A Design of the Group Decision Making Medical Diagnosis Expert System Based on SED-JD Algorithm

Na Zong[1,2], Wuyungerile Li[1,2(✉)], Pengyu Li[1,2], Bing Jia[1,2], and Xuebin Ma[1,2]

[1] Inner Mongolia University, Hohhot 010021, Inner Mongolia, China
gerile@imu.edu.cn

[2] Inner Mongolia A.R. Key Laboratory of Wireless Networking and Mobile Computing, Hohhot 010021, China

Abstract. Medical expert system not only has a lot of medical professional knowledge, but also has inference ability. The inference engine is not only one of the cores of the expert system, but also the key to designing the expert system. We focus on inference engine. In order to improve the diagnostic accuracy of medical diagnostic expert system, we propose the Group Decision Making (GDM) medical diagnosis expert system based on the Standardized Euclidean Distance-Jaccard Distance (SED-JD) algorithm. The mainly research content of inference engine is similarity measurement algorithm (that is SED-JD) and inference engine rule scheme (that is GDM). In order to get more accurate diagnosis, data preprocessing was performed before our experiments. In the design of inference engine, the selection of the Group Decision Making Objects (GDMOs) depends on the maximum similarity distance (MaxDist). The final decision result depends on the average similarity distance of each subgroup. By comparing the similarity scheme and GDM scheme, the experimental results show that GDM scheme is more effective and accurate. By comparing the Standardized Euclidean Distance (SED) algorithm, the Jaccard Distance (JD) algorithm and SED-JD algorithm, the experimental results show that SED-JD algorithm is more accurate.

Keywords: Medical expert system · Group Decision Making · Similarity measurement

1 Introduction

Artificial Intelligence (AI) is a new kind of intelligence which its respond is similar to human intelligence. The expert system is one of the important branches of AI

Supported by the National Natural Science Foundation of China (Grants No. 61761035, 41761086, 61461037, 61661041) and "Scientific and Technological Innovation Project of Inner Mongolia Autonomous Region System Development and Product Application of Urban Flood Disaster Monitoring and Early-warning Management".

J. Li et al. (Eds.): SPNCE 2019, LNICST 284, pp. 421–432, 2019.
https://doi.org/10.1007/978-3-030-21373-2_33

application, and has a important directions that is medical expert system [1]. In this paper, an improved Group Decision Making (GDM) scheme, Subgroup-based Group Decision Making (SGDM), is proposed for inference engine. The expert system and GDM are introduced below.

1.1 Medical Expert System

Expert system is a computer (software) system that can solve difficult and complex problems like human experts [2]. Medical expert system has not only a lot of medical professional knowledge, but also inference ability. Therefore, medical expert system conducts medical diagnosis and medical related inference by simulating the process of analyzing problems in the medical field. Medical expert system, as the assistant of medical system, can lighten the burden of medical workers and make medical work more efficient.

The typical structure of medical expert system [3] is similar to general expert system, including man-machine interface, inference engine, explain module, knowledge base, dynamic database and knowledge base management system. The typical structural diagram of the medical expert system is shown in Fig. 1. Man-machine interface refers to the interaction interface between users and expert system. Inference engine refers to the realization of (generalized) inference procedures. Explain module is responsible for explaining the behavior and results of the expert system to users. Knowledge base refers to the set of knowledge which stored in computers. The knowledge base generally includes expert knowledge, domain knowledge and so on. Dynamic database stores initial evidences, inference results and so on. Knowledge base management system is the supporting software of knowledge base. The relationship between them is similar to the function of database management system on database.

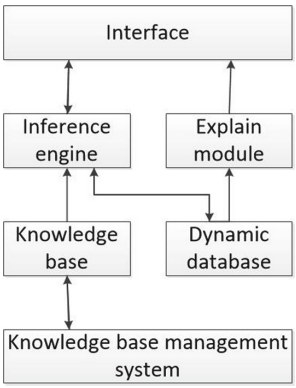

Fig. 1. The Structural diagram of expert system.

