An Improved SVM Classifier for Medical Image Classification*

Yun Jiang[1,2], Zhanhuai Li[1], Longbo Zhang[1,3], and Peng Sun[1]

[1] College of Computer Science, Northwestern Polytechnical University, 710072, Xi'an, P.R. China
[2] College of Mathematics and Information Science, Northwest Normal University, 730070, Lanzhou, P.R. China
[3] School of Computer Science, Shandong University of Technology, Zibo 255049, China

Abstract. Support Vector Machine (SVM) has high classifying accuracy and good capabilities of fault-tolerance and generalization. The Rough Set Theory (RST) approach has the advantages on dealing with a large amount of data and eliminating redundant information. In this paper, we join SVM classifier with RST which we call the Improved Support Vector Machine (ISVM) to classify digital mammography. The experimental results show that this ISVM classifier can get 96.56% accuracy which is higher about 3.42% than 92.94% using SVM, and the error recognition rates are close to 100% averagely.

1 Introduction

Support vector machine (SVM) is a proven success and a state-of-the-art method in many areas, and a promising machine learning technique proposed by Vapnik and his group AT Bell Laboratories[1]. It is based on VC dimensional theory and statistical learning theory. For many practical problems, including pattern matching and classification[2][3], function approximation[4], data clustering and forecasting[5][6], support vector machine has drawn much attention and been applied successfully in recent years because of its greater generalization performance. An interesting property of SVM is that it is an approximate implementation of the structural risk minimization induction principle that aims at minimizing a bound on the generalization error of a model, rather than minimizing the mean square error over the data set[7]. SVM is considered as a good learning method that can overcome the internal drawbacks of neural networks[8]. But there exists a drawback which can not distinguish the importance of training sample attributes. Furthermore, it will take up more storage space when there are a large number of sample attributes. Although SVM has strong capabilities of recognizing patterns and good capabilities of fault-tolerance and generalization, SVM cannot reduce the input data and select the most important information.

* This paper is supported by National Science Foundation No. 60573096 and Gansu province Science Foundation of China No. 3ZS 051-A25-042.

M. Kryszkiewicz et al. (Eds.): RSEISP 2007, LNAI 4585, pp. 764–773, 2007.

Several techniques aim to reduce the prediction complexity of SVM by expressing the SVM solution with a smaller kernel expansion. Since one must compute the SVM solution before applying these post-processing techniques, they are not suitable for reducing the complexity of the training stage[9].

Rough Set Theory(RST), introduced by Pawlak[10] in his seminal paper of 1982, is a new mathematical approach to uncertain and vague data analysis and is also a new fundamental theory of soft computing [11]. In recent years, RST becomes an attractive and promising issue. RST can mine useful information from a large amount of data, generate decision rules without prior knowledge, and eliminate redundant information. It is used generally in many fields[12], such as knowledge discovery, machine learning, pattern recognition and data mining. In this paper, a new classification algorithm based on SVM and RST is proposed, which we call Improved Support Vector Machine (ISVM). ISVM inherits the merits of both SVM and RST. We apply ISVM to medical images classify. It is tested on real datasets MIAS[13](the Mammographic Image Analysis Society)and can get 96.56% accuracy which is higher about 3.42% than 92.94% using SVM, and the error recognition rates are close to 100%averagely.

The rest of the paper is organized as follows: Section 2 describes the theory of SVM. Section 3 presents rough set theory, the reduction algorithm and the Improved SVM algorithm–ISVM. In section 4, data pre-processing and feature extraction are introduced. In section 5, we present our experiments and results. Finally, in section 6, we show our conclusions and future work.

