Fuzzy-Rough Nearest-Neighbor Classification Approach

Haiyun Bian, Lawrence Mazlack
Applied AI Lab
ECECS Department
University of Cincinnati
Cincinnati, OH 45221
{bianh, mazlack}@ececs.uc.edu

Abstract

This paper proposes a new fuzzy-rough nearest-neighbor (NN [1]) approach based on the fuzzy-rough sets theory. This approach is more suitable to be used under partially exposed and unbalanced data set compared with crisp NN and fuzzy NN approach. Then the new method is applied to China listed company financial distress prediction, a typical classification task under partially exposed and unbalanced learning space. Results suggest that the compared with crisp and fuzzy nearest neighbor classification methods, this method provides more accurate prediction result under this research design.

1. Introduction

Pattern recognition is defined as a search for structure in data [2], with cluster analysis, classification and feature selection as the three main tasks. Classification is basic to all of our intellectual activities [9]. Consequently, Classification is the most basic element in human's representation the reasoning process. Since the representation and reasoning process of human being is inherited uncertain, a great deal of research effort has been devoted to develop theories to modeling the uncertainties involved in it.

Considering a classification problem, if the prior probabilities and the state conditional densities of all classes are known, the Bayesian decision theory produces the optimal results in the sense that it minimizes the expected misclassification rate [7]. However, in empirical classification problems, the actual probability distribution of the population is unknown. Under this circumstance, many non-Bayesian classification techniques, such as clustering and discriminant analysis, are designed based on the notion of the similarity or distance in the feature space that describes the observations. K nearest-neighbor (K-NN) algorithm [5] is one of them for its simplicity and its quite satisfactory results obtained in many small sample size problems. It is argued that K-NN classification is preferred for those classification problems with data that is only partially exposed to the system prior employment [11]. Rough sets theory is based on the assumption that the misclassification caused by the classifier results from the imperfect learning space, i.e. imperfect feature vector description about the elements in the universe. So an interesting research question proposed here is to incorporate the rough uncertainty into the fuzzy K-NN classifier [11], which we name as fuzzy-rough nearest neighbor classification approach.

This following paper has three sections. First we introduce the fuzzy rough NN algorithm based on the fuzzy rough sets theory, and simulation result for this algorithm is listed. Secondly, we apply this method to China listed company financial distress prediction application. Finally comes the conclusion part.

2. Fuzzy Rough NN Approach

2.1. Fuzzy Rough NN Algorithm

Conventional fuzzy K-NN algorithm [2] assigns an unlabeled pattern x to the class which appears the most among its k nearest labeled neighbors. The algorithm is described as follows.

Conventional fuzzy NN algorithm:

Part A: get the k nearest neighbors of the test pattern x

Let $X=\{x_1, x_2,\ldots, x_n\}$ be the set of already labeled data (training data), and $C=\{c_1,c_2,\ldots c_c\}$ is the result classification space. Let x be the unlabeled test data.

[1] NN in this paper is a short form for nearest neighbor.

Input x;
Set K, $1 \le K \le n$;
Set the iteration counter *count*=1;
For all $x_j \in X$ $(1 \le j \le n)$ **Do**
Compute $\| x\text{-}x_j \|$
If $(i \le K)$
include x_j in the set of K-nearest neighbors and increase *count* by 1
else if (x_j is closer to x than any previous nearest neighbor)
Begin
Delete the farthest of the K-nearest neighbors
Include x_j in the set of K-nearest neighbors
End
End For

Part B: approximate x by the k-nearest neighbors

For all $c_j \in C$ $(1 \le i \le c)$ **Do**
Compute $u_i(x)$
End For

Where

$$u_i(x)=\frac{\sum_{j=1}^{K} u_{ij}\left(\|x\text{-}x_j\|\right)^{\frac{-2}{m-1}}}{\sum_{j=1}^{K}\left(\|x\text{-}x_j\|\right)^{\frac{-2}{m-1}}}$$

Here, u_{ij} represents the membership of x_j to the i^{th} class, and m determines how heavily the distance is weighted when calculating each neighbor's contribution to the membership value.