In the whole system, the most important modules are the inference engine and the knowledge base. Inference engine is the logical core of expert system.

Knowledge base is the foundation and support of expert system. We focus on the design of the inference engine and the knowledge base.

1.2 Group Decision Making

Science of Decision Making [4] is a comprehensive subject that takes decision making as the research object. That mainly studies the principle, procedure and method of decision making, and explores how to make correct decisions. Group Decision Making (GDM) [5] is a research field with a long history in Science of Decision Making. GDM was proposed by Duncan in 1948 [6]. Hwang gave a clear definition of GDM in 1987 [7] that GDM is decision scheme. Firstly, different members propose their own decision plans. Then form all the plans into a set. Finally form a consistent decision plan based on individual preferences and certain rules. The above is GDM connotation. Meanwhile, the people who participate in decision making constitute the decision making group. The simplest scenario of GDM is election.

For medical diagnostic expert system, it is not enough to diagnose according to similarity measurement only. We can think of each sickness as a group, and the boundaries between groups are not very clear, as shown in Fig. 2. According to the similarity measurement, the input sample is the most similar to a sample in class A and should be classified as class A. But in fact, class A is adjacent to class B. This lead to a miscalculation. In order to reduce or even eliminate sample classification errors on the boundary, GDM is introduced into the medical diagnostic expert system. The majority rule commonly [8] applied in GDM is that the preference of most people is the group preference.

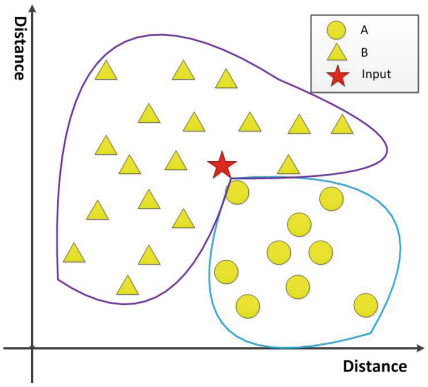

Fig. 2. Similarity judgement diagram.

In Sect. 2, we will introduce the work related to this research. In Sect. 3, we will introduce the system design, and in Sect. 4 implement the design and give the results. Finally, the conclusion and prospect is made in Sect. 5.

2 Related Work

The related work of this paper mainly involves medical expert system, data preprocessing, similarity measurement and GDM.

2.1 Medical Expert System

The inference model of expert system mainly includes prescription production inference model [9–11], fuzzy inference model [12–16] and machine learning inference model [17]. According to the application scenarios, the type of the expert system can be divided into diagnostic [18], explanatory [19], predictive [20], decision-making [21], design [22] and control [23]. And the type of the medical expert system can be divided into diagnostic, explanatory and predictive. According to the classification of output, the type of the expert system is mainly divided into analysis and design. And medical expert system is mainly analytical type. According to the classification of knowledge representation [24], the type of the expert system is mainly divided into production rule representation, predicate logic representation, frame representation, semantic Web representation and so on. Medical expert system is mainly production rule representation type. According to the classification of knowledge, the type of both expert system and medical expert system can be divided into precise inference and imprecise inference [25].

The inference mechanism of medical expert system [26] mainly includes simple production system inference engine, Bayesian theory-based inference engine, MYCIN inference model-based inference engine, fuzzy inference theory-based inference engine, machine learning theory-based inference engine. In this paper, we prefer to a simple production system inference engine.

2.2 Data Preprocessing

Incomplete (with missing values), inconsistent and abnormal raw data will bring obstacles to the research, so data preprocessing operations should be carried out before the research. The operations of data preprocessing mainly include data cleaning, data integration, data transformation and data specification [27].

