The Design of Improved Neighborhood Rough Set Algorithm and Its Application in Diabetes Prediction Research

Jing Gao\textsuperscript{1, a}, Ming Cheng\textsuperscript{1, b}, Ruifang Shen\textsuperscript{1, c} and Wei Hu\textsuperscript{1, d}

\textsuperscript{1} Capital University of Economics and Business, Beijing, China

\textsuperscript{a}gaojing@cueb.edu.cn; \textsuperscript{b}waq13241600615@163.com; \textsuperscript{c}ruifangshen@163.com; \textsuperscript{d}892125187@qq.com

Abstract. In order to reduce the importance of neighboring rough set only by single attribute, this paper proposes an improved neighborhood rough set attribute reduction algorithm (INRS), which increases the dependence of conditional attributes based on considering the importance of individual features. Relationship, that is, whether it affects the effect of other conditional attributes on decision attributes after deleting the attribute. The importance of a certain attribute to the decision attribute is divided into two parts: direct influence (after the attribute is deleted, the obtained decision attribute depends on the degree of reduction of the conditional attribute) and indirect influence (relative to the influence of other conditional attributes on the role of the decision attribute when there is no such attribute), so that the importance of each attribute can be clearly identified. When the attribute is reduced, the potential attributes are not reduced. In this paper, the data of diabetes in a hospital of the National Population and Health Science Data Sharing Service Platform was collected. The attribute reduction algorithm was used to attribute the diabetes dataset, and the random forest (RF) was used for classification prediction, which formed high precision. The Diabetes Prediction Model aims to provide support for doctors’ clinical diagnosis and disease research to improve the level of clinical diagnosis and treatment.

1. Introduction

In recent years, artificial intelligence has swept the world. Data mining, as a practice branch of artificial intelligence, has also swept across all occupations. The mining and analysis of medical data has become a hot topic at present, so intelligent diagnosis and treatment has emerged. The neighborhood rough set has great advantages in reduction attribute, which can make up for the lack of rough set processing numerical attribute data set, reduce the noise of continuous data attributes, improve the attribute reduction ability. However, there are also defects that are not fully considered, and the improved neighborhood rough set can find the relationship between attributes to reduce the error caused by the error reduction attribute.

Given that diabetes is one of the major diseases that threaten global human health and life safety, and it is the fuse of various incurable diseases, early diagnosis of diabetes is essential. However, the consumption of diabetes medical resources is a huge and growing global problem. Therefore, taking diabetes as the object, using the mature random forest technology in data mining, the diagnosis information of patients diagnosed with diabetes in a hospital in China is extracted, cleaned, mined and analyzed, diabetes prediction model is established, and the results are evaluated. The feasibility and effectiveness of using data mining technology to establish a high-precision disease diagnosis and
prediction model will help doctors diagnose diseases in a short period and predict possible complications in advance. It provides reference for medical diagnosis and disease control in China.

2. Correlation research on attribute reduction of neighborhood rough sets

Feature selection is an important task of data mining and pattern recognition. Its goal is to select the optimal feature subset with minimum redundancy and maximum recognition ability. Rough set theory is a data analysis theory proposed by Z. Pawlak, a scholar of the Polish Academy of Sciences. The main information processing system is characterized by inaccurate, uncertain or fuzzy information. However, the classical rough set requires the data to be discretized, and the improper discretion will have a greater impact on the later mining. In order to better deal with the continuity attribute, the neighborhood rough set came into being. At present, research scholars have studied a lot of neighborhood rough set attribute reduction algorithms.

