ICIC Express Letters
Volume **10**, Number **2**, February **2016**

ICIC International ©2016 ISSN 1881-803X
pp. 519–525

A CREDIT RATING ATTRIBUTE REDUCTION APPROACH BASED ON PEARSON CORRELATION ANALYSIS AND FUZZY-ROUGH SETS

Baofeng Shi, Jingxu Zhao and Jing Wang

College of Economics and Management
Northwest A&F University
No. 3, Taicheng Road, Yangling 712100, P. R. China
{ fengbei_wuyu; zhaojingxu1103 }@163.com; wj66xyx@126.com

Received July 2015; accepted September 2015

Abstract. *This paper introduces a credit rating attribute reduction approach based on Pearson correlation analysis and fuzzy-rough sets. First of all, by calculating the Pearson correlation coefficient between attributes, the attribute of a larger correlation coefficient can be deleted. It can eliminate the influence of duplication information on the classification results. And then, utilizing fuzzy-rough sets method, the approximate classification quality coefficient* γ_R *can be obtained. The attribute* x_i *which the approximate classification quality coefficient* $\gamma_R(x_i)$ *equals one can be deleted. It indicates the attribute* x_i *has no significant effect on customers' classification results. The proposed model has been verified using the data of 106 small private businesses in credit rating. The empirical results show that the proposed approach can accurately reduce attributes. Moreover, our research could provide theoretical basis and practical reference for establishing index system in credit rating.*

Keywords: Attribute reduction, Fuzzy-rough sets, Pearson correlation analysis, Credit rating, Small private business

1. Introduction.

With the coming of the era of big data, the attributes reduction exists in pattern recognition, complex system evaluation, multi-objective decision, fault diagnosis, customer classification and so forth. Therefore, it has become more and more important for researchers to find appropriate methods of attribute reduction in all walks of life. For this purpose, many mathematical models are explored as attributes reduction.

In the literature, there is a mass of attributes reduction methods. Căleanu *et al.* [1] studied the problems of feature extraction and classifier design in facial recognition by combining a feature extraction technique and a k-NN statistical classifier method. Experimental results showed that the approach enables them to achieve both higher classification accuracy and faster processing time. In order to recognize key attributes which can effectively distinguish the good customers and bad customers, Shi and Chi [2] established a model for recognizing key attributes based on collinearity diagnostics and Logistic regression significant discrimination. Ding *et al.* [3] established a new minimum attribute self-adaptive cooperation co-evolutionary reduction algorithm based on quantum elitist frogs. The simulation experiments results indicated that the proposed model can achieve the higher performance than the existing models. Yu and Li [4] established a theorem for judging upper and lower approximation consistent sets that are obtained by using a produced dependence space. Then, a new attribute reduction method is proposed to preserve some invariant characters of upper and lower approximation in each decision class. In order to solve the reduction problems in a decision table with multiple attribute types, An and Liu [5] proposed a two-phase genetic reduction algorithm using the attribute importance as evaluation criteria. Tang *et al.* [6] mined the geography knowledge between parking system's location and a variety of geography factor applying fuzzy rough sets mutual information attribute reduction algorithm. Chen *et al.* [7] proposed a novel attribute reduction algorithm combining the variable precision rough set and optimization extreme learning machine method. And the simulation results based on XV-15 showed the high average recognition rate of the proposed method. Zhang [8] created an attribute reduction method combining the advantages of the double-quantitative rough set model based on logical difference of precision and grade. Jiang *et al.* [9] proposed an approximate decision entropy-based attribute reduction algorithm (ADEAR). The experimental results indicated that ADEAR algorithm could obtain higher classification accuracies than the current algorithms, and the computational cost of ADEAR algorithm was relatively low.

