An Approach for Selective Ensemble Feature Selection Based on Rough Set Theory*

Yong Yang[1,2], Guoyin Wang[2], and Kun He[2]

[1] School of Information Science and Technology,
Southwest Jiaotong University,
Chengdou, 610031, P.R.China
[2] Institute of Computer Science & Technology,
Chongqing University of Posts and Telecommunications,
Chongqing, 400065, P.R.China
{yangyong,wanggy}@cqupt.edu.cn, pandahe_916@yahoo.com.cn

Abstract. Rough set based knowledge reduction is an important method for feature selection. Ensemble methods are learning algorithms that construct a set of base classifiers and then classify new objects by integrating the prediction of the base classifiers. In this paper, an approach for selective ensemble feature selection based on rough set theory is proposed, which meets the tradeoff between the accuracy and diversity of base classifiers. In our simulation experiments on the UCI datasets, high recognition rates are resulted.

Keywords: Rough set, ensemble learning, selective ensemble, feature selection.

1 Introduction

Rough set is a valid theory for data mining. The most advantage of rough set is its great ability for attribute reduction (feature selection). It has been successfully used in many domains such as machine learning, pattern recognition, intelligent data analyzing and control algorithm acquiring, etc [1][2][3][4][5]. Based on the feature selection, some classifiers can be built. There are always over one reduction for an information system. Thus, it is a problem to choose a suitable reduction or integrate several reductions into a system to get better performance.

Ensemble learning has been a hot research topic in machine learning since 1990s'[6]. Ensemble methods are learning algorithms that construct a set of base classifiers and then classify new objects by integrating the prediction of the base classifiers. An ensemble system is often much more accurate than each base classifier. Ditterrich proved the effectiveness of ensemble methods from the viewpoint

* This paper is partially supported by National Natural Science Foundation of China under Grant No.60373111 and No.60573068, Program for New Century Excellent Talents in University (NCET), Natural Science Foundation of Chongqing under Grant No.2005BA2003.

J.T. Yao et al. (Eds.): RSKT 2007, LNAI 4481, pp. 518–525, 2007.

of statistic, computation and representation in [7]. As a popular machine learning method, ensemble methods are often used in pattern recognition, network security, medical diagnosis, etc [7][8][9][10].

A necessary and sufficient condition for an ensemble system to be more accurate than any of its base classifiers is that all base classifiers are accurate and diverse. An accurate classifier is one that has an error rate less than random guessing on new instances. Two classifiers are diverse if they make different prediction errors on unseen objects [7]. Besides accuracy and diversity, another important issue for building an efficient ensemble system is the choice of the function for combining the predictions of the base classifiers. There are many techniques for the integration of an ensemble system, such as majority voting, weighted voting, reliability-based weighted voting, etc [8].

There are many methods proposed for ensemble. The most popular way for ensemble is to get different subset of the original dataset by resampling the training data set many times. Bagging [9], boosting [10] and cross-validation are all such ensemble methods. These methods work well especially for unstable learning algorithms, such as decision trees, neural network. Some other methods are also studied, such as manipulating the output targets [11], injecting randomness into classifiers [12]. Besides these methods, there is another effective approach for ensemble, which uses different feature subsets, and is usually called ensemble feature selection [13]. Ensemble feature selection (EFS) is also a classical ensemble method. It takes different feature subset as the input features for a base classifier construction.

There are two methods for generating base classifiers and integrating the predictions of base classifiers. One is called direct strategy, the other is called over producing and choosing strategy. The direct strategy aims to generate an ensemble of base classifiers directly in the training period. The over producing and choosing strategy is also called selective ensemble, which creates a lot of base classifiers at first, and then chooses a subset of the most suitable base classifiers and generates the final prediction.

In this paper, based on rough set theory and ensemble learning theory, a selective ensemble feature selection method is proposed. The rest of this paper is organized as follows. In Section 2, based on the basic concepts and methods of rough set theory and the diversity of ensemble learning, an algorithm for selective ensemble feature selection is proposed. Simulation experiment results are illustrated in Section 3. Finally, conclusion and future works are discussed in Section 4.

2 Ensemble Feature Selection Based on Rough Set Theory

2.1 Basic Concept of Rough Set Theory

Rough set (RS) is a valid mathematical theory for dealing with imprecise, uncertain, and vague information. It has been applied successfully in such fields as

machine learning, data mining, pattern recognition, intelligent data analyzing and control algorithm acquiring, etc, since it was developed by Professor Pawlak in 1980s [1][2]. Some basic concepts of rough set are introduced here for the convenience of following discussion.