There are many types of neighborhood rough set attribute reduction, which are mainly divided into positive region based, information entropy based on difference matrix, attribute reduction based on importance and other methods [1]. Li et al. proposed a new feature reduction method based on neighborhood rough set and resolvable matrix, and used the evaluation index of neighborhood rough set theory to characterize the reparability of classification problems in specific feature spaces. Then design the feature selection strategy to find the optimal feature set by constructing the difference matrix of the input characteristics [2]. Yao et al. proposed that the existing attribute reduction algorithm is not suitable for processing incomplete data with digital attributes and symbol attributes, and proposes an extended incomplete neighborhood rough set model. He defines the concept of neighborhood hybrid entropy to evaluate [3]. In order to solve the redundant and uncorrelated characteristics of high-dimensional low-sample data and low-dimensional high-sample data, Dong et al. proposed an improved binary genetic algorithm using feature granulation (IBGAFG) to select salient features. Improved neighborhood rough set [4].

3. Improved neighborhood rough set attribute reduction algorithm

3.1. Classic neighborhood rough set attribute reduction

In the neighborhood rough set algorithm, the method of feature selection based on importance is used. The importance degree refers to the degree of influence on the decision attribute before and after a certain feature is removed [5]. If the feature is removed, the degree of influence is small, the importance is close to 0; and the greater the importance, the greater the influence of the feature on the decision attribute. Classic feature selection is divided into the following steps:

- For the data set decision system $DS(U, C \cup D, V, f)$ determine the value of its neighborhood parameter and importance lower limit parameter.
- Calculate the neighborhood radius and then calculate the neighborhood of each sample under each attribute subset based on the neighborhood radius. Among them, the neighborhood radius $= \frac{\text{attribute sample standard deviation}}{\text{neighborhood parameter}}$. The sample neighborhood is centered on the value of a certain attribute of a sample, and the radius of the neighborhood is a radius. The sample of all the attribute values contained in the graph is the sample neighborhood of the attribute value.
- The upper and lower approximations of the computational decision subsets of the conditional subsets yield the upper and lower approximations of the decision attributes in the conditional subset.
- The dependence of the decision attribute on the subset of conditional attributes is calculated by the positive region.
- Calculate the importance of decision attributes under each condition attribute, and attribute reduction by attribute importance. Common methods for calculating attribute importance are based on attribute dependency, information entropy, and mutual information. Since the calculation process of calculating the importance based on attribute dependency is simple and
easy to understand, this time based on the dependency-based method: That is, for the decision system \( DS(U, C \cup D, V, f) \), \( \forall H \subseteq C \), if the attribute \( h_i \in H \), the importance of the condition attribute \( C \subseteq D \) to the decision attribute \( D \) is defined as:

\[
sig_{\gamma}(h_i, C, D) = \gamma_c(D) - \gamma_{C-h_i}(D)
\]

(1)

3.2. **Problems in the Reduction of Attributes of Classic Neighborhood Rough Sets**

From the calculation formula of the attribute importance degree above: \( \sig_{\gamma}(h_i, C, D) = \gamma_c(D) - \gamma_{C-h_i}(D) \) the importance degree of an attribute \( h_i \) is equal to the magnitude of the reduction of the decision attribute \( D \) depending on the condition attribute \( H \) after the attribute \( h_i \) is deleted from the condition attribute set \( h_i \). When the importance of an attribute is 0, it means that the attribute has no effect on the decision attribute, so it can be reduced.

Therefore, from the above calculation process of the classical neighborhood rough set, the classical neighborhood rough set only considers the direct influence of the single condition attribute on the decision attribute when calculating the importance through the dependency degree. However, the attributes are not independent, ignoring the interaction between the attributes, which may lead to the reduction of attributes with implicit value, affecting the accuracy of the classification.

For example, the diagnosis of diabetes, doctors manually diagnose diabetes, is also associated with a variety of indicators to make the final diagnosis, and can not only focus on a certain indicator can be diagnosed. And the causes of diabetes are also diverse, not caused by a single cause. Therefore, it is important to consider the relationship between attributes when analyzing data attributes.

3.3. **Improved Neighborhood Rough Set Algorithm**

In the reduction of attributes of classical neighborhood rough set, only the influence of a single feature on decision-making is considered, and the lack of the measurement of the dependency relationship between features may lead to excessive reduction of some features, which will have a great negative effect on the subsequent model effect. In this paper, an improved neighborhood rough set attribute reduction algorithm is proposed to increase the relationship between conditional attributes on the basis of considering the importance of a single feature (that is, after deleting this attribute, whether the effect of other conditional attributes on decision attributes will be affected).

