International Journal of u- and e- Service, Science and Technology
Vol.9, No. 1 (2016), pp.123-128
http://dx.doi.org/10.14257/ijunesst.2016.9.1.14

Liver Function Diagnosis Based on Artificial Bee Colony and K-Means Algorithm

Zhang Lin[2], Li Peng[1,2]* and Qiao Pei-li[2]

*[1] School of Software, Harbin University of Science and Technology,
150080 Harbin, China*

[2] School of Computer Science and Technology, Harbin University of Science and Technology, 150080 Harbin, China

Email:e_roc@126.com

Abstract

The traditional K-Means clustering is sensitive to random selection of initial cluster centroids, easily into the local optimal solution. In this paper, an efficient aggregation algorithm which combined with Artificial bee colony and K-Means algorithm is proposed to apply to the diagnosis of liver function. The algorithm reduced the dependence on the initial cluster centroids and the probability to be trapped by local optimal solution, thus assigning data points to their appropriate cluster more efficient. The experimental results show that algorithm proposed in this paper is superior to the K-Means clustering in diagnosis of liver function.

Keywords: *K-Means clustering; Artificial bee colony; liver disorders diagnosis; aggregation algorithm*

1. Introduction

With the rapid development of computer and database technology, the amount of information in medical has been growing explosively. The development of database technology has resolved the problem of data storage and data retrieval efficiency. As the systemic theory of the latest achievements in data processing, data mining is suitable for the multidimensional medical data analysis which lack of prior knowledge.

As the most common and serious disease, liver disease include viral hepatitis, fatty liver, liver cirrhosis, liver cancer. According to the International Agency for Research on Cancer, liver cancer is the fifth most common cancer in men (523,000 cases per year, 7.9% of all cancers) and the seventh in women (226,000 cases per year, 6.5% of all cancers)[1]. The method of ultrasonic imaging and nuclear medicine diagnosis has obtained ideal effect in liver diagnosis recently[2]. Using the advantage of computer data storage, analyze the diagnosis result of liver patients in the form of chart. This method is conducive to track the ebb and flow of liver patients and has very important significance in liver disease diagnosis. But this method is just make a brief comparison which can’t become the direct basis of liver disease.

Tan and Tang proposed a K-Means clustering based on particle swarm optimization[3]. This method has resolved the problem that the traditional K-Means clustering too dependent on the selection of the initial cluster centroids. But the proposed algorithm has low efficiency and does not guarantee global convergence. Bagirov attempt to overcome the sensitivity to the initialization through eliminate the dependence on random initial conditions, the method is select one cluster at first, and at each circle add a new cluster deterministically to the solution basis of an

ISSN: 2005-4246 IJUNESST

appropriate criterion[4,5]. Although the algorithm can tackle the initialization problem, it has a high complexity. Tzortzis and Likas proposed a novel approach called MinMax k-Means which overcome the drawback of K-Means by altering its objective[6]. The algorithm give weights to each cluster, and the higher the weight the larger the variance. By limiting large variance clusters to tackle initialization problem. In this paper, we propose a novel aggregation algorithm based on k-means and ABC algorithm. Our method is more stable and accurate than traditional clustering algorithms but has the same computational complexity.

2. Computer-aided Diagnosis of Liver Function

The routine technology is feature extraction, feature selection, training classifier and classification with computer-aided diagnosis of liver function. Feature extraction: As an important step in computer-aided diagnosis, the result of feature extraction has directly effect on the accuracy of the final classification. Those who are tested must be fasting for 8 hours before feature extraction, collect 2-millititer blood samples at 4 degree centrifuge, centrifugal 5 min at a speed of 2000 r/min, sera diagnosis within 4 hours. ALT were detected by using microtiter plate pyruvate oxidase method, AST were detected by using immunoinhibition, γ -GT were detected by using the method of Szasz.

Feature selection: We need a number of indicators of liver function for judging when use serum enzymology for liver function diagnosis. Remove those features which have little or nothing to do with the judgment of liver function in serum enzymology. Typically, clinical testing consists of the following 10 key indicators: total protein (TP), albumin (ALB), globulin(GLB), albumin/globulin ratio, total bilirubin(TBIL), direct bilirubin (DBIL), indirect bilirubin (IBIL), aspartate aminotransferase (AST), alanine aminotransferase (ALT), and AST/ALT[7,8].

Because the sample set may contain duplicate data, so we need to deduplicate and classify the sample.

Training classifier: Obtain the prior knowledge of liver function diagnosis.

