International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.11 (2015), pp.133-142
http://dx.doi.org/10.14257/ijmue.2015.10.11.13

Missing Value Imputation Method Based on Density Clustering and Grey Relational Analysis

Li Peng[1, 2], Zhang Ting-ting[2], LiangTian-ge[2] and Zhang Kai-hui[3]

[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
[3]Journal Center, HeiLongJiang University, 150080 Harbin, China
e_roc@126.com

Abstract

In the computer-aided medical diagnosis, the problem of missing attribute values in many medical data sets brings a great challenge to data mining. To solve the problem, this paper proposes a method based on density clustering and grey relational analysis. It provides an effective solution for missing medical data. The method uses the characteristic and degree of data samples dynamic relation and the existing attribute value information to impute the missing value, in order to alleviate the difficulty brought by missing data for aided medical diagnosis. By comparison with the experiment, the proposed method can effectively solve the classification problem which causes by missing medical attribute value, accurately predict the patient's health and provide the help to doctor's diagnosis.

Keywords: *Computer-aided medical diagnosis; data mining; DBSCAN; grey relational analysis*

1. Introduction

In recent years, data mining as a method of computer-aided medical diagnosis, has been widely used in medical fields [1]. With the rapid development of medical science, a lot of medical data arises at the same time. Due to the limitation on the data acquisition, information omitting and getting information at a great cost etc,. There are a large amount of data missing problems in many medical data sets. For such data, some traditional algorithms have serious effect of mining potential knowledge from them. If we want to predict the disease trend of patients through the missing medical data set, we need to process the data further [2]. There are three kinds of missing mechanisms: Missing Completely At Random (MCAR), refers to the situation that the data missing doesn't depend on any complete data or any incomplete data; Missing At Random (MAR) implies that data missing only depend on complete data, and not rely on incomplete data. In this situation, we require the data set to contain enough samples so that we can find the implicit information from other complete data. The missing data mechanism discussed in this paper is the MAR. Not Missing At Random (NMAR) refers to that data missing not only depend on the complete data but also rely on the incomplete data [3]. There are many common methods to process missing data, for example direct deletion method which deletes the missing samples from the data sets. At the expense of reducing number of data, we will get a complete data set and then analyze it. But at the same time it also deletes some implicit information and affects the effect of data analysis [4]. Mean substitution method, it replaces the missing values by the mean of all the

ISSN: 1975-0080 IJMUE

observational values and regards it as the new value of the missing data. The multiple imputation in statistics provides several complete data sets and finds the better imputation [5]. KNN imputation method according to the Euclidean distance calculates K samples which is closest to the missing value. Then regard the weighted average of the K values as the missing value[6].

Calculating the similarity degree between attribute values as one of common methods to impute the missing values, has been widely applied to various fields, for example using association rules to impute data. But some traditional method of association rules imputation may be because of few rules (*i.e.*, use one or a few auxiliary information), leading to some missing value can't to fill. It reduces the effect of missing value imputation. This paper proposed a method that based on density clustering and grey relational analysis. The density clustering can make the samples in each cluster have high similarity. Make full use of the intrinsic relationship between variables, and find enough auxiliary imputation information. Then we can use grey relational analysis to find the most similar sample with missing data in each cluster and calculate the most suitable imputation. This method will provide help for subsequent data mining in medical science, which provide computer-aided medical diagnosis for the doctor.

2. Aided Medical Diagnosis Based on Data Mining

Traditional medical diagnostic method is generally the doctor applying their knowledge of pathology and clinical experience in diagnosis, according to the patient's examination results, combining with the patient's medical history and other aspects such as living and working environment, making a careful judgment about their disease. Accurate diagnosis can help patient clear the treatment direction, make treatment plan and strategy, and provide necessary help to the patient's recovery. Because of the influence of subjective factors in the traditional medical diagnosis is very great, the judgment of disease is about the patient's health and life safety, so the aided medical diagnosis method based on data mining is widely used [7].

