Support Vector Machine Based Ensemble Classifier

Zhonghui Hu, Yunze Cai, Ye Li and Xiaoming Xu

Abstract—**The strategy that the original input space is partitioned into several input subspaces usually works for improving the performance. Different from conventional partition methods, the partition method, attribute reduction based on rough sets theory, allows the input subspaces partially overlapped. These input subspaces can offer complementary information about hidden data patterns. In every subspace, a SVM sub-classifier is learned. Then, those SVM sub-classifiers with good performance are selected and combined to construct an ensemble classifier. The proposed method is applied to decision-making of medical diagnosis. Comparison between our method and several other popular ensemble methods is done. Experimental results demonstrate that the proposed approach can make full use of the information contained in data and improve the decision-making performance.**

I. INTRODUCTION

FOR medical diagnostic decision-making problem, a great deal of information from many data sources is collected. Some popular learning methods can be applied in a large input space consisting of all the information. However, a mass of information in one input space may not improve the performance, and even worse affect the performance. Reasonable input space partition may improve the performance. This is demonstrated with the experimental results in this paper. Furthermore, the contributions of different information are different. Sometimes there exists redundant information, which has no contribution to find the data patterns. There usually also exists minor information with little contribution. In general, without the cost of decreasing accuracy, the efficiency can be improved by deleting the minor or redundant information.

For improving the performance, the input space consisting of all the information is partitioned into several input subspaces. Different from conventional partition methods, the partition method proposed in this paper allows the input subspaces partially overlapped. These input subspaces can offer complementary information about hidden data patterns. In every subspace, a sub-classifier can be learned. The performance can be improved by combining these learned sub-classifiers. For the synergistic use of information from these sub-classifiers, the information fusion techniques are considered for obtaining a global decision. Our proposed approach can make full use of the information from many data sources and improve the decision-making performance.

The input subspace can be obtained by using information reduction algorithm to delete the redundant or minor information. Usually more than one input subspace can be obtained. Generally, the learned sub-classifier with the best performance is selected as the global classifier, and the rest sub-classifiers are abandoned. However, in our proposed method, all the learned sub-classifiers with better performance than that of the learned classifier in global input space are selected, and then an ensemble classifier is constructed by combining these sub-classifiers. Our proposed method is used in the situation that at least more than one subspace are obtained. The improved algorithm of information reduction proposed by [1] is a typical one for obtaining exhaustive information subspaces.

Other methods for producing ensemble classifiers, such as Bagging and Boosting, have been investigated and shown to be helpful for improving the performance [2]. The multiple support vector machine decision model (MSDM), using the majority-vote scheme, is proposed in [3] to improve the classification performance.

The support vector machine (SVM) proposed by Vapnik [4] is a new kind of learning machine, which derives from statistical learning theory and VC-dimension theory [5] and has become another research hotspot. Based on the structural risk minimization principle, SVM has good generalization performance. As compared to other popular learning methods, SVM is more appropriate for small sample problems. Therefore, it is used as the basic classifier for comparison in this paper.

The rest of this paper is organized as follows. In Section II the basic theory of SVM for classification is introduced. The

Manuscript received September 12, 2004. This work was supported by the National Key Fundamental Research Program of China (2002cb312200), the National High Technology Research and Development Program of China (2002AA412010), and in part supported by the Natural Science Foundation of China (60174038).

Zhonghui Hu is with the Department of Automation, Shanghai Jiaotong Universtiy, Shanghai 200030, P. R. China. (Tel/Fax: +86.21.62826946; e-mail: huhzh@sjtu.edu.cn).

Yunze Cai is with the Department of Automation, Shanghai Jiaotong Universtiy, Shanghai 200030, P. R. China. (e-mail: yzcai@sjtu.edu.cn).

Ye Li is with the Department of Automation, Shanghai Jiaotong Universtiy, Shanghai 200030, P. R. China. (e-mail: liyemail@ sjtu.edu.cn).