Although the existing researches have made great progress in attributes reduction, there are still some shortcomings. First of all, the repeated information between attributes, which could induce the information chaos of evaluation system, cannot be excluded in the existing attributes reduction. And secondly, there is rare literature studying on attribute reduction application in credit rating. In order to overcome the above drawbacks, this paper creates an attributes reduction approach based on Pearson correlation analysis and fuzzy-rough sets. Using the credit rating data of 106 small private businesses, the proposed model is tested.

The contribution of this paper could be summarized as follows. Firstly, the paper develops a novel method which combines fuzzy-rough set and Pearson correlation coefficients for mining the key attributes in credit rating. By calculating the Pearson correlation coefficient between attributes, the attribute of a larger correlation coefficient can be deleted. It can eliminate the influence of duplication information on the classification results. And then, this paper selected the attributes which have significant influence on customers' classification by using fuzzy-rough sets. The empirical results by using the data of 106 small private businesses show that the created model has good information extraction ability. Secondly, our research could provide reference of attribute reduction for establishing index system in credit rating. It could lower the cost of evaluating for commercial banks or cooperative enterprises.

The rest of the paper is organized as follows. Section 2 introduces the methodology of this paper. Section 3 presents the data and empirical analysis of our proposed model. We conclude in Section 4.

2. An Attribute Reduction Model Based on Pearson Correlation Analysis and Fuzzy-Rough Sets.

2.1. Attribute data standardization.

In order to eliminate the influence of the differences of attributes units and dimensionality on attribute reduction, the original data should be transformed into real numbers within the interval [0, 1]. The attribute data can be divided into four classes. That is positive attributes, negative attributes, interval attributes and qualitative attributes. The positive attributes are attributes whose values are the bigger, the better. The negative attributes are attributes whose values are the smaller, the better. The interval attributes are attributes which are reasonable only when they lie in certain intervals. And the qualitative attributes are attributes whose values cannot be directly expressed in real numbers.

It has to be noted that no technique has been proved to be optimal for all kinds of data, so this paper adopts the method to transform the attributes data in literature [10]. Let x_{ij} denote the standard score of the j-th object on the i-th attribute. Let v_{ij} denote the attribute original data of the j-th object on the i-th attribute. Let n denote the number of objects. Let q_1 denote the left boundary of the ideal interval. Let q_2 denote the right boundary of the ideal interval. The standardization equations of positive attributes, negative attributes and interval attributes are shown as Equation (1), Equation (2) and

Equation (3) respectively [10].

$$x_{ij} = \frac{v_{ij} - \min\limits_{1\le j\le n}(v_{ij})}{\max\limits_{1\le j\le n}(v_{ij}) - \min\limits_{1\le j\le n}(v_{ij})} \tag{1}$$

$$x_{ij} = \frac{\max\limits_{1\le j\le n}(v_{ij}) - v_{ij}}{\max\limits_{1\le j\le n}(v_{ij}) - \min\limits_{1\le j\le n}(v_{ij})} \tag{2}$$

$$x_{ij} = \begin{cases} 1 - \frac{q_1 - v_{ij}}{\max\left(q_1 - \min\limits_{1\le j\le n}(v_{ij}), \max\limits_{1\le j\le n}(v_{ij}) - q_2\right)}, & v_{ij} < q_1 \ (a) \\ 1 - \frac{v_{ij} - q_2}{\max\left(q_1 - \min\limits_{1\le j\le n}(v_{ij}), \max\limits_{1\le j\le n}(v_{ij}) - q_2\right)}, & v_{ij} > q_2 \ (b) \\ 1, & q_1 \le v_{ij} \le q_2 \ (c) \end{cases} \tag{3}$$

By rational analysis and expert investigation for qualitative attributes, the scoring standard of qualitative attributes can be obtained.