2 Support Vector Machine(SVM)[1]

Consider the problem of separable training vectors belonging to two separate classes,

$$G = \{(x_i, y_i)\}_{i=1}^{l}, \quad x_i \in R^n, \ y_i \in \{-1, 1\}, \ i = 1, \cdots, l \tag{1}$$

We should find a linear function,

$$y = f(x) = \omega\varphi(x) + b \tag{2}$$

That is to say, we should make the margin between the two classes points as possible as big, it is equal to minimize $\frac{1}{2}\|\omega\|^2$, so the optimal classification problem is transformed into a convex quadratic programming problem:

$$\min \quad \frac{1}{2}\|\omega\|^2 \quad s.t. \quad y_i((\omega \cdot x_i) + b) \geq 1, i = 1, \cdots, l \tag{3}$$

when the training points are non-linearly separable, (3) should be transformed into (4).

$$\min \quad \frac{1}{2}\|\omega\|^2 + c\Sigma_{i=1}^{l}\xi_i \quad s.t. \quad y_i((\omega \cdot x_i) + b) \geq 1 - \xi_i, i = 1, \cdots, l \tag{4}$$

The solution to the above optimization problem of equation (4) is transformed into the dual problem (5) by the saddle point of the Lagrange functional,

$$\begin{aligned} \max \quad & \Sigma_{i=1}^{l}\alpha_i - \frac{1}{2}\Sigma_{i=1}^{l}\Sigma_{j=1}^{l}\alpha_i\alpha_j y_i y_j k(x_i, x_j) \\ s.t. \quad & \Sigma_{i=1}^{l}\alpha_i y_i = 0, \quad 0 \le \alpha_i \le c, \ i = 1, \cdots, l \end{aligned} \tag{5}$$

We can get the decision function:

$$f(x) = \Sigma_{i=1}^{l}\alpha_i y_i K(x_i, x) + b \tag{6}$$

kernel function $K(x_i, x) = (\Phi(x_i) \cdot \Phi(x))$ is a symmetric function satisfying Mercer's condition, when given the sample sets are not separate in the primal space, we can be used to map the data with mapping Φ into a high dimensional feature space where linear classification is performed.

There are three parameters in SVM model that we should choose, they make great impact on model's generalization ability. It is well known that SVM generalization performance depends on a good setting of hyperparameters C, the kernel function and kernel parameter. Moreover, kernel function and kernel parameter's selection connects with feature selection in SVM, so feature selection is very important.

3 The Improved Support Vector Machine Algorithm (ISVM)

In this section, the theory of rough sets has been used in the first stage to reduce the original feature sets. In the second stage the SVM algorithm has been executed with the reduced feature sets. The reduction of original feature sets results in a smaller structure and quicker learning of the SVM and as a whole the hybrid algorithm provides better performance than the SVM algorithm from individual paradigm. The following are the basic concepts of the rough set theory, the algorithms of reduction and the improved SVM (ISVM).

3.1 Rough Set Theory

The original Rough Set Theory was proposed by Pawlak [10][14]. This theory is concerned with analysis of deterministic data dependencies.

Information Systems. In the Rough Set Theory, information systems are used to represent knowledge. An information system $S = (U, A, V, f)$ consists of U which is a nonempty, finite set named universe, which is a set of objects, $U = \{x_1, x_2, \cdots, x_m\}$; A is a nonempty, finite set of attributes, $A = C \cup D$, in which C is the set of condition attributes, and D is the set of decision attributes; $V = \bigcup_{a \in A} Va$ is the domain of a; $f : U \times A \to V$ is an information function. For each $a \in A$ and $x \in U$, an information function $f(x, a) \in Va$ is defined, which means that for each object x in U, f specify its attribute value.

Lower and Upper Approximation. Due to imprecision which exists in the real world data, there are always conflicting objects contained in a decision table. Here conflicting objects refers to two or more objects that are undistinguishable by employing any set of condition attributes, but they belong to different decision classes. Such objects are called inconsistent. Such a decision table is called inconsistent decision table. In the rough set theory, the approximations of sets are introduced to deal with inconsistency. If $S = (U, A, V, f)$ is a decision table, suppose $B \subseteq A$, and $X \subseteq U$, then the *B-lower* and *B-upper* approximations of X are defined as:

$$\begin{aligned} \underline{B(X)} &= \bigcup \{Y \in U/IND(B) : Y \subseteq X\}, \\ \overline{B(X)} &= \bigcup \{Y \in U/IND(B) : Y \cap X \neq \emptyset\} \end{aligned} \tag{7}$$