Part A is to choose some of the training data points that are similar to the test data point as its neighbors. And part B is to use the membership functions of the selected neighbors to compute the approximated membership of the test data point.

The new fuzzy-rough nearest neighbor classification approach is a further generalization of the fuzzy nearest neighbor approach [11]. The main idea of the algorithm is listed in the flowing.[2]

Let $X = \{x_1, x_2, \ldots, x_n\}$ be the training data set, and x be the test data. Let $\Phi=(F_1,F_2)$ be a fuzzy partition on X, where:

$F_1=(f_{11},f_{12},\ldots,f_{1n})$ $\quad 0 \le \sum_{j=1}^{n} f_{1j} \le n$

$F_2=(1\text{-}f_{11},1\text{-}f_{12},\ldots,1\text{-}f_{1n})$.

Here f_{1j} means the membership degree that x_j is similar to the test pattern x compared with all elements in the training set.

Then we can approximate the output class C_c over this fuzzy partition over X.

$\mu_{\underline{Cc}}(F_1)=\inf_j \max(1\text{-}f_{1j}, u_{jc}')$ $\qquad 1 \le j \le n$

$\mu_{\overline{Cc}}(F_1)=\sup_j \min(f_{1j}, u_{jc}')$ [3] $\qquad 1 \le j \le n$

Finally $\mu_{\underline{Cc}}(F_1)$ and $\mu_{\overline{Cc}}(F_1)$ can be used as the lower and upper approximation of the membership of F_1 to C_c, which is also the membership of x to C_c.

Compared with the conventional fuzzy K-NN algorithm, this new algorithm can also be divided into two parts. The fisrt part is exactly the same as part A of conventional fuzzy K-NN algorithm. But its second part adopts a different approximation method to get the fuzzy-rough memberships of the test data point. This new approximation method includes not only the fuzzy uncertainties, but also the rough uncertainties.

2.2. Simulation Result

Simulation for the fuzzy-rough NN algorithm is done on some popular used data sets from the UCI machine learning repository [14]. Comparison is made between crisp NN, fuzzy NN and this new fuzzy rough NN approach.

The number of training cases and test cases are determined in the following way. For Weld data, training cases and test cases are provided separately by the author [13], so we follow his sample size in our research. For the remaining data set, since no separating is provided by the source [14], we adopts the popular used 70/30 criteria, i.e., 70 percent of the cases are used for training, and remaining 30 percent are for testing. Comparisons of the simulation result between crisp nearest neighbor, Fuzzy NN and Fuzzy-Rough NN are listed in table 1.

From the worst performance in Tyroid and Breast-cancel-Wisconsin, we notice that fuzzy NN approach is most sensitive to unbalances in different classes[4]. For example, there are 50 cases belonging to

[2] For detailed information refer to [3] and [4].

[3] Here the inf and sup is defined in [8].

[4] Partially exposed space refers to the situation when some of the classes do not occur in the training data set, or when some attributes are absent from the feature space. Unbalanced learning space refers to the situation when the numbers of cases for each classes have large differences.

class 1 in Thyroid data set, while only 10 for class 2 and 3. So the errors come from the unbalance in the class distribution. From the simulation result, we get that fuzzy-rough NN approach is more robust than fuzzy NN approach when dealing with unbalanced data set.

Table 1 Simulation Result

	Crisp NN				Fuzzy NN	Fuzzy-Rough NN
	K=1	K=3	K=5	K=7	K=5	K=5
Iris	1/45[5]	1/45	1/45	1/45	1/45	1/45
Weld	8/45	5/45	4/45	7/45	10/45	9/45
Thyroid	4/70	9/70	5/70	6/70	20/70	6/70
Breast_cancel	0/140	5/140	5/140	6/140	21/140	0/140

In conclusion, when the sample is partially disclosed or unbalanced, this new fuzzy-rough NN method outperforms crisp NN and fuzzy NN method based on the simulation results. It may be more suitable when dealing with partially exposed and unbalanced domain.