Classification: The unknown samples will be detected by the trained classifier. The trained classifier classify the current sample with present information.

The process of computer-aided liver function diagnosis as shown in the figure below:

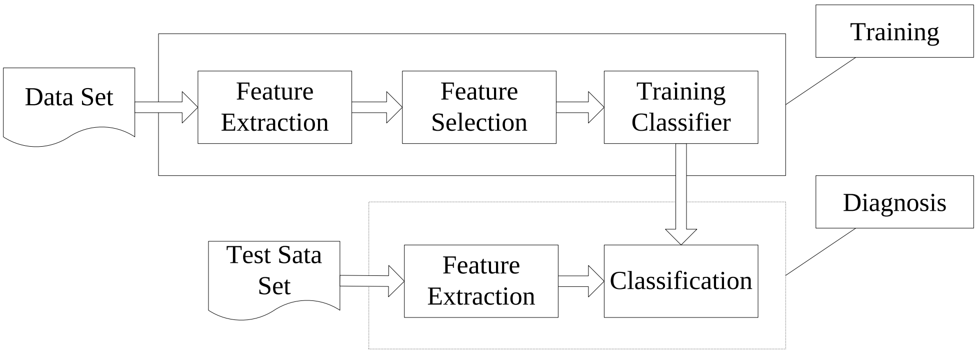

Figure 1. Computer-aided Diagnosis of Liver Function Process

3. Aggregation Algorithm Combined with ABC and K-Means Clustering

3.1. Artificial Bee Colony Algorithm

Artificial bee colony (ABC) algorithm is motivated by food foraging behavior of honey bees. ABC consists of three groups of bees: employed bees, onlookers and scouts[9]. Different groups have respective honey task. The optimization problem is represented by the position of a food source, the function value of the associated solution is equivalent to the nectar amount of a food source. The task of bees be summarized in three stages: At the first step, employed bee chooses a new food source in the neighborhood according to the position of food source in the memory. Onlooker bee chooses a food source according to the information from employed bee, and change the food source position in memory. Employed bee becomes a scout to discover a new food source if the food source was exhausted.

The nectar amount of food source corresponds to the fitness, onlookers chooses a food source is implemented by using:

$$P_i = \frac{fit(i)}{\sum_{i=1}^{SN} fit(i)} \tag{1}$$

Where SN is the amount of food source, $fit(i)$ denote the fitness value of solution.

Employed bees and onlookers produce a new food source according to the memory by using:

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \tag{2}$$

Where $k \in \{1,2,\cdots,S_N\}$, $j \in \{1,2,\cdots,D\}$ are randomly chosen indexes, ϕ_{ij} is a random number in the range[-1,1].

As a very important static parameters in ABC, Limit can prevent the situation that search fall into local best. If food source x_i didn't change after Limit time loops, indicate that this food source has been fallen into local best, and the employed bee becomes a scout bee. The scout find a new food source randomly by using:

$$x_i^j = x_{\min}^j + rand(0,1)(x_{\max}^j - x_{\min}^j) \tag{3}$$

3.2. K-Means Clustering Algorithm

K-Means is one of unsupervised learning algorithms which proposed by Macqueen in 1967 to solve the problem of the well-known cluster[10]. The idea of this algorithm is partition data sets into k clusters. The algorithm is as follows:

1. select k data points $c_1^0, c_2^0, \cdots, c_k^0$ randomly as initial centroids from data sets, superscript denotes iterative computation times in clustering.

2. Calculate the distance between the centroid of a cluster and data points, then each data point will be assigned to the nearest cluster by using:

$$d(x, z_j^{(r)}) = \min\{d(x, z_i^{(r)}), i = 1,2,\cdots,K\} \tag{4}$$

Where $d(x, z_j^{(r)})$ is the distance between $d(x, z_j^{(r)})$, $x \in S_j^{(r)}$, $S_j^{(r)}$ is the sample set which regard $z_j^{(r)}$ as the centroid. Partition all data sets into k clusters with this principle of minimum distance.

3. Recalculate the centroid of each cluster:

$$z_j^{(r+1)} = \frac{1}{n_j^{(r)}} \sum_{x \in S_j^{(r)}} X,\ j = 1,2,\cdots,K \tag{5}$$

Where $n_j^{(r)}$ is the number of data points in $S_j^{(r)}$.

4. steps 3 to 4 are run until $z_j^{(r+1)} = z_j^r,\ j = 1,2,\cdots,K$.