- Data cleaning is mainly to delete irrelevant and duplicate data, smooth noise data, filter out data irrelevant to mining topics and deal with missing and abnormal values in the original data set.
- Data integration is a process which combines multiple data sources into a consistent data storage. Entity recognition and attribute redundancy should be considered in this process.
- Data transformation mainly carries on the standardization processing to the data. Its methods include the function transformation, the data standardization (normalization), the continuous attribute discretization, the attribute construction, the wavelet transformation and so on.

– Data specification can produce new data set which is smaller than the original data but retains the integrity of the original data. Data specification can reduce the impact of invalid and error data on mining, improve the accuracy, reduce the time of data mining, and reduce the cost of data storage.

2.3 Similarity Measurement

For medical diagnosis expert system, many inference engines are designed based on similarity measurement between medical samples. Similarity measurement [28] is a measurement that comprehensively judges the similarity between two things. The more similar two samples are, the more increased similarity measurement is. There are many kinds of methods for similarity measurement, which are usually selected according to practical problems. Similarity measurements which commonly used include correlation coefficient and similarity coefficient. Correlation coefficient measurements the degree of proximity between variables. Similarity coefficient measurements the degree of proximity between samples. The degree of similarity between samples can be expressed by the following functions.

(1) Similarity coefficient function: the range of similarity coefficient is [0, 1], and the values are positively correlated with similarity. That is to say, the more similar the samples are, the closer the similarity coefficient value is to 1; the more dissimilar the samples are, the closer the similarity coefficient value is to 0.
(2) Distance function: each sample is regarded as a point in n-dimensional space. Distance is used to represent the similarity between samples and is negatively correlated with similarity.

2.4 Group Decision Making

Peng et al. [29] used similarity as the only selection factor for Group Decision Making Object (GDMO). In this scheme, a certain number of GDMO are set firstly. Then GDMOs according to similarity are selected. Finally the minority is subject to the majority.

3 The System Design

This section mainly introduces the main research content of this paper, including data preprocessing, similarity measurement algorithm (that is SED-JD), and GDM scheme (that is SGDM).

3.1 Privacy Data Preprocessing

In this paper, GDM is made based on similarity measurement. Therefore, samples need to be set up in advance for similarity measurement and GDM. Data preprocessing is very important, which is related to the correctness of the whole algorithm. We mainly did the following work.

- Data integration. We merge multiple data sources in this paper. The merge process takes the consultation number as the primary key and merges the data with the same consultation number into one piece of data.
- Data cleaning. We delete private data, irrelevant data, duplicate data, etc.
- Data transformation. We normalize the data, including the standardization of numerical values and the transformation of character values.

3.2 Proposed SED-JD Algorithm

In this paper, similarity measurement between samples is used as the first step of inference. For any sample to be diagnosed, the similarity will be calculated with all samples in the sample set. Due to the sample data of this paper there are both numeric values and character values, in this paper, the similarity measurement will combine the Standardized Euclidean Distance (SED) algorithm and the Jaccard Distance (JD) algorithm, called SED-JD algorithm. SED is used to calculate numeric values, and JD is used to calculate character values. Then combine the two results to get the final similarity.

SED Algorithm. SED is an improvement of the Euclidean Distance (ED). ED is the most commonly used distance calculation formula. Since ED measurements the absolute distance between samples in a multidimensional space, that is, the calculation is based on the absolute value of the characteristics of each dimension, so ED's calculation needs to ensure that all dimensional indicators are at the same scale level. Therefore, the values of each dimension should be standardized before calculation. The standardized formula is formula (1), and the standardized result is expressed by ${x_i}^*$, x_i refers to the ith value of a attribute, $\bar{x}$ is the mean of the attribute, and s is the standard deviation of the attribute.

$$ {x_i}^* = \frac{x_i - \bar{x}}{s} \tag{1} $$

SED's calculation formula is formula (2), where $d(X, Y)$ represents the values between sample X and sample Y calculated by SED, n is the number of attributes, x_i is the ith value of sample X, y_i is the ith value of sample Y, s_i represents the standard deviation of the ith attribute.

$$ d(X, Y) = \sqrt{\sum_{i=1}^{n} (\frac{x_i - y_i}{s_i})^2} \tag{2} $$

Since SED calculates the absolute distance between samples, the larger the value of $d(X, Y)$ is, the smaller the similarity is. When $d(X, Y) = 0$, it means that the two samples coincide exactly.