Here, $U/IND(B)$ denotes the family of all equivalence classes of B; $IND(B) = \{(x, y) \in U \times U | \forall\ a \in B, f(x, a) = f(y, a)\}$ is the *B-indiscernibility relation.* $\underline{B(X)}$is the set of all elements of U which can be certainty classified as elements of $\overline{X}$, employing the set of attributes B. The *Positive Region* of X is defined as:

$$POS_B(X) = \underline{B(X)} \tag{8}$$

$\overline{B(X)}$ is the set of elements of U which can be possibly classified as elements of X using the set of attributes B. The set $Bnd_B(X) = \overline{B(X)} - \underline{B(X)}$ is called the *B-boundary* of X. If $Bnd_B(X) = \emptyset$, then we say that X is definable on B; otherwise we say that X is non-definable on B, which is also named as *rough set.*

Attribute Reduction. An important issue in the Rough Set Theory is about attributes reduction. The process of finding a smaller set of attributes than original one with same classify ability as original set is called attribute reduction. *Core* is the intersection of all reductions. Given an information system S, for a given set of condition attributes $P \subseteq C$, we can define a positive region $POS_P(D) = \bigcup_{x \in U/D} \underline{P}X$, which contains all objects in U, which can be classified without error into distinct classes defined by $IND(D)$ based only on information in the $IND(P)$. Another important issue in data analysis is discovering dependencies between attributes. Let D and C be subsets of A. D depends on C in degree as denoted in the following:

$$\gamma(C, D) = card(POS_C(D))/card(U),\ \gamma(C, D) \in [0, 1] \tag{9}$$

The set of attributes reduction is described as:

$$R = \{R : R \subseteq C, \gamma(R, D) = \gamma(C, D)\} \tag{10}$$

Thereby, the equality of the attributes dependency can be used as the end condition of iterative operation.

3.2 Attribute Reduction Algorithm

For a given decision information table $S' = (U, C \cup D, V, f)$, the subset $C' \subseteq C$ is the smallest reduction of C. If C' satisfy two conditions as below: (1).$POS_c(\gamma) = POS_{C'}(\gamma)$, (2).Not exist $C'' \subset C'$, so as to $POS_{c''}(\gamma) = POS_{C'}(\gamma)$. Based on the definition of attributes dependency, the importance of an attribute $a \in C-R$ can be defined as:

$$\theta(a, R, D) = \gamma(R \cup \{a\}, D) - \gamma(R, D) \tag{11}$$

where $R = \emptyset, \theta(a, D) = \gamma(\{a\}, D)$. Based on the hereinabove definition, we design the attributes reduction algorithm-Algorithm1.

Algorithm1: Reduce (S', R)-Attributes Reduction Algorithm.
Input: Decision information table $S' = (U, C \cup D, V, f)$
Output: An attribute reduction set R of S'
1). $R = \emptyset$;
2).For every attribute $a_i \in C-R$ calculating its attribute importance $\theta(a_i, R, D)$;
3).Choosing the attribute a_i which the value of $\theta(a_i, R, D)$ is the largest, and $R \Leftarrow R \cup \{a_i\}$;
4). If $\gamma(R, D) = \gamma(C, D)$ then goto 5) else goto 2);
5).Return (R); // Return the attribute set R which has been reduced.

Obviously,the complexity of the above algorithm is $O(m^2)$, m is the number of condition attributes in decision table S'.

3.3 The Algorithm of Improved Support Vector Machine(ISVM)

The ISVM algorithm is composed with two stages. Firstly, the condition attributes of the information set is reduced by running the reduction algorithm. Then, the reduced information set will be classified by the SVM classifier. The ISVM algorithm-Algorithm2 is as following:

Algorithm2: Improved Support Vector Machine-ISVM(S, Y)
Input: A decision information table $S = (U, C \cup D, V, f)$
Output: The classify result Y
1).Discrete(S);// Discrete the decision information table S
2).Reduce (S, R); // Running the reduction algorithm1, R is the reduced //condition attributes set.
3).$S = R \cup D$; //S is a new information table which condition attributes has //been reduced
4).SVM (S, Y);// Executing SVM classifier. Its input is the new information //table S, and Y is the classified result.
5).Return (Y);

4 Data Pre-processing and Feature Extraction

This section summarizes the mammography collection and the techniques used to enhance the mammograms as well as the features that were extracted from images.