2.2. **The first attribute reduction based on Pearson correlation analysis.** The Pearson product-momentum correlation coefficient was developed by Karl Pearson from a related idea introduced by Pearson in the 1880s [11]. It is a measure approach of the linear correlation between two random variables. By calculating the Pearson correlation coefficient between attributes, the attribute of a larger correlation coefficient can be deleted. It can eliminate the influence of duplication information on the classification results, so the attribute system can be simplified. The Pearson correlation coefficient r_{ij} of two attributes x_{ki} and x_{kj} is defined as the covariance of the two variables divided by the product of their standard deviations. And it can be equivalently defined by:

$$r_{ij} = \frac{\sum\limits_{k=1}^{m}(x_{ki} - \bar{x}_i)\sum\limits_{k=1}^{m}(x_{kj} - \bar{x}_j)}{\sqrt{\sum\limits_{k=1}^{m}(x_{ki} - \bar{x}_i)^2 \sum\limits_{k=1}^{m}(x_{kj} - \bar{x}_j)^2}} \tag{4}$$

where x_{ki} and x_{kj} denote the standard score of the corresponding attributes, and $\bar{x}_i$ and $\bar{x}_j$ denote the average value of the corresponding attributes.

Equation (4) is applied to calculating the Pearson correlation coefficient. The bigger the absolute value $|r_{ij}|$ of Pearson correlation coefficient is, the higher the correlation coefficient between two attributes is. As a matter of experience, the critical value of correlation coefficient is equal to 0.8 [10]. That is to say, if the absolute value of correlation coefficient between two attributes is bigger than 0.8, it means the two attributes reflect the duplication information and one of the attributes should be deleted.

2.3. **The second attribute reduction based on fuzzy-rough set.** Fuzzy-rough set methodology was developed as a non-parametric data-mining approach. It has been applied to a variety of fields, such as pattern recognition, sustainable and green supply chain, and operations management concerns [12,13]. In this paper, we use the fuzzy-rough set methodology for attribute reduction. The steps of fuzzy-rough set attribute reduction are as follows.

Step 1: define a fuzzy similarity relation. Let o_s and o_i denote the evaluation objects respectively. Let U denote the evaluation object set. Let x_{ij} denote the standard score of the j-th object on the i-th attribute. Let n denote the number of objects. The fuzzy similarity relation $R(U, \alpha)$ between o_s and o_i is given by Equation (5).

$$R(U, \alpha) = \left\{ (o_s, o_i) \in U \times U \middle| \frac{1}{n}\sum_{j=1}^{n}|x_{ij} - x_{sj}| \le \alpha \right\} \tag{5}$$

In Equation (5), if (o_s, o_i) belong to R, then 1-α is called the similarity degree of object o_s and o_i. The parameter α is used for measuring the distance between the two evaluation objects o_s and o_i. The bigger the similarity degree 1-α is, the smaller the distance of the two evaluation objects o_s and o_i is, and the greater the similarity degree of the two objects o_s and o_i is. Then the two objects can be classified as the same class. This paper selects 0.7 as the threshold value of the similarity degree 1-α [13,14]. It means that if the distance of the two evaluation objects is less than 0.3, then the two objects can be divided into the same class.

Step 2: calculate the corresponding fuzzy similarity class. Define all the objects which have the same similarity class with o_i as the fuzzy similarity class of the object o_i, denoted by $FR(o_i, \alpha)$. The equation is given by

$$FR(o_i, \alpha) = \left\{ o_s \in U \middle| \frac{1}{n} \sum_{j=1}^{n} |x_{sj} - x_{ij}| \leq \alpha \right\} \tag{6}$$

According to the given similarity degree 1-α, we can divide objects into different categories utilizing Equation (6).