3. Financial Distress Prediction

3.1. Background Introduction

Corporate financial distress or failure has been a focal point of issue in financial analysis. The use of financial data, or financial ratios more specifically, to predict corporate financial distress/failure has been the major methodology for this research topic. Since these financial ratios have been long considered as objective indicators, different sets of ratios have been used to distinguish between financial distressed/failed and non-distressed/non-failed companies since mid 1960's. Many statistical based approaches have been used to do the prediction task, mainly including discriminant analysis, logit analysis, linear probability model [1].

More Recently, new methods for predicting financial distress/failure are developed due to the advantage of computer and information science, such as neural networks, decision trees and rough sets approach [6]. These methods come mainly from the artificial intelligence and machine learning area, and have been proved to be successful in many other application areas besides financial distress/failure prediction. Comparisons between these new techniques and the traditional statistical methods are been one of the hot debate. However, no univariate conclusions have been made yet [12].

Since no unifying theory of corporate failure has been developed up till now, all the models available do not constitute an explanatory theory of failure/distress. Rather they summarize, via statistics aggregation or other techniques, information contained in a firm's financial statements to determine whether or not the firm's financial profile most resembles the financial profiles of previously failed/distressed or non-failed/non-distressed firms. They can, therefore, be more accurately classified as descriptive tools of a pattern recognition nature [10]. On the other hand, the percentage of financial distressed companies is very small. So this application domain can be characterized as unbalanced and imperfect.

3.2. Data Sampling

This research focuses on the China listed companies in telecommunications and computer industry sector. We identified 56 ratios from literature review, and then after we delete those are duplicate, we got 28 financial ratios. Among the 28 ratios, we further find that 4 of them are unavailable in the annual report of the companies. So finally the number of variables ends with 24, as shown in appendix A. These variables are computed directly from the annual report of the listed companies according to them definitions in the literature, and all of them are interval variables. No noise data is considered in this application given two reasons. First of all, all the financial data from the annual report are audited, so there should be little chance to have wrong data contained in it. Secondly, our double check on the data furtherly reduces the possibility of having dirty data. Another reason is my method here does not deal with noisy data perspective. So here we take the sampled data as clean and accurate.

Decision tree is used to select 7 significant variables out of 24, which are listed in the following table

.

Table 2 Variables selected by the decision tree method

CA_CL: current assets / current liabilities, also called current ratio
NI_TA: net income / total asset
CF_TD: cash flow / total debt
S_TA: sales / total assets
NI_SE: net income / stockholders' equity
GP_S: gross profit / sales
CL_TD: current liabilities / total debts

[5] 1/45 means that one out of 45 test data is wrongly classified.

3.3 Exploitation of Uncertainties

We can identify at least the following two uncertainties when using nearest neighbor approach to do the listed company profit analysis.

Given a sample of records $X=\{x_1, x_2,\ldots, x_n\}$ for either labeled as non-distressed (class 0 or C_0) or distressed (class 1 or C_1) listed companies, and an unlabeled company x, we describe the problem as following. Find in X those companies whose financial profiles are most similar to x's, and decide the label of x based on the founded companies labels.

Fuzzy uncertainty occurs due to the following two factors.

(1) How typical each $x_i \in X$ $(0 \le i \le n)$ is for C_0 and C_1

C_0 and C_1 are not mutually exclusive crisp sets, but are fuzzy sets that are overlapping. Each x_i does not exclusively belong to C_0 or C_1, but belongs to C_0 and C_1 with a certain membership degree.