**Definition 1** There is a neighborhood decision system \( NDS = (U, C \cup D) \), and decision attribute \( D \) divides the object set \( U \) into \( N \) classes \( (X_1, X_2, ..., X_n) \), \( \forall H \subseteq C \). For any single attribute \( h_i \), the importance of the condition attribute set \( C \) relative to the decision attribute \( D \) is:

\[
sig^*_\gamma(h_i, C, D) = \sig_{\gamma}(h_i, C, D) + \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} [\sig_{\gamma}(h_j, C-h_i, D) - \sig_{\gamma}(h_j, C, D)]
\]

\[
= \gamma_c(D) - \gamma_{C-h_i}(D) + \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} [\gamma_{C-h_i}(D) - \gamma_{C-h_i-h_j}(D) - (\gamma_c(D) - \gamma_{C-h_i}(D))]
\]

(2)

In equation (2), \( \sig^*_\gamma(h_i, C, D) \) is the importance of the original \( h_i \) to the decision attribute \( D \), which is equivalent to the difference between the dependency of the conditional attribute set \( C \) on the decision attribute \( D \) and the dependency of the conditional attribute set after the \( h_i \) on the decision attribute \( D \).

\( \sig_{\gamma}(h_j, C-h_i, D) \) can be regarded as the importance of the attribute \( h_j \) in the conditional attribute set \( C \) after the division of \( h_i \) to the decision attribute, the formula is equivalent to: \( \sig_{\gamma}(h_j, C-h_i, D) = \gamma_{C-h_i}(D) - \gamma_{C-h_i-h_j}(D) \)

It can be seen as the difference between the dependence of the conditional attribute set on the decision attribute \( D \) after \( h_i \) and the dependence of the conditional attribute set on the decision attribute \( D \) after \( h_i \) and \( h_j \).
\( \sigma(\gamma(h_j, C, D)) \) refers to the importance of the attribute \( h_j \) in the conditional attribute set \( C \) relative to the decision attribute \( D \).

\[ \sigma(\gamma(h_j, C - h_i, D)) - \sigma(\gamma(h_j, C, D)) \]

is the difference between the importance of the set of attribute \( h_j \) in condition attribute set \( C \) divided by \( h_i \) relative to decision attribute \( D \) and the importance of the set of attribute \( h_j \) in condition attribute set \( C \) relative to decision attribute \( D \), that is, the role of \( h_i \) in the \( h_j \)'s influence on decision attribute.

So

\[
\frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \left[ \sigma(\gamma(h_j, C - h_i, D)) - \sigma(\gamma(h_j, C, D)) \right]
\]

is the sum of the importance of all other attributes in the conditional attribute set after \( h_i \), relative to the decision attribute, and the difference in the importance of \( h_i \) over all the conditional attributes relative to the decision attribute. That is, the average effect of \( h_i \) on the influence of other conditional attributes on decision attributes.

The above theory can be explained, so it is feasible to define the calculation formula of the neighborhood rough set importance. That is, \( \frac{1}{n-1} \sum_{j=1, j \neq i}^{n} \left[ \sigma(\gamma(h_j, C - h_i, D)) - \sigma(\gamma(h_j, C, D)) \right] \) is the difference between the influence of other condition attributes on the decision attribute \( D \) relative to the removal of \( h_i \) condition attribute set and the influence on the decision attribute set relative to the original condition attribute set. It is advisable to use this to express the potential value of this attribute \( h_i \). Therefore, it can be seen that the importance of attribute \( h_i \) to decision attribute \( D \) consists of direct influence (after the deletion of attribute \( h_i \), the obtained decision attribute \( D \) depends on the degree of reduction of condition attribute \( h_i \)) and indirect influence (the influence of other condition attributes on the action of decision attribute when there is no such attribute).