Def.1 A decision information system is a formal representation of a data set to be analyzed. It is defined as a pair $s = (U, R, V, f)$, where U is a finite set of objects and $R = C \cup D$ is a finite set of attributes, C is the condition attribute set and $D = \{d\}$ is the decision attribute set. With every attribute $a \in R$, set of its values V_a is associated. Each attribute a determines a function $f_a : U \to V_a$.

Def.2 For a subset of attributes $B \subseteq A$, an indiscernibility relation is defined by $Ind(B) = \{(x, y) \in U \times U : a(x) = a(y), \forall a \in B\}$.

Def.3 The lower approximation $B_-(X)$ and upper approximation $B^-(X)$ of a set of objects $X \subseteq U$ with reference to a set of attributes $B \subseteq A$ may be defined in terms of equivalence classes as follows:
$B_-(X) = \bigcup\{E \in U/Ind(B)|E \subseteq X\}$, $B^-(X) = \bigcup\{E \in U/Ind(B)|E \cap X \neq \Phi\}$.

They are also called as the B_- lower and B^- upper approximation respectively. They can also be defined as follows:

$B_-(X) = \{x \in U|[x]_B \subseteq X$, $B^-(X) = \{x \in U|[x]_B \cap X \neq \Phi\}$.

Where, $[x]_B \in U/Ind(B)$ is the equivalence class of object induced by the set of attributes $B \subseteq A$.

Def.4 $POS_P(Q) = \bigcup_{x \in U/Ind(B)} P_-(X)$ is the P positive region of Q, where P and Q are both attribute sets of an information system.

Def.5 A reduction of P in an information system is a set of attributes $S \subseteq P$ such that all attributes $a \in P - S$ are dispensable, all attributes $a \in S$ are indispensable and $POS_S(Q) = POS_P(Q)$. We use the term $RED_Q(P)$ to denote the family of reductions of P. $CORE_Q(P) = \bigcap RED_Q(P)$is called as the Q-core of attribute set P.

Def.6 The discernibility matrix M_DC=$\{c_{ij}\}_{n*n}$ of an information system S is defined as:

$$c_{ij} = \begin{cases} \{a \in C : x(_i) \neq x(_j)\}, D(x_i) \neq D(x_j) \\ 0 \qquad\qquad , D(x_i) = D(x_j) \end{cases} \quad i = 1, 2, ..., n. \tag{1}$$

Where, $D(x_i)$ is the value of the decision attribute. Based on the dicernibility matrix, all possible reducts can be generated. An attribute reduction algorithm based on dicernibility and logical operation is proposed in [3]. The detailed algorithm is introduced in Algorithm 1.

Any attributes combination of C_0 as well as a conjunctive term of P' can be an attribute reduction of the original information system. All possible reducts of the original information system can be generated with the Algorithm 1. Using these reducts, classifiers could be built which have the same classification ability

Algorithm 1. Attribute reduction algorithm based on dicernibility matrix and logical operation

Input : An information system S with its discernibility matrix M_DC.
Output: Reduction of S.
1. Find all core attributes (C_0) in the discernibility matrix M_DC. $Redu = C_0$.
2. Find the set (T) of elements (C_{ij}'s) of M_DC that is nonempty and does not contain any core attribute. $T = \{C_{ij} : C_{ij} \cap C_0 = \Phi \wedge C_{ij} \neq \Phi\}$.
3. Each attribute is taken as a Boolean variable. A logic function (P) is generated as the Boolean conjunction of disjunctions of each components belonging to element (C_{ij}) of T. That is, $P{=}\wedge_{ij}\{\vee_k\{a_k\} : a_k \in C_{ij} \wedge C_{ij} \in T\}$.
4. Express the logic function P in a simplified form (P') of a disjunction of minimal conjunctive expressions by applying the distributivity and absorption laws of Boolean algebra.
5. Choose suitable reduction for the problem.

as the original whole decision table. Therefore, these classifiers can be taken as candidate base classifiers of an ensemble system.

2.2 Diversity in Ensemble Method

2.2.1 Measurement of Diversity in Ensemble

Theoretically speaking, if base classifiers are more diverse between each other, an ensemble system will be more accurate than its base classifiers. There are a number of ways to measure the diversity of ensemble methods. Some of them are called pairwise diversity measures, which are able to measure the diversity in predictions of a pair of classifiers, and the total ensemble diversity is the average of all the classifier pairs of the ensemble. For example, plain disagreement, fail/non-fail disagreement, Q statistic, correlation coefficient, kappa statistic and double fault measures [14][15][16]. Some others are called non-pairwise diversity measures, which measure the diversity in predictions of the whole ensemble only. For example, the entropy measure, measure of difficult, coincident failure diversity, and generalized diversity [14][15][16].

