Constructing Classification Rules Based on SVR and Its Derivative Characteristics

Dexian Zhang[1], Zhixiao Yang[1], Yanfeng Fan[2], and Ziqiang Wang[1]

[1] College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450052, China
zdx@haut.edu.cn
[2] Computer College, Northwestern Polytecnical University, Xi'an 710072, China

Abstract. Support vector regression (SVR) is a new technique for pattern classification , function approximation and so on. In this paper we propose an new constructing approach of classification rules based on support vector regression and its derivative characteristics for the classification task of data mining. a new measure for determining the importance level of the attributes based on the trained SVR is proposed. Based on this new measure, a new approach for clas-sification rule construction using trained SVR is proposed. The performance of the new approach is demonstrated by several computing cases. The experimen-tal results prove that the approach proposed can improve the validity of the extracted classification rules remarkably compared with other constructing rule approaches, especially for the complicated classification problems.

1 Introduction

The goal of data mining is to extract knowledge from data. Data mining is an inter-disciplinary field, whose core is at the intersection of machine learning, statistics and databases. There are several data mining tasks, including classification, regression, clustering, dependence modeling, etc. Each of these tasks can be regarded as a kind of problem to be solved by a data mining approach. In classification task, the goal is to assign each case (object, record, or instance) to one class, out of a set of predefined classes, based on the values of some attributes (called predictor attributes) for the case. In this paper we propose an new constructing approach of classification rules based on support vector regression and its derivative characteristics for the classification task of data mining. The classification rule extraction has become an important aspect of data mining, since human experts and corporate managers are able to make better use of the classification rules for making decision and easily discover unknown relationships and patterns from a large data set than other expression forms of knowledge.

The existing approaches for constructing the classification rules can be roughly classified into two categories, data driven approaches and model driven approaches. The main characteristic of the data driven approaches is to extract the symbolic rules completely based on the treatment with the sample data.

R. Alhajj et al. (Eds.): ADMA 2007, LNAI 4632, pp. 300–309, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The most representative approach is the ID3 algorithm and corresponding C4.5 system proposed by J.R.Quinalan. This approach has the clear and simple theory and good ability of rules extraction, which is appropriate to deal with the problems with large amount of samples. But it still has many problems such as too much dependence on the number and distribution of samples, excessively sensitive to the noise, difficult to deal with continuous attributes effectively and etc. The main characteristic of the model driven approaches is to establish a model at first through the sample set, and then extract rules based on the relation between inputs and outputs represented by the model. Theoretically, these rule extraction approaches can overcome the shortcomings of data driven approaches mentioned above. Therefore, the model driven approaches will be the promising ones for rules extraction. The representative approaches are rules extraction approaches based on neural networks [1-8]. Though these methods have certain effectiveness for rules extraction, there exist still some problems, such as low efficiency and validity, and difficulty in dealing with continuous attributes etc.

There are two key problems required to be solved in the classification rule extraction, i.e. the attribute selection and the discretization to continuous attributes. Attribute selection is to select the best subset of attributes out of original set. The attributes that are important to maintain the concepts in the original data are selected from the entire attributes set. How to determine the importance level of attributes is the key to attribute selection. Mutual information based attribute selection [9-10] is a common method of attribute selection, in which the information content of each attribute is evaluated with regard to class labels and other attributes. By calculating mutual information, the importance levels of attributes are ranked based on their ability to maximize the evaluation formula. Another attribute selection method uses entropy measure to evaluate the relative importance of attributes [11]. The entropy measure is based on the similarities of different instances without considering the class labels. In paper [12], the separability-correlation measure is proposed for determining the importance of the original attributes. The measure includes two parts, the intra-class distance to inter-class distance ratio and an attributes-class correlation measure. Through attributes-class correlation measure, the correlation between the changes in attributes and their corresponding changes in class labels are taken into account when ranking the importance of attributes. The attribute selection methods mentioned above can be classified into the sample driven method. Their performance depends on the numbers and distributions of samples heavily. It is also difficult to use them to deal with continuous attributes. Therefore, it is still required to find more effective heuristic information for the attribute selection and the discretization to continuous attribute in the classification rule extraction.