JD Algorithm. Jaccard Similarity Coefficient (JSC) is a kind of similarity measurement which mainly used for computing symbols or boolean value. Because

the characteristics of the sample attributes are symbols or boolean, so the similarity between samples can only be measured by whether they are the same or not. Therefore, JSC only care about whether the common characteristics between samples are consistent. In short, the proportion of the intersection in the union of set A and set B, which is represented by the symbol $J(A, B)$. The formula is shown in formula (3).

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

In contrast to JSC, the Jaccard Distance (JD) uses the proportion of different elements in two sets to all the elements to measure the divisibility of two sets. It is expressed in symbol $J_\sigma(A, B)$. The formula is shown in formula (4).

$$J_\sigma(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{4}$$

Proposed SED-JD Algorithm. The similarity measurement used in this paper is represented by the symbol $D(X, Y)$, and the formula is shown in formula (5). And $d(X_E, Y_E)$ is the similarity value between sample X and sample Y which calculates by SED. $J_\sigma(X_J, Y_J)$ is the similarity value between sample X and sample Y which compare by JD. JD collect all character attributes as a set. Therefore, we treat $J_\sigma(X_J, Y_J)$ as the average distance of character attributes. SED calculates properties separately. So $d(X_E, Y_E)$ and $J_\sigma(X_J, Y_J)$ have different scales. To calculate the total distance, in this paper, the usual way to combine the two scales is to multiply a and $J_\sigma(X_J, Y_J)$, where a is the number of attributes which participate in JD calculation.

$$D(X, Y) = d(X_E, Y_E) + a \cdot J_\sigma(X_J, Y_J) \tag{5}$$

3.3 GDM Scheme

The GDM rule used in this paper is majority rule. The decision making process can be divided into the following steps.

(1) Set parameter that the number of GDMO that we call it Object_Number.
(2) Calculate the similarity between the input sample and the comparison sample, and put the comparison sample into the Group Decision Making Candidate Set (GDMCS) according to the similarity.
(3) If the similarity calculation between the input sample and all the comparison samples is completed, it will enter the next step; otherwise, it will enter (2).
(4) Select GDMOs. Select the 1st to the Object_Number elements in the GDMCS and put them into the Group Decision Making Object Set (GDMOS).
(5) Organize GDMOS. Calculate the number of elements of the same diagnosis result in GDMOS respectively.
(6) The diagnosis results were obtained. The result with the most elements selected is the diagnostic result.

See Fig. 3 for the flow chart of SGDM.

