ten categories, and given the range of diagnostic values and the percentage of their number, it can better meet the needs of clinical diagnosis.

4. Model Evaluation

In this paper, the K-means algorithm is used for clustering the diagnosis types of the disease. When constructing the disease grouping model, K-means clustering is performed on the data set and the diseases are grouped by combining the various ICD coding ranges. The advantage is that K-means is a classic algorithm for solving clustering problems. Its operation is simple and the data processing speed is fast. For processing large data sets, the algorithm maintains scalability and efficiency. When the result clusters are very dense, and When the difference between clusters is obvious, the algorithm has better processing effect. However, K-means is very sensitive to noise and outlier data. Even if there is a small amount of such data, it will have a great impact on the average value calculated by the overall model.

5. Conclusion

The study of disease diagnosis classification can not only meet the needs of hospital management, but also can be used for hospital medical, scientific research and teaching purposes. The correctness of disease diagnosis classification directly affects the evaluation of medical quality and the allocation of medical resources. The K-means model established in this paper has a certain guiding role in clinical medical applications, and its algorithm has achieved

DOI: https://doi.org/10.30564/jcsr.v2i1.1650

great research achievements in various fields. Therefore, this model can be used not only for disease diagnosis but also for other unlabeled Clustering problem.

References

[1] Xianjing Hu, Li Zhou, Mingyue Hu. Applied research on clinical disease classification system of electronic medical record[J]. World's latest medical information abstract, 2018, 18 (82): 205 + 207.

[2] Juanjuan Cheng. Application of BI-RADS FFDM classification system in the diagnosis of breast benign and malignant diseases[D]. Huazhong University of science and technology, 2014.

[3] Jiayi Li, Mei Wang, Honglu Duan, Xueqin Liu. Application of Omaha problem classification system in the assessment of inpatients with COPD[J]. Journal of nursing, 2013, 20 (08): 12-15.

[4] Yunsi Cen. Research on medical knowledge modeling for clinical diagnosis and treatment[D]. Tsinghua University, 2012.

[5] Juxiang Ouyang, Feixia Chen, Wenfu Liang, Fang Shen, Dongsheng Liu, Mingjian Hu. The establishment and application of Internet-assisted international disease classification system[J]. Chinese Journal of hospital management, 2006 (07): 485 + 494.

[6] Tiancai Deng, Ning Liao, Fan Zhang, Lifang Man, Fangfang Zhu, junxuan Wang, Lang Li. Quality and efficiency analysis of domestic and international disease classification and coding [J]. Chinese medical record, 2017, 18 (01): 28-32.

 DOI: https://doi.org/10.30564/jcsr.v2i1.1650