A lot of information has serious quality problems in the real medical database, such as incomplete data, data with noise, *etc.*, and these problems bring a challenge for aided medical diagnosis. Therefore, before mining the useful information from these data, we need to preprocess them, improve data quality and facilitate subsequent aided diagnostic work [8]. Aided medical diagnosis generally uses the case data from medical information library to obtain some disease diagnosis rules by analysis, in order to aid doctor's diagnosis. In case library, some information needs to put into the medical information library after data pre-process. By selecting the appropriate data mining algorithm, analyze the data of information library, so as to form a data model library. And by analyzing the data of model library, get patient's auxiliary information. Then the information of patient's case will be added to the case library, and the case library becomes richer. This series of process provide the help to the doctors in the diagnosis of patients [9].

The general aided medical diagnosis procedure based on data mining as shown in the figure below:

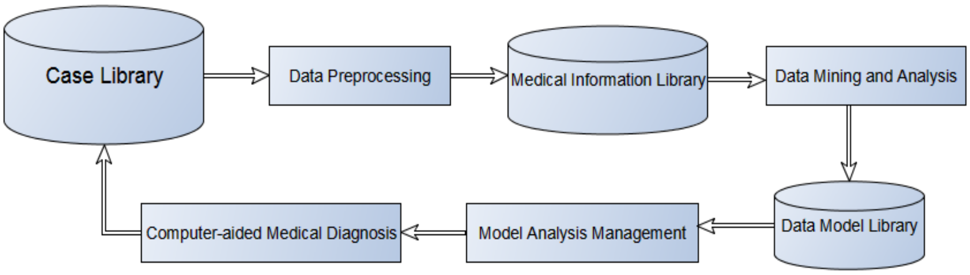

Case Library
Data Preprocessing
Medical Information Library
Data Mining and Analysis
Data Model Library
Model Analysis Management
Computer-aided Medical Diagnosis

Figure 1. Computer-Aided Medical Diagnosis Process

The data in patients' case library is diversity, complexity and uncertainty *etc*. Therefore, organizing the useful medical information library from case library needs to pass the data pre-process, such as solving the problem of concept drift, reducing the noise and dimension. For missing medical data, imputation is the essential step in pre-process. By imputing missing value, it can make missing information fill back to the data sets as much as possible and provide help for future data mining and analysis.

3. Imputation Method Based on Density Clustering and Grey Relational Analysis

3.1. Density-Based Spatial Clustering of Applications with Noise Algorithm

Density-based spatial clustering of applications with noise (DBSCAN) algorithm is a kind of clustering algorithm based on density. It refers to that all the samples in the biggest collection of density are defined as a cluster. High density areas will be separated by low density areas. DBSCAN algorithm needn't to know the shape, density and the number of the clusters. It can get rid of some constraint conditions such as high dimension and outliers [10].

DBSCAN has two parameters: Eps and MinPts. Eps is the radius. The density of a particular point in a given data set refers to that the number of points within a neighborhood which radius is Eps and center is the point. MinPts is the density threshold and represents the minimum number of samples in selected area. If the number of samples in a point's neighborhood has more than a given threshold MinPts, then regard the point as the core. Two core points which distance is within Eps are placed in the same cluster. If one point is not the core and it is in a neighborhood of some cores, we can call it border point. Neither core nor border, we regard the point as the noise [11].

Definition 1 Directly Density-reachable: In a given data set D, sample point x is within the neighborhood area of y and y is core point, so x is directly density-reachable from y .

Definition 2 Density-reachable: In a given data set D, when there is a bunch of sample points $x_1, x_2......x_n$, $Y = x_1$, $X = x_n$, if the point x_{i+1} is directly density-reachable from x_i , then it is density-reachable from X to Y .

Definition 3 Density-connected: In a given data set D, there is a point x , if x is density-reachable to point y_1 and y_2 , then y_1 and y_2 is density-connected.