Step 3: the determination of the low approximate set of variable precision rough set. Let X denote the classification results of evaluation objects with all attributes. Let $R(x_i)$ denote the new classification results of evaluation objects deleting the attribute x_i. X and $R(x_i)$ can be obtained by Equation (6). Let $|\cdot|$ denote the number of elements in a set. Let β denote the threshold value of error. Let $\underline{R_\beta}(x_i)$ denote the set of objects with the same classification results of two categories. Then we have

$$\underline{R_\beta}(x_i) = \cup \left\{ x \in U \middle| \frac{|X \cap R(x_i)|}{|R(x_i)|} \geq \beta \right\} \tag{7}$$

In Equation (7), the ratio of $|X \cap R(x_i)|$ and $|R(x_i)|$ indicates the bigger the ratio value is, the smaller the influence of the attribute x_i on the classification result is. The attribute x_i should be deleted.

Next we will analyze the range of the parameter β. When the two classification results X and $R(x_i)$ are exactly the same, we have $|X \cap R(x_i)|/|R(x_i)| = |R(x_i)|/|R(x_i)| = 1$. When the two classification results X and $R(x_i)$ are completely different, we have $|X \cap R(x_i)|/|R(x_i)| = 0/|R(x_i)| = 0$. As a result, the range of the ratio $|X \cap R(x_i)|/|R(x_i)|$ is belonging to the closed interval $[0, 1]$. Similarly, the range of the parameter β is belonging to the closed interval $[0, 1]$. In this paper, we select 0.9 as the threshold value of the parameter β [14].

Step 4: the calculation of the approximate classification quality coefficient $\gamma_R(x_i)$. Let $\left|\underline{R_\beta}(x_i)\right|$ denote the number of objects in Equation (7). Let $|U|$ denote the number of all of objects. The calculation formula of approximate classification quality coefficient $\gamma_R(x_i)$ is as follows.

$$\gamma_R(x_i) = \left|\underline{R_\beta}(x_i)\right| / |U| \tag{8}$$

Equation (8) is the ratio of the value $\left|\underline{R_\beta}(x_i)\right|$ and the number of all of objects $|U|$. The approximate classification quality coefficient accurately describes the influence of deleting an attribute on the evaluation results in fuzzy-rough set. If the approximate classification quality coefficient is equal to 1 after deleting an attribute, it means that the attribute has no significant effect on the classification results. And the attribute should be deleted.

3. An Application Example.

3.1. Sample and data source.

Data is collected from a Chinese government-owned commercial bank that deals with 106 small private businesses from Shaanxi province [15]. There are 33 attributes for each small private business, as shown in Table 1.

Table 1. The original data and standardized data of attributes for small private business

(*a*) Index	(*b*) Feature layers	(*c*) Attributes	(*d*) Attribute type	Original data of attributes v_{ij}			Standardized data of attributes x_{ij}		
				(1) J. Luo	…	(106) W. Yang	(107) J. Luo	…	(212) W. Yang
1	X_1 Basic information	$X_{1,1}$ Marital status	Qualitative	1	…	1	1.00	…	1.00
…		…	…	…	…	…	…	…	…
8		$X_{1,8}$ Loan purpose	Qualitative	3	…	3	0.50	…	0.50
9	X_2 Capacity of repayment	$X_{2,1}$ Liquidity ratio	Positive	4.15		0.12	0.05		0.00
…		…	…	…	…	…	…	…	…
15		$X_{2,7}$ Repayment to net income ratio	Negative	23.48	…	41.84	0.79	…	0.63
…	…	…	…	…	…	…	…	…	…
27	X_5 Macro environment	$X_{5,1}$ Per capita savings balance	Positive	28329.49	…	28329.49	0.35	…	0.35
…		…	…	…	…	…	…	…	…
33		$X_{5,7}$ Industry cycle index	Positive	96.85	…	122.53	0.00	…	0.74

3.2. **The establishment of attributes extraction model for small private business.**

(1) The standardization of attributes data. According to the attribute type in Column *d* of Table 1, take the original data of positive attributes v_{ij} from Column 1 to 106 of Table 1 into Equation (1), the original data of negative attributes v_{ij} into Equation (2) and the original data of interval attributes v_{ij} into Equation (3), and the standardized data of attributes x_{ij} are obtained. According to the attribute type in Column *d* of Table 1, standardized data of qualitative attributes can be obtained based on the scoring criteria of qualitative attributes [10]. The standardized scoring of attributes x_{ij} is shown in Column 107 to 212 of Table 1.