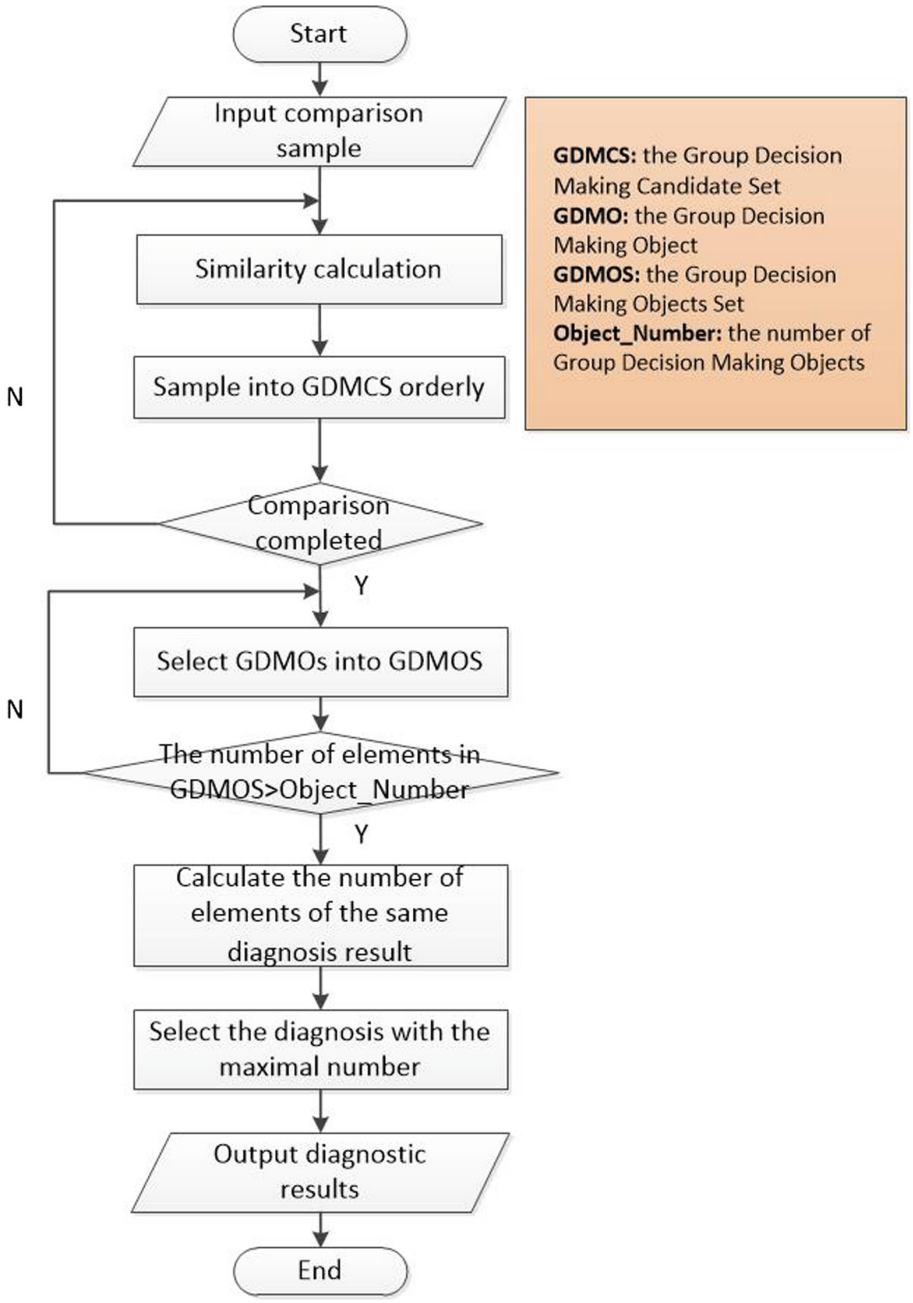

Fig. 3. The flow chart of GDM.

4 Experiment

We compare the accuracy and time consumption of two schemes and three algorithms. Two schemes are similarity scheme and GDM scheme. The three algorithms are SED algorithm, JD algorithm and SED-JD algorithm. At the same time, the accuracy of different Object_Number is compared.

4.1 Experimental Basis

All experiments were run on a Windows 8 64-bit system. The host configuration is as follows: Inter(R) Core(TM) i5-7300HQ CPU, 16 GB RAM.

The experimental data was derived from simulation data at https://github.com/synthetichealth/synthea. After preprocessing these data, we obtained 1066 samples. There were five symptoms in the samples and each with more than 100. We divide the processed samples into two parts, one as the sample set and the other as the test set.