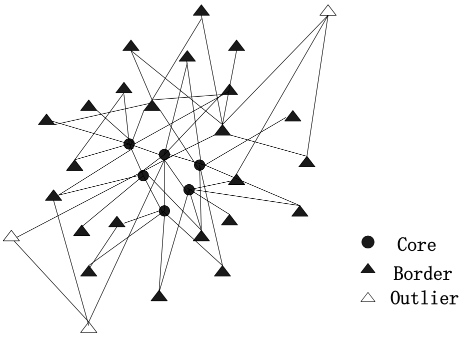
DBSCAN algorithm specific process is as follows:

Input : Data set D

The radius of neighborhood Eps

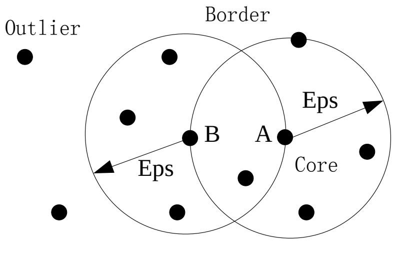
```
        The density of neighborhood MinPts
Initialize all points in the data set D as unmarked objects;
Randomly select a point $x$ as marked object;
    If the neighborhood area of $x$ has at least MinPts points
        create cluster P and put $x$ in P;
            the neighbor objects of $x$ are recorded as set S and all objects are marked in S;
            If the neighborhood area of each point in the S has at least MinPts points
                put these points into S;
                    If the point of S don't belong to any cluster
                        put it into P;
Else record $x$ as noise;
Until without a marked object;
Output : The set of clusters after clustering
```

DBSCAN spatial data cluster and two-dimension points set are shown in the figure below:



Data Cluster

Figure 2. DBSCAN Spatial



Points Set

Figure 3. DBSCAN two-dimension

After using DBSCAN in data set, we can analyze the obtained clusters and calculate grey relational coefficient of missing value in each cluster. Obtain the objects which are closest to missing sample by grey relational grade so as to complete the imputation.

3.2. Grey Relational Analysis

Gray system theory is a kind of theory which is used in uncertainty inference. It is widely applied to solve the problems of special field containing unknown factors. Grey relational analysis is a kind of analysis method in grey system theory. If the system of information completely clear that is called white system. The system of unknown information is called black system. The system which partial information is clear and partial information isn't clear is known as grey system. For a system containing a large number of factors, many factors work together and determine its development momentum. The grey system theory proposes grey relational analysis method for system which has limited data and high grey grade [12]. The basic idea of grey relational analysis is according to the similar of geometric shape of sequence curves to determine whether their relationship is closely. For a given data set, if the tendency of change in two data objects is very similar, we consider they have a high grey relational grade. On the contrary their grey relational grade is low [13].

In a set, the sample which doesn't have missing attribute value is called comparative instance and the sample has missing attribute value is regarded as referenced instance. We can use the grey relational coefficient to compare the tendency between referenced instance and comparative instance. In data set $\mathrm{D} = \{x_0, x_1, x_2, ..., x_n\}$, $x_0, x_1, x_2, ..., x_n$ represent data samples. x_0 is the referenced instance and $x_1, x_2, ..., x_n$ are comparative instances. Each sample contains m attributes, $x_i = (x_i(1), x_i(2), ..., x_i(m))$, $i = 0, 1, 2, ..., n$. First of all, we need to put the data normalization.

The sequence of difference is as following:

$$\Delta_i(k) = \left| x_0(k) - x_i(k) \right| \tag{1}$$

Hereinto, $\Delta_i = (\Delta_i(1), \Delta_i(2), ...\Delta_i(m))$, $i = 0, 1, 2, ..., n$.

The maximum difference and minimum difference is:

$$Y = \max_i \max_k \Delta_i(k) \tag{2}$$

$$y = \min_i \min_k \Delta_i(k) \tag{3}$$

Grey relational coefficient (GRC) is determined to be:

$$GRC(x_0(k), x_i(k)) = \frac{y + \rho Y}{\Delta_i(k) + \rho Y} \tag{4}$$

Hereinto, ρ is discrimination coefficient, $\rho \in (0,1)$, generally $\rho = 0.5$, $k = 1, 2, ..., m$.