This uncertainty is defined by the initialization technique mentioned in Chapter 2 that fuzzifies the memberships of the training data to C_0 and C_1 in the following way:

The K-nearest neighbors of each training pattern x_i are found, and the membership of u_{ij} in each class j is assigned as

$$u_{i0}=\begin{cases} 0.51+(n_0/K)*0.49 & \text{if } x_i \text{ belongs to class 0} \\ (n_0/K)*0.49 & \text{if } x_i \text{ does not belong to class 0} \end{cases}$$

$$u_{i1}=\begin{cases} 0.51+(n_1/K)*0.49 & \text{if } x_i \text{ belongs to class 1} \\ (n_1/K)*0.49 & \text{if } x_i \text{ does not belong to class 1} \end{cases}$$

n_0 and n_1 are the numbers of the neighbors found that belong to class 0 and class 1 respectively.

(2) How similar the test pattern x is to every $x_i \in X$ $(0 \le i \le n)$

The similarity decreases as the distance between the test pattern x and the x_i increases. This similarity is quantified in form of fuzzy membership functions in the following way:

$$f_x(x_i)=\frac{\|x- x_i\|^{-2/(m-1)}}{\sum_{i=0}^{n} \|x- x_i\|^{-2/(m-1)}}$$

Since the fuzzy membership has to be in [0, 1], we normalize $f_x(x_i)$ into the range [0,1] finally. $f_x(x_i)$ is presented as f_i in the following section for simplicity purpose.

Rough uncertainty appears due to incomplete knowledge about the classification process, generally the input representation is not perfect.

Specifically, incomplete representation may cause the following problems:

(1) Two records x_i and x_j may have similar neighbors based on the available feature, thus it is expected that their output class labels should also be similar. Due to the incomplete representation, they actually belong to different classes.

(2) Similarity between the test pattern x and x_i may also computed based on incomplete feature space. Two training cases x_i and x_j who may actually have different similarity to x may get same similarity measures in this incomplete feature space.

The rough uncertainty leads to the problem: how to measure the uncertainty when we want to use the fuzzy similarity neighbors F of the test pattern x to approximate the fuzzy class C_0 or C_1. This kind of uncertainty is solved by the fuzzy-rough sets theory in the following way:

$$\mu_{\underline{Cc}}(F)=\inf_i \max(1- f_i, u_{ic}')$$

$$\mu_{\overline{Cc}}(F)=\sup_i \min(f_i, u_{ic}') \quad (1 \le i \le n \quad 0 \le c \le 1)$$

Here, $\mu_{\underline{Cc}}(F)$ represents the certain membership degree to classify test pattern x to class c (0 or 1) based on its neighborhood. $\mu_{\overline{Cc}}(F)$ represents the possible membership degree to classify test pattern x to class c (0 or 1) based on its neighborhood.

2.4. Result and Discussion

The following table lists the prediction result of fuzzy NN and fuzzy rough NN.

Using the unmatched and unbalanced training data set and test data set, fuzzy-rough NN approach shows the best overall prediction accuracy level at 78.37%, when using the decision tree feature selection method. The result shows that for this China listed company financial distress prediction problem, fuzzy-rough NN approach offers a viable, if not the best alternate approach for our research design compared with the fuzzy NN.

Table 3 Result comparison

Method	Error
Fuzzy NN (K=5)	12/37
Fuzzy-Rough NN (K=5)	8/37

However, overall prediction accuracy may not

represent the actual value in this application since predicting a distressed company into a non-distressed one is much more costive than predicting a non-distressed company into a distressed one. So we compare the accuracy of different approaches by introducing two error types.

Type I errors refer to the situation when actually distressed company is classified as non distressed one, and Type II error refer to non-distressed company is classified into distressed company. It is obviously that Type I error is more costive than Type II error, and our objective should be to reduce Type I error while keep Type II under a satisfactory level. The prediction result is listed in table 3.12 after introducing these two error types.

Table 4 Result comparison: Type I and Type II

Method	Error	
	Type I	Type II
Fuzzy NN (K=5)	12/12	0/25
Fuzzy-Rough NN (K=5)	3/12	5/25

Considering that Type I error is much more costive than Type II error, fuzzy-rough NN approach provides the better results in identifying the non-profitable companies from the whole population.