Xiaoming Xu is with the Department of Automation, Shanghai Jiaotong Universtiy, Shanghai 200030, P. R. China. (e-mail: xmxu@sjtu.edu.cn).

0-7803-9098-9/05/$25.00 ©2005 AACC

attribute discretization methods are discussed in Section III. The attribute reduction method based on rough set theory is presented in Section IV. The decision fusion theory applied to ensemble classifier is described in Section V. The experiment results are given and analyzed in Section VI. Finally, in Section VII our conclusions are presented.

II. Support Vector Machines for Classification

For the training data set $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{l} \in R^n \times \left\{+1,-1\right\}$, where x_i represents condition attribute and y_i represents class attribute. According to basic theory of SVM for nonlinear classification, the original data are projected into a certain high dimensional Euclidean space *H* by a nonlinear map $\Phi : R^n \to H$, so that the problem of nonlinear classification are transformed into that of linear classification in the space *H*. Introducing the kernel function $K\left(x_i, x_j\right) = \left\langle \Phi\left(x_i\right), \Phi\left(x_j\right)\right\rangle$ makes it not necessary to know $\Phi(\bullet)$ [6]. The optimization problem of nonlinear classification can be described by

$$\begin{aligned} &\min \ J\left(W,\xi\right) = \frac{1}{2}\left\|W\right\|^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t. \ y_i\left[W \cdot \Phi\left(x\right) + b\right] \geq 1 - \xi_i, \\ &\qquad \xi_i \geq 0, \quad i = 1,2,\cdots,l \end{aligned} \tag{1}$$

Based on the Lagrange method and duality principle, the problem (1) can be rewritten as

$$\begin{aligned} &\max \ M\left(\alpha\right) = -\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K\left(x_i, x_j\right) + \sum_{i=1}^{l}\alpha_i \\ &s.t. \quad \sum_{i=1}^{l}\alpha_i y_i = 0, \\ &\qquad \alpha_{\mathrm{i}} \in \left[0, C\right], \ i = 1,2,\cdots,l \end{aligned} \tag{2}$$

Solving the problem (2), we can get the optimal hyperplane

$$f\left(x\right) = \sum_{sv}\alpha_i y_i K\left(x, x_i\right) + b = 0 \tag{3}$$

Therefore, the decision function based on SVM for nonlinear classification in the input space is

$$d\left(x\right) = \operatorname{sgn}\left[\sum_{sv} y_i\alpha_i K\left(x_i, x\right) + b\right] \tag{4}$$

III. Attribute Reduction

Different from conventional partition methods, the partition method used in this paper allows the input subspaces partially overlapped, and the original input space not covered entirely, because they are obtained by using attribute reduction algorithm to delete the redundant or minor attribute. The comparison of two partition strategies is illustrated in Figure 1.

At least more than one input subspace where a sub-classifier is learned should be obtained in our proposed ensemble method. The modified rough sets based attribute reduction method proposed in [1] can obtain exhaustive optimal feature subsets by efficiently searching the original input attribute set. The attributes containing more classification information are selected by using feature matrix to compose the reduced attribute sets, namely, the optimal feature subsets, which construct the input subspaces respectively.

Some important basic concepts in rough sets theory are given in the following [7].

Definition 1: The optimal feature subset is the minimal subset composed of attributes, which can differentiate the instances in a positive subset from that in a negative subset.