4. Improved neighborhood rough set case analysis

4.1. Attribute reduction

In order to verify the superiority of the improved rough set attribute reduction, this paper selects the Iris part data in UCI to calculate the improved neighborhood rough set theory. The data is shown in Table 1. In the table, \( x_i (i = 1, 2 ... 8) \) is a sample of objects, where \( a_3 \) is a symbolic attribute, \( a_4 \) and \( a_2 \) are numeric attributes, and \( D \) is a decision attribute.

**Table 1. Iris part of the original data.**

| sample | \( a_1 \) | \( a_2 \) | \( a_3 \) | \( D \) |
|--------|--------|--------|--------|--------|
| \( x_1 \) | 5.1    | 3.5    | 1      | Setosa |
| \( x_2 \) | 4.9    | 3.0    | 1      | Setosa |
| \( x_3 \) | 5.8    | 2.7    | 1      | Virginica |
| \( x_4 \) | 6.3    | 3.3    | 1      | Virginica |
| \( x_5 \) | 7.1    | 3.0    | 2      | Virginica |
| \( x_6 \) | 6.9    | 3.1    | 2      | Versicolor |
| \( x_7 \) | 7.0    | 3.2    | 2      | Versicolor |
| \( x_8 \) | 6.4    | 3.2    | 1      | Versicolor |

In the table 2: When using the neighborhood rough set for attribute reduction, the data has a different magnitude difference, which results in different attribute importance calculation results. Therefore, in order to obtain a more accurate calculation result of the importance degree, the original data is normalized by the minimum maximum method before the calculation, and the data are all in the \([0, 1]\) interval. The normalized data is calculated according to the steps of the improved neighborhood rough set reduction.
Table 2. Dependence of INRS reduction.

| Conditional attribute subset | Dependence |
|------------------------------|------------|
| $a_1$                        | 0.375      |
| $a_2$                        | 0.25       |
| $a_3$                        | 0.25       |
| $a_1a_3$                     | 0.375      |
| $a_1a_2$                     | 0.625      |
| $a_2a_3$                     | 0.625      |

The new and old importance of each attribute is calculated as shown in table 3:

Table 3. Attribute importance before and after neighborhood rough set improvement

| Importance                  | $a_1$ | $a_2$ | $a_3$ |
|-----------------------------|-------|-------|-------|
| $sig_{r}(a_i, C, D)$        | 0     | 0.25  | 0     |
| $sig_{r}^*(a_i, C, D)$      | 0.25  | 0.3125| 0.1875|

As can be seen from table 3, the old attribute importance calculations that only consider the influence of a single attribute on the decision are mostly 0, which is very disadvantageous for feature selection studies. In the improved neighborhood rough set attribute importance result, each attribute has obvious distinction, and its size relationship is $a_2 > a_1 > a_3$. The importance of each attribute can be clearly identified. When the attribute reduced, the potential attributes are not reduced.

4.2. Improved neighborhood rough set attribute reduction and random forest combination classification algorithm

In disease prediction, feature selection is an indispensable part of the mining process. At this stage, selecting the necessary subset of features, ignoring irrelevant features, is the task of feature selection.

Random forests can compensate for the lack of over-fitting of a single decision tree because of the randomness of its sample and feature selection. Moreover, it has higher classification efficiency for multi-feature data sets than other classifiers. Because of its high operating efficiency and accuracy, researchers have loved it.

In order to effectively deal with medical clinical diagnosis big data sets, this paper first uses the improved neighborhood rough set method to reduce the data set, reduce the attribute dimension, and then use the random forest method to achieve classification. The specific ideas are as follows:

- Using the improved neighborhood rough set algorithm (INRS) for attribute reduction and generating new samples as a training sample set.
- The processed sample data is randomly sampled from the sampled data to form k subsample sets.
- Each subsample set can train a decision tree. When each tree is created, m features (generally $m = \sqrt{M}$) are extracted from M features as nodes of the decision tree. Each node uses the standard with the highest information gain rate to select the branch nodes, thereby splitting and growing downward. The M of each decision tree in the forest does not change.
- Every decision tree grows fully without pruning.
- The generated forest predicts the data, and the final prediction result is determined by the voting result.
5. INRS attribute reduction example verification