(2) Attributes extraction using Pearson correlation analysis. Substituting the standardized data of attributes x_{ij} in Table 1 into Equation (4), the correlation coefficients r_{ij} of 33 attributes are obtained, as shown in Table 2.

Table 2. Correlation coefficient of attributes

(*a*) Index	(*b*) Attributes	(1) $X_{1,1}$	…	(5) $X_{1,5}$	…	(33) $X_{5,7}$
1	$X_{1,1}$ Marital status	1.000	…	−0.123	…	0.126
…	…	…	…	…	…	…
5	$X_{1,5}$ Industry of small private business engaged	−0.123	…	1.000	…	**−0.839**
…	…	…	…	…	…	…
33	$X_{5,7}$ Industry cycle index	0.126	…	**−0.839**	…	1.000

As mentioned in Section 2.2, if the absolute value of the correlation coefficient r_{ij} of two attributes is more than the threshold 0.8, it indicates that the two attributes reflect the duplication information and one of them should be deleted. This paper deleted six attributes using Pearson correlation coefficient. They are "$X_{1,5}$ Industry of small private business engaged", "$X_{3,2}$ Net profit", "$X_{5,1}$ Per capita savings balance", "$X_{5,2}$ GDP growth rate", "$X_{5,4}$ GDP per capita" and "$X_{5,6}$ Engel's coefficient". Then, we reserved 27 attributes.

(3) Attributes extraction using fuzzy-rough set. Taking the standardized data of reserved 27 attributes in Table 1 into Equation (5) to Equation (8), the 27 approximate classification quality coefficients $\gamma_R(x_i)$ can be obtained.

As mentioned in Section 2.3, if $\gamma_R(x_i)$ is equal to 1, the attribute x_i should be deleted. This paper deleted twelve attributes utilizing fuzzy-rough set. They are "$X_{1,2}$ Gender",

"$\mathrm{X}_{2,1}$ Liquidity ratio", "$\mathrm{X}_{2,4}$ Owners equity", "$\mathrm{X}_{4,3}$ Return on assets" and so on. Then, the final credit rating index system is established, including 15 attributes.

3.3. **The efficiency analysis of the proposed model.** In order to test the effectiveness of the proposed attributes reduction model, we use a method of information contribution measure in Reference [16].

Substitute the 15 selected attributes data into the information contribution measure model [16], and the information content can be calculated $I(X_1', \cdots, X_{15}') = 0.227$. And in the same way, we can calculate the information content of the total 33 attributes $I(X_1, \cdots, X_{33}) = 0.267$. The information contribution rate of the selected attributes is given by $r = I(X_1', \cdots, X_{15}')/I(X_1, \cdots, X_{33}) = 0.233/0.267 = 0.873$. It means that the selected attributes retained 87.3% information content of audition index information. In a word, the proposed attributes reduction approach has good information extraction ability for the small private business in credit rating.

4. **Conclusion.** It is very important for researchers to find appropriate attribute reduction methods in credit rating. In this paper, the authors create an attributes reduction approach based on Pearson correlation analysis and fuzzy-rough sets. We deleted the attributes with duplication information utilizing Pearson correlation analysis. And then, this paper selected the attributes which have significant influence on customers' classification by using fuzzy-rough sets. The proposed approach has been verified using the data of 106 small private businesses of a Chinese state-owned commercial bank. The empirical results show that the selected attributes retained 87.3% information content of audition attributes information. It indicates that the created model has good information extraction ability. Moreover, our research could provide theoretical basis and practical reference for establishing index system in credit rating.

We just proved the effectiveness of the proposed model using a small private business data set in this paper. An important question to be answered in future research is that the reliability and effectiveness of the proposed model should be tested in a variety of data sets, such as SMEs credit rating, farmers' credit rating and so forth.