Definition 2: The discernibility matrix of information system is defined as

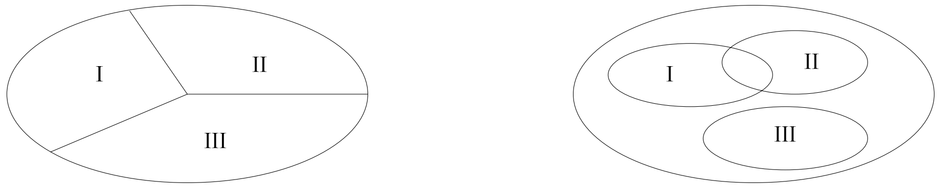
$$\begin{aligned} &M_D = \left(c_{ij}\right)_{n\times n}, \ c_{ij} = \left\{a \in A : a(u_j) \neq a(u_i) \wedge \right. \\ &\left. d(u_i) \neq d(u_j) \wedge u_i, u_j \in U \wedge i,j = 1,2,\cdots,n\right\} \end{aligned} \tag{5}$$

where c_{ij} is the element of M_D, $a \in A$ is the conditional attribute, and d is the class attribute.

Definition 3: The feature matrix of positive set *Pe* opposite to negative set *Ne* is defined by

$$\begin{aligned} &M = \left(m_{ij}\right)_{p\times q}, \\ &m_{ij} = \left\{v_k \mid v_k\left(e^+\right) \neq v_k\left(e^-\right), k = 1,2,\cdots,n\right\}, \\ &e^+ \in P_e, \ e^- \in N_e \end{aligned} \tag{6}$$

where $v_k(\bullet)$ is the value of attribute k, p is the size of P_e and q is the size of N_e. The elements of matrix *M* consist of the feature subsets that can distinguish the positive instance and the negative instance. One important property of the



(a) The conventional partition method

(b) The partition method in this paper

Fig. 1. Comparison of two partition strategies

feature matrix *M* is as below.

Property 1: If no less than one attribute in attribute set $g = (v_1, v_2, \cdots, v_s)(s<n)$ exists in every element of the feature matrix *M*, the attribute set is a reduction set of feature set for decision table. When s is taken the minimal value, the obtained feature subset is an optimal subset.

The procedure of searching optimal feature subsets is described as follows [1].

Suppose the feature set of decision table is A and the initial set of optimal feature subset is $O = \varnothing$ (empty set).

Step 1: Calculate the discernibility matrix M_D and obtain the core set H .

Step 2: Calculate the feature matrix of decision table. If the core set H satisfies the Property 1, $O \leftarrow H$ and go to Step 8; otherwise, go to step 3.

Step 3: Set $B = A - H$. Suppose the i order power set of B is B_i , $i = 1, 2, \cdots, |B|$ (The base of B).

Step 4: $i = 1$, and $Flag = 0$.

Step 5: Randomly select feature (or set) $a \in B_i$. Let $F = H \bigcup a$. If F satisfies Property 1, $O \leftarrow F$ and $Flag = 1$.

Step 6: $B_i = B_i - a$. If $B_i \neq \varnothing$, go to Step 5; otherwise, go to Step 7.

Step 7: If $Flag = 1$, go to Step 8; otherwise, if $i < |B|$, $i = i + 1$ and go to Step 5.

Step 8: Export the optimal feature subset *O*.

Finally, the obtained reduced attribute subsets, i.e., the optimal feature subsets, are as follows

$$OS_i = \left\{ a_j \in A : j = 1, 2, \cdots, k_j \right\},\ i = 1, 2, \cdots, m \qquad (7)$$

where m is the number of obtained optimal feature subsets.

The precondition of applying the above algorithm is that all the attributes are discrete. Therefore, the continuous attributes have to be discretized. The principle of attribute discretization is described as below [7].