5.1. Constructing a predictive model of diabetes characteristics for INRS

Using Matlab tools to program an improved neighborhood rough set for attribute reduction of a diabetes dataset requires the following two steps:

5.1.1. Data import. According to the above research, after removing the two properties of Bp2s and Bp2d, the Diabetes.mat file of the remaining attributes is imported into the Matlab tool to perform the neighborhood rough set attribute reduction.

Table 4. Introduction to Diabetes Data Fields

| The serial number | The field names                     | English names |
|-------------------|------------------------------------|---------------|
| 1                 | Total cholesterol                  | Chol          |
| 2                 | Stable glucose                     | Stan.Glu      |
| 3                 | High density lipoprotein           | HDL1          |
| 4                 | Cholesterol/high density lipoprotein| Ratio         |
| 5                 | Age                                | Age           |
| 6                 | Gender                             | Gender        |
| 7                 | Height                             | Height        |
| 8                 | Weight                             | Weight        |
| 9                 | BMI (Body mass index)              | BMI           |
| 10                | Bone size                          | Frame         |
| 11                | First systolic pressure            | Bp.1d         |
| 12                | First diastolic pressure           | Bp.1s         |
| 13                | waistline                          | Waist         |
| 14                | Hip circumference                  | Hip           |
| 15                | W/H= waistline / Hip circumference | WH            |
| 16                | Genetic history                    | YiChuan       |
| 17                | Smoking                            | XiYan         |
| 18                | Foot symptoms                      | KouGan        |
| 19                | Dry mouth symptoms                 | ZuMa          |
| 20                | Physical fatigue symptoms          | ShenTiFaLi    |
| 21                | Vision loss symptoms               | ShiLiXiaJiang |
| 22                | Glycated hemoglobin                | Glyhb         |

5.1.2. Data normalization. When using the neighborhood rough set for attribute reduction, the data has a different magnitude difference, which results in different attribute importance calculation results. Therefore, in order to obtain a more accurate calculation result of the importance degree, the original data is normalized by the minimum maximum method before the calculation, and the data are all in the [0, 1] interval. Its formula is:

\[ f(x_i) = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \quad (i = 1, 2, \ldots, n) \]  

(3)

Where A and B are the maximum and minimum values of the attribute values, respectively.

5.2. INRS attribute reduction parameter tuning

In the improved neighborhood rough set attribute reduction process, there are mainly two parameters that need to be adjusted:
One calculates the parameters for the neighborhood radius. When the continuous data is discretely divided, the rationality of the partition space needs to be considered. The neighborhood radius is an indicator used by the neighborhood rough set to discretize the data. The sample is divided by a circle with a radius of the neighborhood radius. Neighbor radius = standard deviation / L, L is an adjustable parameter that can be adjusted according to the accuracy of the excavation, usually between 0.5 and 4.

The second is the importance lower limit parameter Sig. The importance is the basis for assessing whether the attribute is preserved in the neighborhood rough set, and is generally infinitely close to zero. The lower importance parameter refers to the limit of the reduction, and the attribute whose importance is greater than this number will be left behind. Since the importance value becomes larger after the algorithm is improved, there is noise, so the importance of too small to be considered when setting the number can also be reduced.

In order to obtain the optimal attribute reduction, the two parameters of the radius and importance of the domain are analyzed, and the following experiments are designed:

- The L parameter is taken in an exhaustive manner, from 0.5 to 4, with an interval of 0.1.
- Sig parameters are set to 0.001/0.01 respectively for comparison.