The double fault measure (DF) can characterize the diversity between base classifiers. It is the ratio between the number of observations on which two classifiers are both incorrect. It was proposed by Giacinto in [17]. It is defined as follows.

$$Div_{i,j} = \frac{N^{00}}{N^{11} + N^{10} + N^{01} + N^{00}}. \tag{2}$$

Where, N^{ab} is the number of instances in the data set, classified correctly (a=1) or incorrectly (a=0) by classifier i, and correctly (b=1) or incorrectly (b=0) by classifier j. The denominator in (2) is equal to the total number of instances N.

2.2.2 Relationship Between Diversity Measure and Integration Method

As discussed above, there are many diversity measure and integration methods. What is the relationship between a diversity measure and an integration

method? Is it effective when we choose a diversity measure and an integration method randomly for an ensemble system? The relationship between diversity measure and integration method is a hot research topic in ensemble learning [14][15][16][18]. In [18], the relationship between 10 diversity measures and 5 integration methods were discussed. It was found that there was little correlation between the integration methods and diversity measures. In fact, most of them showed independent relationship. Only the double fault measure and the measure of difficult showed some correlations greater than 0.3. The measure of difficult showed stronger correlation with the integration methods than the double fault measure. Unfortunately, it was more computationally expensive.

In this paper, the double fault measure and integration method of majority are used in this proposed ensemble method since they showed higher correlation and they both have lower computation complexity.

2.3 Selective Ensemble Feature Selection Base on Rough Set Theory (SEFSBRS)

Based on rough set theory and the diversity measure of the double fault measure, a selective ensemble feature selection method based on rough set theory is proposed here.

Firstly, all possible reducts are generated based on the discernibility matrix of a training set. All candidate base classifiers are generated with the reducts. Secondly, based on the diversity measure defined in Equation (2), all base classifiers are clustered on validation set, and then, a pair of base classifiers, which are the most diverse among two clusters, are chosen from each two clusters. Therefore, the classifiers which are more accurate and more diverse among all the classifiers are chosen for the ensemble system. At last, the majority voting is taken as the integration method for ensemble, and the final prediction can be taken on the testing set. The detailed algorithm is introduced in Algorithm 2.

Algorithm 2. Selective ensemble feature selection base on rough set theory

Input : Decision tables of the training set, validation set and testing set.
Output: Final ensemble prediction.
Apply Algorithm 1 on the training set to generate its all reducts.
Construct all the classifiers using the reducts.
for *each classifier* **do**
  Calculate $div_{(i,j)}$ of each two classifiers on the validation set according to Equation (2).
end
Based on all $div_(i,j)$, all classifiers are clustered.
for *each two clusters* **do**
  Select a pair of classifiers which are the most diversity among all pairwise classifiers of the two clusters.
end
Generate the final prediction of the ensemble system based on the majority voting of the selected classifiers on the testing set.

3 Experiment Results

Several experiments have been done on UCI data sets to test the validity of the proposed method. The data sets used are shown in Table 1.

Table 1. Datasets used in the experiments

Dataset	Data size	Concept	Condition attribute
Breast Cancer Wisconsin	699	2	9
Vote	435	2	16
Iris	150	3	4
Credit screening	690	2	15

Several comparative experiments are done.

The first experiment is done using the proposed method (SEFSBRS).

The second experiment uses an ensemble strategy based on all the classifiers. It is named ensemble all in this paper. It includes the following steps. All the possible classifiers are created based on all reductions generated from the dicernibility matrix at first. Then, final prediction is obtained based on all the classifiers.

The third experiment is based on the feature selection algorithm of MIBARK [19]. A reduct is generated using the MIBARK algorithm. Then, a classifier is constructed according to the reduct. The final prediction is gotten using the single classifier only.

The forth experiment is based on the feature selection algorithm proposed in [20], the detailed experiment process is similar to the third one.

The fifth experiment is based on SVM, and the detailed experiment process is similar to the third one too.

Each dataset is divided into a training set, a validation set and a testing set for all the experiments. Each set contains 60%, 20%, and 20% of the total data set respectively. The 5- fold validation method is carried out for each dataset. The correct recognition rates of each method for these datasets are shown in Table 2.