Grey relational grade(GRG) is:

$$GRG(x_0, x_i) = \frac{1}{m} \sum_{q=1}^{m} GRC(x_0(k), x_i(k)) \text{ , } i = 1, 2, ..., n \tag{5}$$

Grey relational coefficient represents the level of similarity of two samples on attribute. If $GRG(x_0, x_g) > GRG(x_0, x_h)$, it shows that the level of similarity between x_0 and x_g is larger than that between x_0 and x_h . The sample of maximum GRG is the closest to referenced instance. GRG is closer to1, the samples has better relationship.

3.3. The Method Based on DBSCAN-GRA

After DBSCAN clustering in data set, we need to use the grey relational analysis method to calculate the missing value in cluster. In data set S contain n samples, $S = \{S_1, S_2, ..., S_n\}$, missing data is separated from complete data. $I_1 = \{S_1, S_2, ..., S_r\}$ is complete sample set and called comparative instance set. $I_2 = \{S_{r+1}, S_{r+2}, ..., S_n\}$ is missing sample set and regarded as referenced instance set. Each sample has m attribute values, the complete data set after imputation is $C = \{C_1, C_2, ..., C_n\}$.

The process of DBSCAN-GRA method is described as following:

```
Input : Data set   S
          The number of maximal $GRG(S_i, S_j)$   q
1. Use DBSCAN to get clusters in S and record the number of clusters as $k$ ;
2. For  the Nth clusters( $N \in (1, k)$ )
      For  each missing attribute of sample to calculate:
         if   the attribute value is discrete or symbol
              using the value of the highest frequency to impute;
         if   the attribute value is continuous
              using the mean of value to impute;
3. Calculate the $GRG(S_i, S_j)$ of missing samples in each clusters;
4. Descending order of $GRG(S_i, S_j)$ , choose the front q maximum values;
5. For  q maximum $GRG(S_i, S_j)$
         if    the attribute value is discrete or symbol
               using the value of the highest frequency to instead of the first imputation;
         if    the attribute value is continuous
               using the mean of value instead of the first imputation;
6.Repeat step(3)(4)(5), until the algorithm convergence(predicted value is invariable);
Output : The data set C after imputation
```

In the given missing data set, DBSCAN can accurately cluster samples, although there are noise instances and it can find clusters of different shape in spatial database. Grey relational analysis is introduced to compensate for the regret caused by using the mathematical statistics method. There isn't a high requirement for the number of data and whether have rules. Grey relational grade as similarity metrics between missing values data and complete data can accurately impute the missing attribute values and effectively improve the accuracy of aided medical diagnosis. Using grey relational analysis in the clusters can reduce complexity for missing data imputation relative to the whole data set and better mine the useful information in the existing data.

4. Experimental Results and Analysis

4.1. Effect Verification of Data Imputation

The experiment randomly generates missing data sets from complete data set and imputes the missing data by proposed method. We can get the validation results through comparing imputation and actual value.

At first, we select the Iris data set in UCI to validate. Randomly generate missing data sets and miss rate is 5%, 10%, 15%, 20%. Root Mean Square Error (RMSE) is taken as the evaluation standard of prediction accuracy. The value of RMSE is smaller, the imputation result is better. On the contrary, the result is bad. Our method is compared with Mean imputation and KNN method respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e}_i)^2} \tag{6}$$

Hereinto, e_i is the initial actual attribute value, $\bar{e}_i$ is the predicted attribute value, n is the total quantity of predicted value.