The above experiments use the classification accuracy as the excellent criterion for evaluating attribute selection, that is, the percentage of accurate classification after classification using random forest, as shown in Table 5 below:

Table 5. Attributes of L and S adjustment reduction

| L   | Reduction result Sig=0.001                     | Reduction result Sig=0.01 |
|-----|------------------------------------------------|---------------------------|
| 0.5 | {1,2,3,4,5,6,9,10,11,12,14,15,16,17,19,22}     |                           |
| 0.6 | {1,3,5,6,8,9,10,11,12,13,15,16,17,19,20,21,22}|                           |
| 0.7 | {1,3,5,6,7,8,9,10,11,12,14,16,17,19,20,21,22} |                           |
| 0.8 | {1,4,5,6,7,8,9,10,16,17,18,19,21,22}          | {5,10,16}                 |
| 0.9 | {1,3,4,5,6,7,8,10,12,14,15,16,18,19,21,22}    | {4,5,10,16}               |
| 1.0 | {1,3,4,5,6,7,8,10,12,14,15,16,19,20,21,22}    | {1,5,10}                 |
| 1.1 | {1,3,4,5,6,7,10,11,12,14,15,16,20,22}        | {1}                      |
| 1.2 | {1,3,5,6,7,11,14,16,20,22}                   | {1,5}                    |
| 1.3 | {1,3,6,7,14,16,20,22}                        | Null                    |
| 1.4 | {1,3,21}                                      | {3}                     |
| 1.5 | {1,4,11,22}                                   | Null                    |
| 1.6 | {1,9,11,22}                                   | Null                    |
| 1.7 | {1,9,11,22}                                   | Null                    |
| 1.8 | {1,9,11,22}                                   | Null                    |
| 1.9 | {9,11,22}                                     | Null                    |
| >=2 | Meaningless                                   | Meaningless             |

The reduced results are classified using a random forest classifier to obtain the respective classification accuracy, as shown below:
Figure 1. Relationship between classification accuracy and L and Sig.

It can be seen from the analysis in Figure 1 that when $L = 0.6$ and $Sig = 0.01$, the improved neighborhood rough set (INR) achieves the optimal reduction, and the reduction result is \{1, 5, 6, 9, 10, 12, 16, 17, 19, 20, 21, 22\}. Namely: Total cholesterol, age, gender, BMI (Body Mass Index), bone size, first diastolic blood pressure, genetic history, smoking, dry mouth symptoms, symptoms of physical fatigue, symptoms of decreased vision, glycated hemoglobin.

6. Summary

This paper analyzes the current research on the attribute reduction of neighborhood rough sets, and optimizes the shortcomings of considering only single attributes, considers the interaction between conditional attributes, and then uses simple real data sets to verify the examples. Optimization has obvious advantages for attribute reduction. Finally, a classification prediction algorithm based on improved neighborhood rough set and random forest combination is proposed, and its idea is briefly introduced. An example verification was carried out using the diabetes dataset to provide an auxiliary method for the correct diagnosis of early diabetes, which will greatly help the later treatment.

Acknowledgement

Project of Education Committee in Beijing (No.SM201810038005, title: Research on the Development of Medical Service System of Elderly in Beijing Based on Deep Belief Network.); Project on Reform of education in Capital University of Economics and Business in 2018 (No. 00791854210160, title: Research and Practice on the Informationization Method of Computer Courses.)

References

[1] Song Zhipeng, Liu Yi and Liu Yu 2012 Same attribute reduction algorithm. J. Computer Engineering and Applications. B 48 52-54
[2] Li B et al 2018 Feature Reduction for Power System Transient Stability Assessment Based on Neighborhood Rough Set and Discernibility Matrix. J. Energies. B 11 185
[3] Yao Sheng et al 2018 Uncertainty measurement and attribute reduction in incomplete neighborhood rough set. J. Journal of Computer Applications. B 38 97-103
[4] Dong H et al 2018 A Novel Hybrid Genetic Algorithm with Granular Information for Feature Selection and Optimization. J. Applied Soft Computing.
[5] Che Xunjian, Hu Qinghua, Yu Daren 2009 Feature selection based on Neighborhood rough set Multi-granularity estimation C. China Joint Conference on Grain Computing.