Consider a decision-making system $S = (U, A, d)$, where the data set is $U = \{x_q\}_{q=1}^{n}$, the condition attribute set is $A = \{a_1, \cdots, a_k\}$, and the class attribute set is $d : U \rightarrow \{1, \cdots, r\}$. The set of truncation points are defined as $I_a = \{C_i^a \mid C_i^a \in R,\ a \in A,\ 0 \leq i \leq k_a\}$, where C_i^a is the truncation point to divide the interval $[C_l^a, C_u^a]$. Correspondingly, $I'_a = \{[C_0^a, C_1^a), [C_1^a, C_2^a), \cdots, [C_{k_a}^a, C_{k_a+1}^a)\}$ is a subsection set in the interval $[C_l^a, C_u^a]$. Let $P = \bigcup_{a \in A} I'_a$. Then we can define a new discrete decision system $S^p = (U, A^p, d)$, where $A^P = \{a^P \mid a \in A, a^p(x) = i\}$ if and only if $a(x) \in [C_i^a, C_{i+1}^a)$, $x \in U$, $0 \leq i \leq k_a$. The truncation points can be obtained by using various attribute discretization methods.

The simplest method to discretize continuous attributes is equal width interval, which divides the range of observed values into *k* equal sized bins. *k* is a user-defined parameter [8]. In this paper, this approach is used, for its low computational complexity. Empirical rule about how to determine the parameter *k* is not available. The parameter *k* is heuristically determined based on the interval number of the other discrete attributes in the feature space.

IV. Decision Fusion

In every subspace corresponding to each optimal feature set, a SVM sub-classifier can be trained. For improving the performance, these learned SVM sub-classifiers can be combined to obtain a global decision using information fusion techniques. In practice, if we have no a prior knowledge, the appropriate number of sub-classifiers have to be probed [9]. In this paper, the SVM sub-classifiers that work better than or equal to the learned SVM classifier using all features in original input space are combined to construct a SVM ensemble classifier. The scheme is shown in Figure 2. The majority vote is used for combining the individual decisions from selected SVM sub-classifiers. For an unknown instance x , suppose there are m SVM sub-classifiers and the decision of the *i*th SVM sub-classifier is $d_i(x)$, the global decision will be [3]

$$d(x) = \arg\ \max \left\{ \sum_{d_i(x)=1} d_i(x),\ \sum_{d_i(x)=-1} \left| d_i(x) \right| \right\} \qquad (9)$$

where $i = 1, \cdots, m$.

However, if the number of selected SVM sub-classifiers is even, it is possible that the votes of two categories are equal. The expression (9) does not work in this situation. Therefore, we set (9) as the main rule. When it does not work, we apply the following complementary rule for decision fusion.

$$d(x) = \arg\ \max \left\{ \sum_{f_i(x) \geq 0} f_i(x),\ \sum_{f_i(x)<0} \left| f_i(x) \right| \right\} \qquad (10)$$

where $f_i(x)$ is the distance from instance x to the optimal classification hyperplane of the *i*th SVM sub-classifier, $i = 1, \cdots, m$, and m is the total number of SVM sub-classifiers comprising the SVM ensemble classifier.

V. Several Other Ensemble Methods

The multiple support vector machine decision model (MSDM) is proposed in [3] to improve the robustness and reliability of SVM classifiers. The original training set is randomly partitioned into several training subsets with the same size. The individual SVM classifiers are trained respectively using these training subsets. Then the majority-vote scheme is used to combine all the learned