Acknowledgements. The research was supported by National Natural Science Foundation of China (Nos. 71503199, 71373207 and 71301017), Project Funded by China Postdoctoral Science Foundation (No. 2015M572608), Doctoral Scientific Fund Project of the Ministry of Education of China (No. 20130204110034), Basic Business Project of Humanities Social Sciences for Central University (No. 2015RWYB09), Shaanxi Province Postdoctoral Science Foundation Funded Project, Young Talent fund of University Association for Science and Technology in Shaanxi, China.

REFERENCES

[1] C.-D. Căleanu, X. Mao, G. Pradel, S. Mogad and Y. Xue, Combined pattern search optimization of feature extraction and classification parameters in facial recognition, *Pattern Recognition Letters*, vol.32, pp.1250-1255, 2011.

[2] B. Shi and G. Chi, A model for recognizing key factors and applications thereof to engineering, *Mathematical Problems in Engineering*, vol.2014, pp.1-9, 2014.

[3] W. Ding, J. Wang and Z. Guan, A minimum attribute self-adaptive cooperation co-evolutionary reduction algorithm based on quantum elitist frogs, *Journal of Computer Research and Development*, vol.51, no.4, pp.743-753, 2014.

[4] C. Yu and J. Li, Attribute reduction in variable precision rough set based on dependence space, *Pattern Recognition & Artificial Intelligence*, vol.27, no.12, pp.1065-1070, 2014.

[5] L. An and S. Liu, Two-phase genetic algorithm for attributes reduction, *Systems Engineering – Theory & Practice*, vol.34, no.11, pp.2892-2899, 2014.

[6] M. Tang, X. Wang et al., Site selection decision-making with GIS for mechanical parking system based on mutual information attribute reduction, *Systems Engineering – Theory & Practice*, vol.35, no.1, pp.175-182, 2015.

[7] X. Chen, X. Wang, Y. Huang and X. Wang, Fault diagnosis for tilt-rotor aircraft flight control system based on variable precision rough set-OMELM, *Control and Decision*, vol.30, no.3, pp.433-440, 2015.

[8] X. Zhang, Attribute reduction for the double-quantitative rough set model based on logical difference of precision and grade, *Systems Engineering – Theory & Practice*, vol.35, no.1, pp.1-7, 2015.

[9] F. Jiang, S. Wang, J. Du and Y. Sui, Attribute reduction based on approximation decision entropy, *Control and Decision*, vol.30, no.3, pp.65-70, 2015.

[10] B. Shi, J. Wang, J. Qi and Y. Cheng, A novel imbalanced data classification approach based on logistic regression and fisher discriminant, *Mathematical Problems in Engineering*, vol.2015, pp.1-12, 2015.

[11] K. Pearson, Note on regression and inheritance in the case of two parents, *Proc. of the Royal Society of London*, vol.58, pp.240-242, 1895.

[12] C. Bai and J. Sarkis, Integrating sustainability into supplier selection with grey system and rough set methodologies, *International Journal of Production Economics*, vol.124, no.1, pp.252-264, 2010.

[13] C. Bai and J. Sarkis, Green supplier development: Analytical evaluation using rough set theory, *Journal of Cleaner Production*, vol.18, no.12, pp.1200-1210, 2010.

[14] Z. Pawlak, Rough sets, *International Journal of Computer and Information Sciences*, vol.11, no.5, pp.341-356, 1982.

[15] *Credit Risks Decision and Evaluation System of Microcredit for Merchants for Postal Savings Bank of China (the 2nd Edition)*, Chi Guotai Research Group of Dalian University of Technology, 2011.

[16] G. Chi and Z. Li, Model of information contribution of evaluation index system based on principal component-entropy, *Science Research Management*, vol.35, no.12, pp.137-144, 2014.