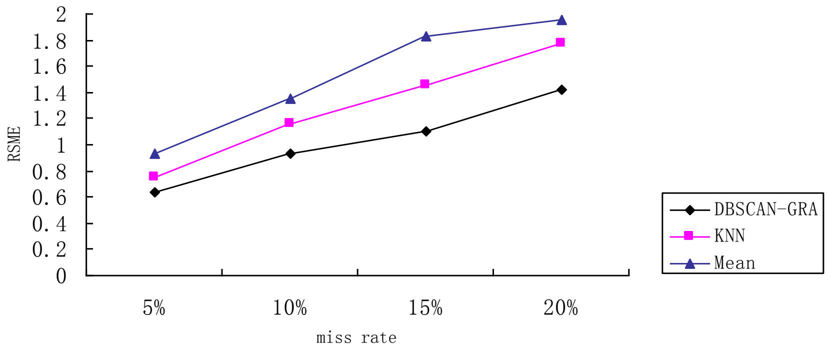


Figure 4. The Comparison of Three Methods

In the paper, the DBSCAN-GRA method can impute the missing attribute value accurately and RSME value is smaller. It can provide effective help for subsequent data mining.

4.2. The Effect Validation of Aided Medical Diagnosis

The medical data is Hepatitis in UCI. Hepatitis has 155 samples, including the basic information of patients, symptoms of disease and whether dead. The numbers of death samples are 32 and survival samples are 123. In this paper, DBSCAN-GRA is used to impute the missing values and predict the diseases of patients by support vector machine (SVM). SVM is based on Vapnik-Chervonenkis (VC) theory and structural risk minimization (SRM) principle. When the condition is linear separable, it constructs the objective function which can separate data set as much as possible. To the linear inseparable, it transforms to a high-dimensional feature space by non-linear map algorithm. After it becomes linear separable, we can use linear algorithm to analysis. SVM can get a better accuracy and adaptability; it also can use kernel function flexibly. The dimension disaster caused by too much features is not obvious [14].

Our experiment uses ROC curve and AUC (the area under the curve) to evaluate the prediction performance. ROC is a graphical method which is showed the compromise between true positive rate and false positive rate. The abscissa is FPR and ordinate is TPR . It intuitively shows the corresponding relation between FPR and TPR .

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{N} \tag{8}$$

TP : the number of true positives. FP : the number of false positives. TN : the number of true negatives. FN : the number of false negatives. The number of virtual positive $P = TP + FN$, while the number of negative $N = TN + FP$.

The range of AUC value is between 0 and 1. When the AUC value is 1, it represents that the classifier achieves the most optimal result. The better the classification result, the nearer ROC curve to the upper left and the closer AUC value to 1.

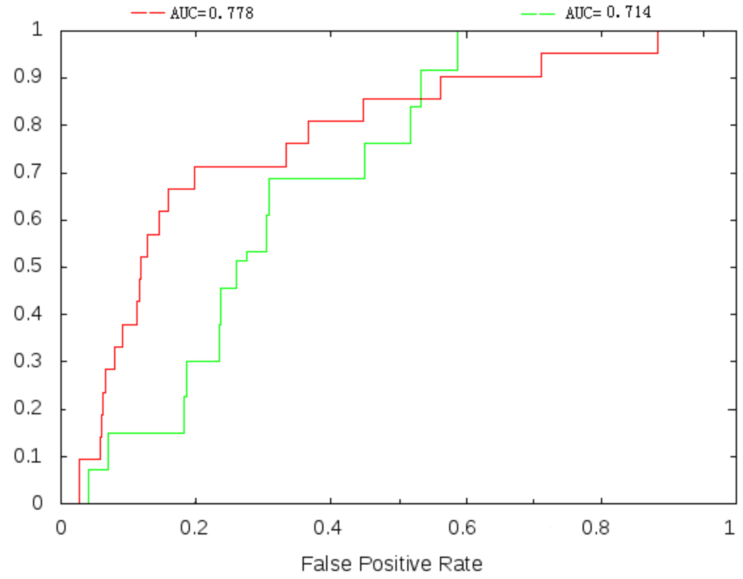


Figure 5. The ROC of DBSCAN-GRA Imputation and Simple Processing

In the picture, green curve represents the classification result after simple processing of missing data such as using the unified value to impute in each attribute. AUC value is 0.714. Red curve is the result by using DBSCAN-GRA to impute and SVM to classify. AUC value is 0.778. The classified effect has obvious improvement than previous simple processing. So the method proposed in this paper can effectively impute missing medical data and improve the effect of the subsequent classification.