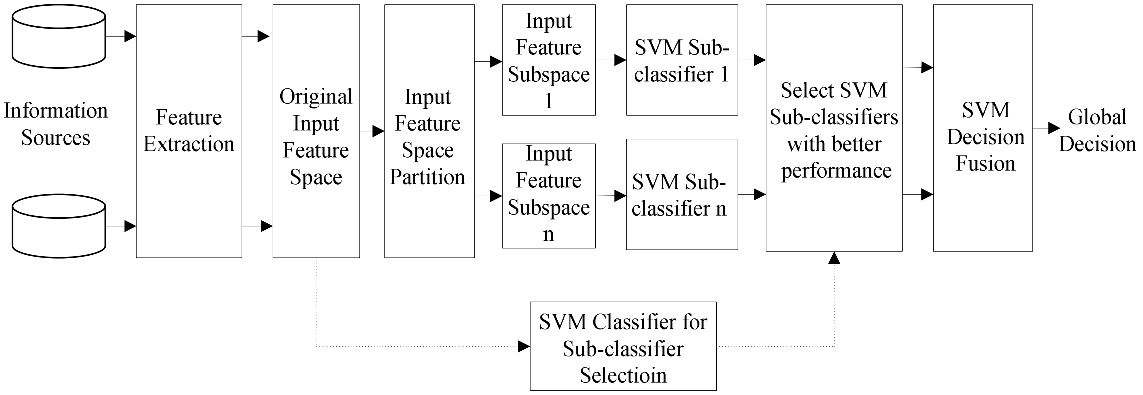


Fig. 2. SVM ensemble scheme

SVM classifiers. Suppose that there are *m* individual learned SVM classifiers and the output of the *i*th SVM classifier is O_i, the final output will be

$$O_d = sign(\sum_{i=1}^{m} O_i) \tag{11}$$

Bagging is a method for generating multiple versions of a classifier and using these to get an ensemble classifier [10]. The multiple versions are formed by making bootstrap replication of the learning set and using these as new learning sets. Given an original training dataset S of size N, a training dataset S_t ($t = 1, 2, \cdots, T$) is sampled with replacement from the original dataset S [11]. Then a SVM classifier C_t is learned for each training subset S_t. To classify an unknown sample x, each SVM classifier C_t returns its class prediction, which counts as one vote. The bagged SVM classifier $C*$ counts the votes and assigns the class with the most votes to X. That is

$$C^*(x) = sign\left(\sum_{t=1}^{T} \mathrm{C_t}(x) \right) \tag{12}$$

Boosting encompasses a family of methods [12]. The focus of these methods is to produce a series of classifiers. The training set is chosen based on the performance of the earlier classifier in the series. The subsequent classifiers will pay more attention to the samples misclassified by the earlier classifier, which are more weighted. However, some learning algorithms, such as SVMs, do not allow training using weighted samples. In this case, sampling with replacement can be used. Samples that are incorrectly predicted by previous classifiers in the series are chosen more often than samples that were correctly predicted. The AdaBoost algorithm using basic SVM classifier is shown in Figure 3.

VI. Experimental Results

Two data sets, the Cleveland heart disease database (CHDD) provided by Robert Detrano, and the breast cancer database (BCD) donated by Olvi Mangasarian in UCI machine learning datasets, are used. 10-fold cross validation is applied for obtaining the average performance.

The CHDD has total 303 records, among which 297 records have no absent attributes. Only 13 attributes are used. The status of heart disease has two categories: presence (1, 2, 3, 4) and absence (0) [13]. The attributes with continuous

```
Input: training set S of size $N$, class predictor $CP_t$, interger $T$ (number of iteration)
  Set the probability of picking each sample to be $1/N$
  For $t$ = 1 to $T$
    Training set $S_t$ is obtained by using updated probabilities and sampling with replacement
    Training SVM classifier $CP_t$ with training set $S_t$
    Let $\varepsilon_t$ be the sum of the probability of the misclassified instances for the currently trained classifier
                $\beta_t = (1 - \varepsilon_t)/\varepsilon_t$                (13)
    Multiply the probabilities of incorrectly classified instances by the factor $\beta_t$ and then renormalize them.
  End
  Output:       $C^*(x) = sign\left(\sum_{t=1}^{T} \log(\beta_t) O_{CP_t}\right)$                (14)
```

Fig. 3. The AdaBoost algorithm

TABLE I
COMPARISON OF SEVERAL SVM CLASSIFIERS (%)

Classifier	Training accuracy	Testing accuracy
Single	86.50	82.47
MSDM	88.33	83.50
Bagging	87.50	83.50
Boosting	88.00	82.47
Our method	90.50	83.91

TABLE II
COMPARISON OF DIFFERENT SVM CLASSIFIERS (%)

Classifier	Training accuracy	Testing accuracy
Single	86.50	82.47
MSDM	88.33	83.50
Bagging	87.50	83.50
Boosting	88.00	82.47
Our method	90.50	83.91

values are discretized by using the method of equal width interval before attribute reduction. The exhaustive optimal feature subsets, i.e., 9 partially overlapped input subspaces of 8 dimensions are obtained by using the search algorithm of optimal feature subsets.