4.2 Experimental Evaluation

Figure 4 is a comparison diagram of the results of the two schemes. SED-JD was used as the similarity measurement algorithm in the comparison. Meanwhile, in the GDM scheme, we select the result (based on accuracy) when Object_Number is 4. As can be seen from Fig. 4, although the two schemes have little difference in result, GDM scheme has better result than similarity scheme on the whole. When the number of samples is about 360, the result of the schemes is the best. This indicates that the selection of samples' number should be appropriate. In addition, the accuracy rate of both schemes has some twists and turns. This is because the samples are disordered, and the distribution of symptoms varies with the number of samples.

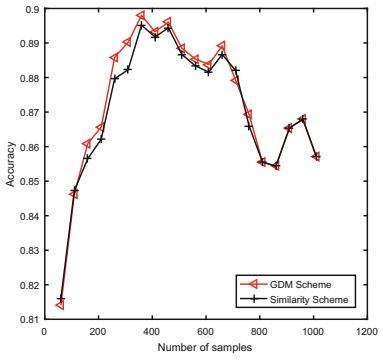

Fig. 4. Schemes comparison diagram.

Figure 5 is a comparison diagram of the three algorithms. GDM scheme that Object_Number is 4 is used in the comparison. As can be seen from Fig. 5, the accuracy of all three algorithms increases rapidly with the increase of the number of samples and then tends to be stable. Meanwhile, we can see that SED-JD algorithm is much better than the other two algorithms.

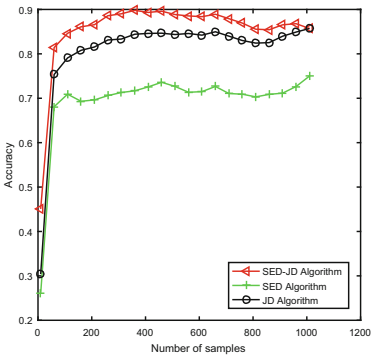

Fig. 5. Algorithms comparison diagram.

Figure 6 shows the selection of different Object_Numbers in the GDM scheme. As we can see from Fig. 6, the accuracy of all three algorithms increases with Object_Number and then decreases slightly. So we can know that the Object_Number selection should not either too large or too small, which means that the Object_Number selection should be appropriate. As can be seen from the figure, SED-JD algorithm and SED algorithm work best when Object_Number is 4, and JD algorithm works best when Object_Number is 3.

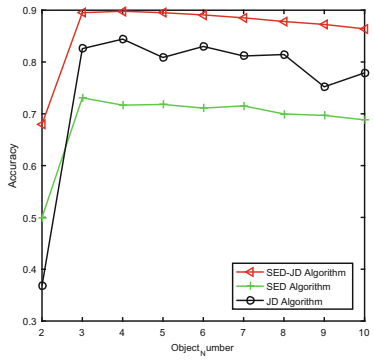

Fig. 6. The selection for Object_Number diagram.

5 Conclusion and Future Work

We mainly study two aspects: SED-JD algorithm and GDM scheme. We compare the accuracy of two schemes and three algorithms. Experimental results show that the SED-JD algorithm proposed in this paper is superior to the other two algorithms. GDM scheme is superior to similarity scheme, but the difference is not large. Therefore, how to optimize GDM scheme is the next step.