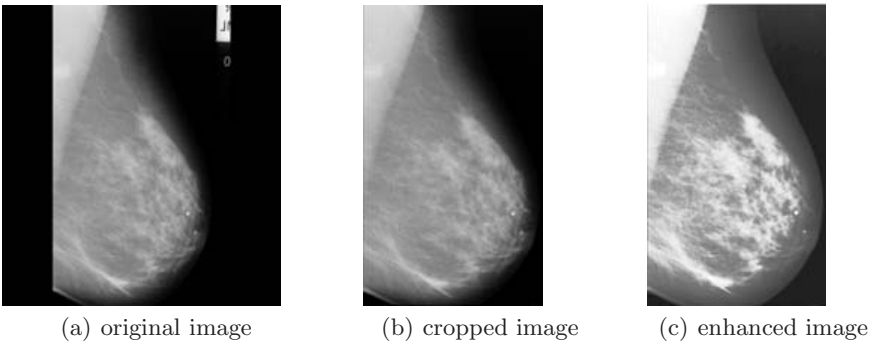
4.1 Mammography Collection

The data collection used in our experiments was taken from the Mammographic Image Anlysis Society (MIAS)[13]. We selected this dataset because it is freely available, and to be able to compare our method with other published work like [15], since it is a commonly used database for mammography categorization.

MIAS consists of 322 images, which belong to three big categories: normal, benign and malign. There are 208 normal images, 63 benign and 51 malign, which are considered abnormal. In addition, the abnormal cases are further divided in six categories: microcalcification, circumscribed masses, speculated masses, ill-defined masses, architectural distortion and asymmetry. All the images also include the locations of any abnormalities that may be present. The existing data in the collection consists of the location of the abnormality (like the center of a circle surrounding the tumor), its radius, breast position (right or left), type of breast tissues (fatty, fatty-glandular and dense) and tumor type if it exists (benign or malign). All the mammograms are medio-lateral oblique view.

4.2 Data Pre-processing

Pre-processing is always a necessity whenever the data to be mined is noisy, inconsistent or incomplete. Pre-processing significantly improves the effectiveness of data mining techniques [16].The type size of the images in MIAS is 1024x1024 and almost 50% of the whole image comprised the background with a lot of noise. In addition, these images are scanned at different illumination conditions, and therefore some images appeared too bright and some were too dark. The first step toward noise removal was pruning the images with a cropping operation.The second step was an image enhancement. Thus, we eliminated almost all the background information and most of the noise. An example of cropping that eliminates the artefacts and the black background is given in Figure 1 (a-b). Since the resulting images had different sizes, the x and the y coordinates were normalized to a value between 0 and 255. The cropping operation was done



(a) original image (b) cropped image (c) enhanced image

Fig. 1. Pre-processing phase on an example image

automatically by sweeping horizontally through the image. Then we applied the Histogram Equalization method to enhance the image in order to diminish the effect of over-brightness or over-darkness in images. Histogram Equalization increases the contrast range in an image by increasing the dynamic range of grey levels [16]. Figure 1 (c) shows an example of histogram equalization result after cropping.