The criterion of selecting SVM sub-classifiers is that the testing performance of SVM sub-classifier should be better than that of SVM trained by using all the attributes in the original input space (Table I: Single SVM). The experiment shows that the performance of some SVM sub-classifiers is better than that of the single SVM classifier. Therefore, these SVM sub-classifiers are selected to construct an SVM ensemble classifier. Compared with that of several other methods, the performance of proposed method is also satisfactory, which provides another optional approach to improve the classification accuracy.

Another data set used is the breast cancer database [14]. The database has total 699 records, among which 16 instances have incomplete attributes. All the nine attributes are discrete. Eighteen optimal feature subsets of 4 dimensions are obtained. The SVM sub-classifiers with better performance are used to create a SVM ensemble classifier. The comparison of performance among several SVM methods is given in Table II. The proposed method has a little better performance than the other SVM classifiers evaluated.

VII. CONCLUSIONS

The ensemble method based on SVMs is proposed in this paper. The strategy of partitioning the original input space into several input subspaces usually works for improving the performance. The method of rough sets based attribute reduction allows the input subspaces partially overlapped. These input subspaces can offer complementary information about hidden data patterns. Each sub-classifier is learned corresponding to the input subspace. Those SVM sub-classifiers with good performance are selected and combined to construct an ensemble classifier.

The proposed method is applied to decision-making of medical diagnosis. The experimental results demonstrate that our proposed approach can make full use of the information contained in data and improve the decision-making performance. When a basic classifier with the problem of curse of dimensionality is used, the proposed method can reduce the space dimensionality and overcome the problem.

REFERENCES

[1] T. Pan, W. D. Zhao and Z H. H. Sheng, "Optimal Feature Subset Selection for Decision Tables - a Heuristic Algorithm Based on Rough Set," *Journal of Southeast University*, Vol. 30, no. 5, pp. 118-122, 2000.

[2] E. Bauer and R. Kohavi, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants," *Machine Learning*, Vol. 38, 1998.

[3] W. W. Yan, , Z. G. Chen and H. H. Shao, "Multi Support Vector Machines Decision Model and its Application," *Journal of Shanghai Jiaotong University*, Vol. E-7, no. 2, pp. 220-222, 2002.

[4] V. N. Vapnik, *Statistical Learning Theory*, New York: Wiley, 1998.

[5] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, 2000.

[6] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, Vol. 2, no. 2, pp. 121-167, 1998.

[7] Q. Liu, *Rough Set and Reasoning*, Science Press, Bei Jing, 2001. (In Chinese)

[8] J. Dougherty, R. Kohavi and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features," *In Armand Prieditis & Stuart Russell, eds., Machine Learning: Proceedings of 12th International Conference*, Mrogan Kaufmann Publishers, San Francisco, 1995.

[9] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd Edition, John Wiley & Sons, Inc., 2001.

[10] L. Breiman, "Bagging Predictors," *Technical Report No. 421*, Department of Statistics University of California, Berkeley, California, 1994.

[11] H. Jiawei and M. Kamber, *Data Mining-Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[12] D. Opitz and R.Maxlin, "Popular Ensemble Methods: An Empirical Study," *Journal of Artificial Intelligence Research*, Vol. 11, pp. 169-198, 1999.

[13] L. P. Proben1, "A Set of neural benchmark problem and benchmark rules," *Fakulitat fur informatik, Technical Report 21/94*, Univ. Karlsruhe, Germany, 1994.

[14] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," *SIAM News*, Vol. 23, no 5, pp. 1-18, 1990.