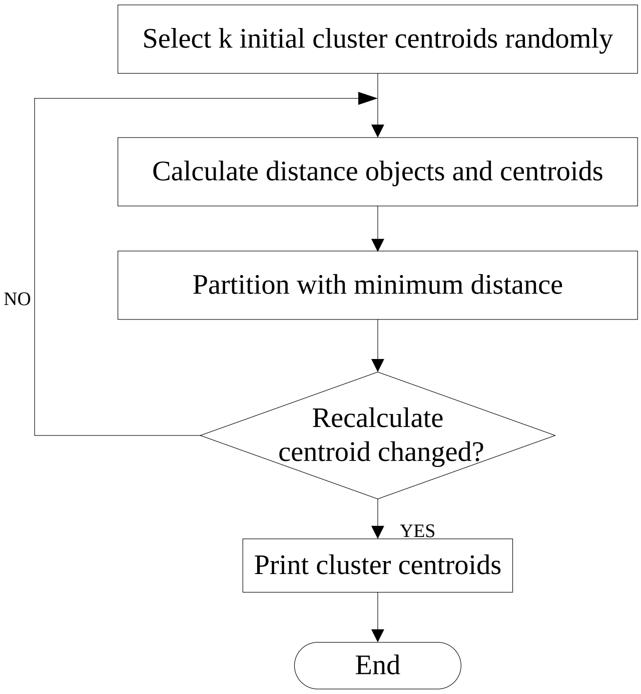
Select k initial cluster centroids randomly
Calculate distance objects and centroids
Partition with minimum distance
NO
Recalculate centroid changed?
YES
Print cluster centroids
End

Figure 2. Process of K-Means Algorithm

3.3. Aggregation Algorithm

In this paper, we combined with ABC and K-means algorithm to implement the clustering process. The concrete progress is: In the process of each iteration, using ABC algorithm optimizes each cluster centroid, then optimizing each cluster centroid with K-Means clustering. Alternate those two algorithms until the end of the clustering. On the one hand, ABC algorithm can reduce the dependence on the initial cluster centroid to K-Means algorithm, on other hand, K-Means enhance the convergence rate.

In this paper, the steps of aggregation algorithm are as follows:

1. Set the number of employed bees, onlookers and scouts, SN denotes the size of food source and it is equal to the number of employed bees. MCN denotes the maximum Iterations, limit is a parameter, k denotes the number of clustering;

2. Initialize the colony, produce $\{Z_1, Z_2, \cdots, Z_{SN}\}$ randomly as the initialized colony. Compute fitness of each food source with (6).

$$fit_i = \frac{1}{1 + \sum_{j=1}^{k} \sum_{\forall x_i \in c_j} d(x_i, c_j)} \tag{6}$$

3. Sorting the colony with the fitness of bees, the top 50% as employed bees and other 50% as onlookers. Employed bees discover food source in the neighborhood with (2), if

fitness of the new position of food source has changed, then replace the old position with v_i, otherwise stated.

4. Onlookers select food source by using roulette wheel selection. If the position of food source no further change, then find a new food source replace for the old one.

5. Make a K-Means Iteration to the food source which on behalf of the cluster centroid.

6. If the current cycle number is less than the maximum cycle number, then go to 3, otherwise output the final cluster centroids.

4. Experiments

Our computing system is Intel Core i5 processor 430M 2.26 GHz, 2 GB of RAM, Matlab R2010a, and Windows 7 Home Premium Edition.

The dataset we used is the BUPA Liver Disorders in the UCI machine learning repository that contains 345 test samples in 7 classes with six features and a selector field. The selector field is used split the data into 2 sets with 138 instances and 207 instances respectively. The 7 attributes of dataset are mean corpuscular volume(mcv), alkaline phosphatase (alkphos), gamma-glutamyl transpeptidase (gammagt), number of half-pint equivalents of alcoholic beverages(drinks) , aspartate amino-transferase (sgot), alamine aminotransferase(sgpt) .

Parameter settings are as follows: population size is 50, maximum iterations MCN=100, Limit=15. Our algorithm is compared with K-Means clustering with the 345 data points of liver disorders. Each algorithm is repeated 10 runs independently, and take one of these experimental results. Compute the minimum, maximum, and average value of clustering error of aggregation and K-Means clustering. Clustering error represents the clustering quality of algorithm which indicates the smaller the better. The clustering quality of both algorithm in the 100 iteration of Liver Disorders show as follows:

Table 1. The Comparison Result of K-Means and Proposed Algorithm

Algorithm	Minimum (e + 004)	Maximum (e + 004)	Average (e + 004)
K-Means	1.7244	2.3545	1.8025
Proposed algorithm	1.7376	2.0625	1.7841

Table 1 show the comparison result of K-Means and proposed algorithm.