4.3 Feature Extraction

After pre-processing the images, features relevant to the classification are extracted from the cleaned images. The extracted features are organized in a database, which is the input for the mining phase of the classifier. This database is also constructed by merging some already existing features like the type of the tissue (dense, fatty and fatty-glandular) and the location of the abnormality (like the center of a circle surrounding the tumor). The extracted features are four statistical parameters: mean, variance, skewness and kurtosis. The formula for the statistical parameters computed is the following: $Mean$ is $\mu = \Sigma_{k=1}^{N} f_k p_f(f_k)$; $Variance$ is $\sigma^2 = \Sigma_{k=1}^{N}(f_k - \mu)^2 p_f(f_k)$; $Skewness$ is $\mu_3 = \frac{1}{\sigma^3}\Sigma_{k=1}^{N}(f_k - \mu)^3 p_f(f_k)$; $Kurtosis$ is $\mu_4 = \frac{1}{\sigma^4}\Sigma_{k=1}^{N}(f_k - \mu)^4 p_f(f_k)$. Where N denotes the number of gray levels in the mammogram, f_k is the kth gray level and $p_f(f_k) = \frac{n_k}{n}$, where n_k is the number of pixels with f_k gray level and n is the total number of pixels in the region.

All these extracted features are computed over smaller windows of the original image. The original image is first split in four parts. For a more accurate extraction of the features we split each of these four regions in other four parts. The statistical parameters were computed for each of the sixteen sub-parts of the original image [15]. After that, we get sixty-four statistical features.

5 Experimental Results

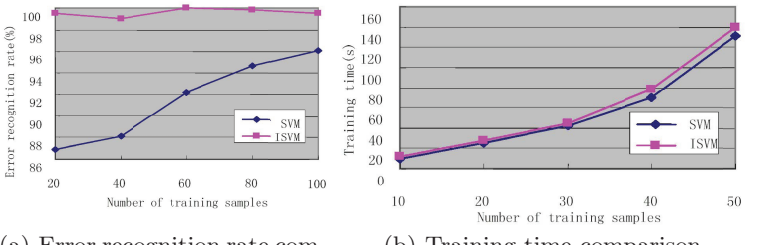
We used the 10 fold cross validation techniques to evaluate the algorithm performance. We divided the features database in ten splits. For each split we selected about 90% of the dataset for training and the rest for testing. That is 288 images in the training set and 34 images in the testing set. The features database is composed with the extracted features and the existing data of 322 images in MIAS. All the numeric attributes are discrete using algorithm DBChi2[17]. In the training phase, the ISVM was applied on the training data. Then, for an image in the testing set, the classification process searches in this ISVM for finding the class that is the closest to be attached with the object presented for categorization. At the same time, the number of choosing attributes is recorded. The SVM program is from LIBSVM[18].The experimental results is in table1.

Table 1. The comparison of SVM and ISVM algorithms on MIAS

Ten splits	SVM Accuracy(%)	ISVM Attributes number	ISVM Accuracy(%)
1	93.56	21	96.42
2	90.21	16	97.12
3	92.19	18	97.56
4	93.88	15	96.87
5	93.47	23	96.06
6	94.66	20	96.44
7	92.25	13	95.15
8	90.83	26	94.96
9	93.64	19	97.34
10	94.75	15	97.69
Average	92.94	18.6	96.56

Table 1 represents the comparison in terms of the choosing attributes number and classifying accuracy of the present algorithm ISVM and the algorithm SVM. The first column is the ten splits of MIAS. The second and the fourth columns are the classified accuracy of SVM and ISVM based on ten splits. The third volume is the number of choosing attributes of ISVM. At the bottom of the table, each column's average is shown. The table shows that the ISVM performs better than only SVM algorithm in terms of the classifying accuracy. At the same time, because the data set is reduced firstly, there are only 18.6 condition attributes averagely inputted to SVM classifier, which makes it easy for the SVM classifying.

To compare the capabilities of classification and the training time of SVM with ISVM on MIAS, we give the experiment results about error recognition rate and training time on small samples simultaneously. Figure 2(a) is the comparison of error recognition rate on different samples including 20,40,60,80 and 100 applied to train SVM and ISVM respectively. Figure 2(b) is the comparison of training time on training samples varies from 10 to 50. Figure 2(a) shows that error



(a) Error recognition rate comparison (b) Training time comparison

Fig. 2. The experimental results comparison of SVM and ISVM

recognition rate of ISVM classifier is close to 100%, which is higher than SVM classifier obviously. When the training samples are smaller than 50, the error recognition rate of SVM can not reach to 90%. From Figure 2(b), the ISVM classifier needs more time than that of SVM classifier because of the attributes reduction stage of ISVM. But the training time of them is closed. Because of calculating with second, the distinction of training time is very little and can be ignored in practice.