From the experiment results, we can find that the proposed method (SEFSBRS) is valid. It can get high recognition rate. By comparing SEFSBRS and Ensemble all, we can find that the selective ensemble is almost as accurate as

Table 2. Experiment results

Dataset	SEFSBRS	Ensemble all	MIBARK	Feature selection	SVM
Breast Cancer Wisconsin	96.37%	96.23%	70.97%	87.23%	86.40%
Vote	94.10%	94.00%	91.11%	78.98%	87.14%
Iris	84.13%	86.44%	72.37%	50.83%	94%
Credit screening	68.60%	68.87%	41.63%	19.69%	56%
Average	85.80%	86.39%	69.02%	59.18%	80.89%

the ensemble using all possible candidate classifiers. Sometimes, it can get higher accuracy. However, SEFSBRS often use less classifiers than Ensemble all. So, it will be more efficient in real application. By comparing SEFSBRS, MIBARK, and Feature selection, we can find that the ensemble is more accurate. It proves that the ensemble of base classifiers is more effective than a single classifier. We can also find that the ensemble method even over performs SVM in most cases. Thus, the method can be taken as a useful method in machine leaning, pattern classification, etc.

4 Conclusion and Future Works

In this paper, a selective ensemble feature selection method based on rough set theory is proposed. All candidate classifiers for ensemble are produced based on the disernibiltiy matrix. The classification ability of each base classifiers is guaranteed. For the purpose of selecting uncorrelated base classifiers for ensemble, cluster method is used. It can ensure more diversity on the selective classifiers. Experiment results show its validity.

In the future, the proposed method will be used in real pattern recognition problems, such as emotion recognition. At the same time, improvement of the proposed method should be discussed too.

References

1. Pawlak, Z.: Rough sets. *International J. Comp. Inform. Science*. 11 (1982) 341-356.
2. Pawlak, Z.: Rough Classification. *International Journal of Man-Machine Studies*. 5 (1984) 469-483.
3. Wang, G.Y., Wu, Y., Fisher, P.S.: Rule Generation Based on Rough Set Theory. In: Proceedings of SPIE, Washington (2000) 181-189.
4. Skowron, A., Polkowski, L.: Decision Algorithms: A Survey of Rough Set - Theoretic Methods. *Fundamenta Informaticae*. 3-4 (1997) 345-358.
5. Zhong, N., Dong, J.Z., Ohsuga, S.: Using Rough sets with Heuristics for Feature Selection. *Journal of Intelligent Information Systems*. 3 (2001) 199-214.
6. Ditterrich, T.G.: Machine learning research: four current direction. *Artificial Intelligence Magzine*. 4 (1997) 97-136.
7. Ditterrich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F., (Eds.), *Multiple Classifier Systems*. LNCS 1857, Springer, Berlin (2001) 1-15.
8. Tsymbal, A., Pechenizkiy, M., Cunningham, P.: Diversity in search strategies for ensemble feature selection. *Information Fusion*. 1 (2005) 83-98.
9. Breiman, L.: Bagging predictors. *Machine Learning*. 2 (1996) 123-140.
10. Freund, Y.: Boosting a weak algorithm by majority. *Information and Computation*. 2 (1995) 256-285.
11. Ditterrich, T.G., Bakiri, G.: Solving multi-class learning problem via error-correcting output codes. *Journal of Artificial Intelligence Research*. 2 (1995) 263-286.
12. Ditterrich, T.G.: An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*. (1998).

13. Opitz, D.: Feature selection for ensembles. In: Proceedings of 16th National Conference on Artificial Intelligence, AAAI Press, Florida (1999) 379-384.
14. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*. 2(2003) 181-207.
15. Kuncheva, L.I.: That elusive diversity in classifier ensembles. In: Proceedings of First Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA), Mallorca (2003) 1126-1138.
16. Brow, G., Wyatt, J., Harris, R., Yao, X.: Diversity Creation Methods: A Survey and Categorisation. *Journal of Information Fusion*. 1 (2005) 1-28.
17. Giacinto, G., Roli, F.: Design of effective neural network ensembles for image classification processes. *Journal of Image Vision and Computing*. 9 (2001) 699-707.
18. Shipp, C.A., Kuncheva, L.I.: Relationships between combination methods and measures of diversity in combining classifiers. *Journal of Information Fusion*. 2 (2002) 135-148.
19. Miao, D.Q., Hu, G.R.: A Heuristic Algorithm for Reduction of Knowledge. *Journal of Computer Research and Development*. 6 (1999) 681-684 (in Chinese).
20. Hu, X.H., Cercone, N.: Learning Maximal Generalized Decision Rules via Discretization, Generalization and Rough set Feature Selection. In: Proceedings of 9th International Conference on Tools with Artificial Intelligence (ICTAI '97), California (1997) 548-556.