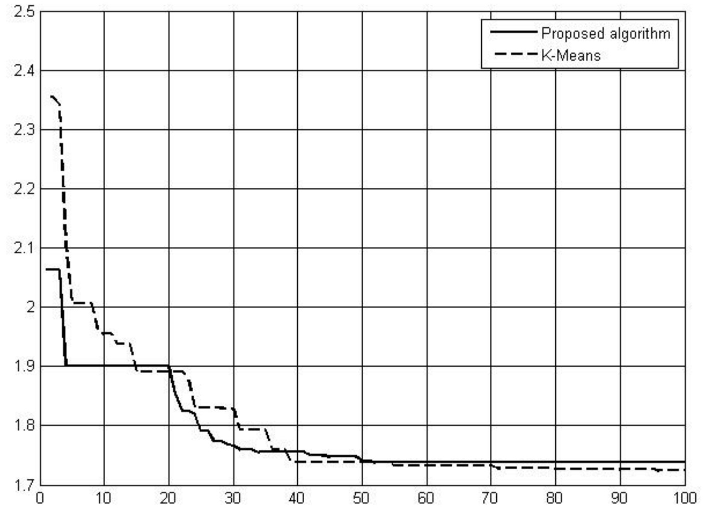

Figure 3. The Clustering Quality of Two Kinds of Algorithm

From Figure 3 and Table 1, it is clear that K-Means too dependent on the selection of the initial cluster centroids, so the stability and accuracy of the algorithm is rather poor. In this paper, the proposed algorithm with combined with ABC and K-Means algorithm to produce results which are more stable and accurate than K-Means algorithm.

5. Conclusion

In this paper, we propose a aggregation algorithm combined with ABC and K-Means algorithm to apply to the diagnosis of liver function. The proposed algorithm overcome the drawback of the K-Means which is too dependent on the selection of the initial cluster. The results show that our algorithm performs much better performance on stability and clustering quality. This algorithm worked marvelously for diagnosis of liver function which the classification characteristic between data points is not obvious. As future works, aggregation algorithm will be applied on large data set.

Acknowledgements

This paper is partially supported by Foundation for University Key Teacher of Heilongjiang Province (1252G023). Li Peng is the corresponding author of this paper.

References

[1] H. B. El Serag, "Epidemiology of viral hepatitis and hepatocellular carcinoma", Gastroenterology. vol. 142, no. 6, **(2012)**, pp. 1264-1273.

[2] U. Eberlein, J. H. Bröer and C. Vandevoorde, "Biokinetics and dosimetry of commonly used radiopharmaceuticals in diagnostic nuclear medicine–a review", European journal of nuclear medicine and molecular imaging, vol. 38, no. 12, **(2011)**, pp. 2269-2281.

[3] D. Tan and D. Tang, "Application of k-means algorithm in diagnosis of liver diseases based on PSO", Computer Era, vol. 12, no. 6, **(2013)**, pp. 34-38.

[4] A. M. Bagirov, "Modified global k-means algorithm for minimum sum-of-squares clustering problems", Pattern Recognition, vol. 41, no. 10, **(2008)**, pp. 3192-3199.

[5] A. M. Bagirov, M. Adil and J. Ugon, "Dean Webb. Fast modified global k-means algorithm for incremental cluster construction", Pattern Recognition, vol. 44, no. 4, **(2011)**, pp. 866-876.

[6] G. Tzortzis and A. Likas, "The MinMax k-Means clustering algorithm", Pattern Recognition, vol. 47, no. 7, **(2014)**, pp. 2505-2516.

[7] Z. Fang-yin, L. Shu-fen and J. L. Caffrey, "Domestic sampling survey of biochemical tests of liver disease", Laboratory Medicine & Clinic, vol. 14, no. 7, **(2014)**, pp. 19-29.

[8] Y. S. Lin, G. Ginsberg and Z. Peng, "Association of body burden of mercury with liver function test status in the US population", Environment international, vol. 16, no. 6, **(2014)**, pp. 88-94.

[9] I. Brajevic and M. Tuba, "An upgraded artificial bee colony (ABC) algorithm for constrained optimization problems", Journal of Intelligent Manufacturing, vol. 24, no. 4, **(2013)**, pp. 729-740.

[10] N. Singh and D. Singh, "The improved K-Means with Particle Swarm Optimization", Journal of Information Engineering and Applications, vol. 3, no. 11, **(2013)**, pp. 1-7.