6 Conclusions

In this paper, we have presented a hybrid classifier based on Rough Set Theory(RST) and Support Vector Machine(SVM) which is called Improved Support Vector Machine-ISVM. ISVM makes great use of the advantages of SVM's greater generalization performance and RST in effectively dealing with vagueness and uncertainty information. By data-analyzed method of RST, it can remove large amount of redundancy, and decrease volume of SVM training data. The preprocessing step enhances the efficiency of SVM in training and testing phases and strengthens classification and generation capabilities of SVM. Finally, ISVM was applied to medical image classification, and the evaluation of the ISVM was carried out on MIAS dataset. The experimental results show that the accuracy of the ISVM classifier can reach 96.56% than 92.94% which execute SVM classifier, and the error recognition rate values of ISVM tend to 100% in more than half the splits. There are some future research directions to be studied. To cooperate with medical staff would get more interesting results. In addition, the extraction of different features or a different database organization could lead to improved results.

References

1. Vapnik, V.N.: The nature of statistical learning theory. Springer, Heidelberg (1995)
2. Osareh, A., Mirmehdil, M., Thomas, B., Markham, R.: Comparative Exudate Classification Using Support Vector Machines and Neural Networks. In: Dohi, T., Kikinis, R. (eds.) MICCAI 2002. LNCS, vol. 2489, pp. 413–420. Springer, Heidelberg (2002)
3. Foody, G.M., Mathur, A.: A Relative Evaluation of Multiclass Image Classification by Support Vector Machines. IEEE Transactions on Geoscience and Remote Sensing 42(6), 1335–1343 (2004)
4. Ma, J.S., Theiler, J., Perkins, S.: Accurate On-line Support Vector Regression. Neural Computation 15(11), 2683–2703 (2003)
5. Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V.: Support Vector Clustering. Journal of Machine Learning Research 2(2), 125–137 (2001)
6. Kim, K.J.: Financial Time Series Forecasting Using Support Vector Machines. Neuro-computing 55(1), 307–319 (2003)
7. Dibike, Y.B., Velickov, S., Solomatine, D.: Support Vector Machines: Review and Applications in Civil Engineering. In: Proc. of the 2nd Joint Workshop on Application of AI in Civil Engineering, pp. 215–218 (2000)

8. Wang, L.P. (ed.): Support Vector Machines: Theory and Application. Springer, Heidelberg (2005)
9. Schlkopf, B., Smola, A J.: Learning with Kernel-Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
10. Pawlak, Z.W.: Rough Sets. International Journal of Information and Computer Science 11(5), 341–356 (1982)
11. Lin, T.Y.: Introduction to the Special Issue on Rough Sets. International Journal of Approximate Reasoning 15(4), 287–289 (1996)
12. Wang, G.Y.: Rough Set Theory and Knowledge Acquisition. Xi'an Jiaotong University Press, Xi'an (2001)
13. (2006-9) http://www.wiau.man.ac.uk/services/MIAS/MIASweb.html
14. Pawlak, Z.W.: Rough Sets and Intelligent Data Analysis. Information sciences (147), 1–12 (2002)
15. Antonie, M.-L., Zaiane, O.R., Coman, A.: Application of data mining techniques for medical image classification. In: Proc. of Second Intl. Workshop on Multimedia Data Mining (MDM/KDD'2001) in conjunction with Seventh ACM SIGKDD, San Francisco, pp. 94–101 (2001)
16. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Addison-Wesley, Reading (1993)
17. Hu, X., Cercone, N.: Data Mining Via Generalization, Discretization and Rough Set Feature Selection [J]. Knowledge and Information System: An International Journal, vol. 1(1) (1999)
18. Chang.C., Lin, C. (2001) LIBSVM
http://www.csie.ntu.edu.tw/~cjlin/libsvm,2006-9