5. Conclusion

In the real data set, especially some medical data, the problems of missing attribute value is very common. This paper proposes a new imputation method to reduce the classified effect by missing attribute values. The most similar samples in data set are clustered together by DBSCAN algorithm. For the samples in each cluster, use grey relational analysis to find the most useful auxiliary information. Find the sample which has the most similar attribute value to missing value and predict missing value by known information. Applying this method to the medical data, it can effectively solve the problem of patient information missing which happens in transition from case library to medical information library. And it improves the prediction accuracy of disease in aided medical diagnosis.

But many medical data sets may have other problems, such as concept drift. It may cause the lack of auxiliary information and bring the problems for imputation. So we will further study this problem.

Acknowledgments

This paper is partially supported by Foundation for University Key Teacher of Heilongjiang Province (1252G023).

References

[1] Jerez J. M., Molina I. and García-Laencina P. J., "Missing data imputation using statistical and machine learning methods in a real breast cancer problem", Artificial intelligence in medicine, vol. 50, no. 2, **(2010)**, pp. 105-115.

[2] Tremblay M. C., Dutta K. and Vandermeer D., "Using data mining techniques to discover bias patterns in missing data", Journal of Data and Information Quality (JDIQ), vol. 2, no. 1, **(2010)**, pp. 2-12.

[3] Tuan L. W., Chen Y. J., Li P. L. and Lin K. C., "Evaluation of Multiple Imputation for Longitudinal Ordinal Data under MCAR and MAR Missing-Data Mechanisms", ICIC Express Letter, 2011, 5(6): 1833-1838.

[4] Cismondi F., Fialho A. S. and Vieira S. M., "Missing data in medical databases: Impute, delete or classify", Artificial intelligence in medicine, vol. 58, no. 1, **(2013)**, pp. 63-72.

[5] White I. R., Daniel R. and Royston P., "Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables", Computational statistics & data analysis, vol. 54, no. 10, **(2010)**, pp. 2267-2275.

[6] Batista G. E. and Monard M. C., "A Study of K-Nearest Neighbor as an Imputation Method", HIS, vol. 8, no. 7, **(2002)**, pp. 251-260.

[7] S. Lin, C. Qianhong and T. Hongzhuan, "Identification and treatment of missing data", Journal of Central South University (Medical Science), vol. 38, no. 12, **(2013)**, pp. 1289-1294.

[8] Palaniappan R., Sundaraj K. and Sundaraj S., "A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals", BMC bioinformatics, vol. 15, no. 1, **(2014)**, pp. 2-23.

[9] Srinivas K., Rani B. K. and Govrdhan A., "Applications of data mining techniques in healthcare and prediction of heart attacks", International Journal on Computer Science and Engineering (IJCSE), vol. 12, no. 2, **(2010)**, pp. 250-255.

[10] Pan D. and Zhao L., "Uncertain data cluster based on DBSCAN", Multimedia Technology (ICMT), 2011 International Conference on. IEEE, **(2011)**, pp. 3781-3784.

[11] Bordogna G. and Ienco D., "Fuzzy Core DBScan Clustering Algorithm", 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU), **(2014)**, pp. 100-109.

[12] T. Jing, Y. Bing, Y. Dan and M. Shilong, "Missing data analyses: A hybrid multiple imputation algorithm using Gray System Theory and entropy based on clustering". Applied Intelligence, vol. 40, no. 2, **(2014)**, pp. 376-388.

[13] Huang C. C. and Lee H. M., "A grey-based nearest neighbor approach for missing attribute value prediction", Applied Intelligence, vol. 20, no. 3, **(2004)**, pp. 239-252.

[14] Mishra B. K., Lakkadwala P. and Shrivastava N. K., "Novel Approach to Predict Cardiovascular Disease Using Incremental SVM", Communication Systems and Network Technologies (CSNT), **(2013)**, pp. 55-59.