When using decision tree feature selection method, Fuzzy-rough NN approach has the lower Type I error at 3/12 (25%). Although fuzzy NN approach has the lowest Type II error at 0/25 (0%), it is achieved by classifying all the test instances into non-distressed companies. Overall, fuzzy-rough NN approach has the better performance in minimizing Type I error, while having satisfactory Type II error.

4. Conclusion

This research presents a fuzzy-rough NN classification approach which performs better under partially exposed and unbalanced domain compared with the crisp NN and fuzzy NN approach. Since the result of this fuzzy-rough NN approach contains not only upper but also lower membership degree, more meaningful interpretation can be drawn from the output of the new approach, which in return provides the decision maker more valuable information. Both the simulation on machine learning database and the application in China stock market distressed company prediction suggest that fuzzy-rough NN approach not only provides a comparable prediction accuracy with fuzzy NN method, it advantages over fuzzy NN are best demonstrated when dealing with incomplete feature space and unbalanced classification problem.

This research has some limitations. Firstly, we limit the simulation for fuzzy-rough NN approach to small data set no larger than 1000 instances, thus may restrict its performance estimation in larger data sets. Future research may focus on simulating this approach on larger data sets.

Secondly, we limit the simulation on data set that has no missing value or having missing value less than 5 percent of all the instances. Future research is to test the performance on data sets that have large volume of missing values, since the performance of algorithms may deteriorate very fast with the number of missing values increase. Since missing value is ineluctable in many application domains, bad performance in handling larger percentage of missing values may prohibit its applicability.

Thirdly, China stock market profit analysis limits to one Telecommunications and computer industry sector only. It restricts its explanation capability in other industry sectors. Due to the insufficiency of data, this research is done by using contemporary out-of-sample testing. Future research may do the test on future dated data, which will be more convincing.

[1] Altman E.I., *Corporate financial distress and bankruptcy: a complete guide to prediction & avoiding distress and profiting from bankruptcy*, John Wiley & Sons, 1993

[2] Bezdek J.C., "*Pattern Recognition with Fuzzy Objective Function Algorithms*," Plenum Press, New York, 1981

[3] Bian H.Y., "Fuzzy-Rough Nearest Neighbor Classification: an Integrated Framework," *proceedings of IASTED International Symposium on Artificial intelligence and Applications*, 2002, 160-164

[4] Bian H. Y., "Fuzzy-rough NN Approach and its Application to China Listed Companies Profit Prediction," *Master of Philosophy Thesis*, City University of Hong Kong, 2002

[5] Cover T.M., and Hart P.E, (1967), "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, IT vol(13), January 1967, 21-27

[6] Dimitras A.I., Zanakis S.H. and Zopounidis C., "A survey of business failures with an emphasis on prediction methods and industrial applications", *European Journal of Operational Research*, 1996, 487-513

[7] Dubois D. and Prade H., (1990), "Rough Fuzzy Sets and Fuzzy Rough Sets," *International Journal of General Systems*, 17, 1990, 191-209

[8] Duda R.O. and Hart P.E., *Pattern Classification and Scene Analysis*, New York: Wiley, 1973

[9] Estes W. K., *Classification and cognition: Oxford Psychology series 22*, Oxford University Press & Clarendon Press, 1994
[10] Keasey K. and Watson R., (1991), "Financial distress prediction models: a review of their usefulness", *British Journal of Management*, Vol 2, 1991, 89-102
[11] Keller J.M., Gray M.R. and Givens J.A., "A Fuzzy K-Nearest Neighbor Algorithm," *IEEE Transactions on Systems, Man and Cybernetics*, 15(4), 1985
[12] Lin F.Y. and McClean S., (2001), "A data mining approach to the prediction of corporate failure", *Knowledge-Based Systems*, 14, 2001, 189-195
[13] Mitchell H.B., and Schaefer P.A., "A "soft" K-nearest neighbor voting scheme", *International Journal of Intelligent Systems,* Vol. 16, 459-468, 2001
[14] Murphy P.M., and Aha D. W., *UCI Repository of Machine Learning Databases*, http://www.ics.uci.edu/~mlearn/MLRepository.h