References

1. David, J.S., Rodney, C.G.F., Kate, B.: Assessment, criticism and improvement of imprecise subjective probabilities for a medical expert system. In: Machine Intelligence & Pattern Recognition, pp. 285–294 (1990)
2. Liao, S.: Expert system methodologies and applications—a decade review from 1995 to 2004. Expert Syst. Appl. **28**(1), 93–103 (2005)
3. Peiyu, W.: The medical expert diagnoses and analogous system. Jilin University (2008)
4. Si, L., Zijian, L., Tian, L.: Decision making research. J. Hubei Univ. Natl., 87–105 (1988)
5. Fubing, R., Lingling, L.: Advance of group decision making abroad research. J. Mod. Inf., 172–177 (2018)
6. Duncan, B.: On the rationale of group decision-making. J. Polit. Econ. **56**(1), 23–24 (1948)
7. Hwang, C.L., Lin, M.L.: Group Decision Making Under Multiple Critical. Springer, New York (1987). https://doi.org/10.1007/978-3-642-61580-1
8. Yuda, H.: Impossibility theorem and majority rule of group decision making. Science **56**(6), 50–52 (2004)
9. De Silva, N.T., Jayamanne, D.J.: Computer-aided medical diagnosis using Bayesian classifier-decision support system for medical diagnosis. Int. J. Multidiscip. Stud. (2016)
10. Mottalib, M.M., Rahman, M.M., Habib, M.T.: Detection of the onset of diabetes mellitus by Bayesian classifier based medical expert system. Trans. Mach. Learn. Artif. Intell. (2016)
11. Buntine, W.: Learning classification rules using Bayes. In: Proceedings of the Sixth International Workshop on Machine Learning, pp. 94–98 (1989)
12. Oad, K.K.: A fuzzy rule based approach to predict risk level of heart disease. Central South University (2014)
13. Dennis, B., Muthukrishnan, S.: AGFS: adaptive genetic fuzzy system for medical data classification. Appl. Soft Comput. **25**, 242–252 (2014)
14. Sanz, J.A., Galar, M., Jurio, A., Brugos, A.: Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. Appl. Soft Comput. **20**, 103–111 (2014)
15. Reddy, P.V.S., Sadana, A.: Fuzzy medical expert systems for clinical medicine learning through the fuzzy neural network. Int. J. Clin. Med. Res. **2**(5), 54–60 (2015)
16. Agrawal, P., Madaan, V., Kumar, V.: Fuzzy rule-based medical expert system to identify the disorders of eyes, ENT and liver. Int. J. Adv. Intell. Paradig. **7**(3/40), 352–367 (2015)

17. Inbarani, H.H., Azar, A.T., Jothi, G.: Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. Comput. Methods Programs Biomed. **113**(1), 175–185 (2014)
18. Ying, X., Huailong, L., Haitao, W.: Design of diagnosis expert system for children's motor skill disorder. In: Modern Educational Technology, pp. 121–126 (2015)
19. Yunfeng, M.: Design and development of an interpretation expert system for the prevention and control of equine disease. Northeast Agricultural University (2016)
20. Lin, D., Wenru, L.: Biblimetrics analysis of experts system research for rice disease forecast. In: Digital Agricultural Machinery and Equipment, pp. 76–83 (2012)
21. Bo, H., Qing, C., Pengji, Y.: Research on fuzzy decision expert system. In: Mechanical Engineering (1991)
22. Lizhi, H., Zhonghai, Y.: Expert system of hydraulic wrench based on examples. In: Electronic Science and Technology, pp. 122–124+12 (2016)
23. Da, L., Jian, D., Gang, Q.: Control expert system and its application in power plants. In: Northeastern Electric Power Technology (1997)
24. Pan, Z., Bo, W., Xiaoxia, Q.: Integrated application of multiple knowledge representation methods in expert systems. In: Microcomputer Applications, pp. 4–5+18–4 (2004)
25. Zhao, W., Zhaoqing, Y.: Fuzzy expert system inference engine design. Wuhan Institute of Chemical Technology (2003)
26. Xin, Z., Yundou, W., Lijun, G.: Prospects and general methods of designing medical expert system inference machine. Chin. Med. Equip. J., 100–102+11 (2013)
27. Zhigang, G., Xu, J.: Research on data preprocess in data mining and its application. Appl. Res. Comput. **7**, 117–119 (2004)
28. Baosheng, L., Liping, Y., Donghua, Z.: Comparison of some classical similarity measures. Appl. Res. Comput. **23**, 1–3 (2006)
29. Peng, X., Na, L., Changqing, J.: A medical diagnosis expert system based on correlation analysis of features. In: Computer Engineering and Applications (2018)