In this paper, we use trained SVR to obtain the position and shape characteristics of the classification hypersurface. Based on the analysis of the relations among the position and shape characteristics of classification hypersurface, the partial derivative distribution of the outputs of trained SVR to its corresponding inputs and the importance level of attributes to classifications, this paper mainly

studies on the measure method of the classification power of attributes on the basis of differential information of the trained SVR and develops new approach for the rule extraction.

The rest of our paper is organized as follows. Section 2 discusses the representation of the classification rules . Section 3 describes the classifier construction based on SVR. Section 4 presents the measure method for attribute importance ranking. The rules extraction method is presented in section 5. Experimental results and analysis are reported in Section 6. Finally, we give the conclusion in Section 7.

2 Representation of the Classification Rules

Classification rules should be not only accurate but also comprehensible for the user. Comprehensibility is important whenever classification rules will be used for supporting a decision made by a human user. After all, if classification rules is not comprehensible for the user, the user will not be able to interpret and validate it. In this case, probably the user will not trust enough the classification rules to use it for decision making. This can lead to wrong decisions.

In this paper, the classification rule is often expressed in the form of IF-THEN rules which is commonly used, as follows: IF <conditions > THEN < class>.The rule antecedent (IF part) contains a set of conditions, connected by a logical conjunction operator (AND). In this paper we will refer to each rule condition as a term, so that the rule antecedent is a logical conjunction of terms in the form: IF term 1 AND term 2 AND ... Each term has two kind of forms. One kind of the form is a triple <attribute, operator, value>. The operator can be $<$,$\geq$ or $=$. Another kind of the form is a triple <attribute, $\in$, value range>. The rule consequent (THEN part) specifies the class label predicted for cases whose attributes satisfy all the terms specified in the rule antecedent.

This kind of classification rule representation has the advantage of being intuitively comprehensible for the user, as long as the number of discovered rules and the number of terms in rule antecedents are not large.

3 Constructing the Classifier Based on SVR

SVR is a new technique for pattern classification , function approximation and so on. The SVR classifiers have the advantage that we can use them for classification problem with more than 2 class labels. In this paper, we use an SVR classifier to determine the importance level of attributes and construct classification rules.

Given a set of training sample points,$\{(x_i, z_i), i = 1, .., l\}$,such that $x_i \in R^n$ is an input and $z_i \in R^1$is a target output, The primal form of Support vector Regression is

$$min_{w,b,\xi,\xi^*,\varepsilon} \frac{1}{2} w^T w + C(\nu\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)) \quad (1)$$

subject to

$$(w^T\phi(x_i)+b)-z_i \geq \varepsilon+\xi_i \tag{2}$$

$$z_i-(w^T\phi(x_i)+b) \geq \varepsilon+\xi_i^* \tag{3}$$

$$\xi_i,\xi_i^* \geq 0, i=1,..,l,\varepsilon>0 \tag{4}$$

Here sample vector x are mapped into a higher dimensional space by the function ϕ.$C > 0$is the penalty parameter of the error item. In this paper we usually let $C = 10 \sim 10^5$.The ν and ε are two parameters. The parameter $\nu, \nu \in (0,1]$,control the number of support vectors. In this paper we usually set $\nu = 0.5$,$\varepsilon = 0.1$. Furthermore, $K(x_i,x_j) \equiv \phi(x_i)^T\phi(x_j)$. is the kernel function. In this paper, we use the following radial basis function (RBF) as kernel function.

$$K(x_i,x_j) = exp(-\gamma||(x_i-x_j)^2||), \gamma>0 \tag{5}$$

Here,γ is kernel parameter. In this paper we usually let $\gamma = (0.1 \sim 1)/n$, $n$is the number of attributes in training sample set. The formula (1) dual is

$$min_{\alpha,\alpha^*}\frac{1}{2}(\alpha-\alpha^*)^TQ(\alpha-\alpha^*)-z^T(\alpha-\alpha^*) \tag{6}$$

subject to

$$e^T(\alpha-\alpha^*)=0, e^T(\alpha+\alpha^*) \leq C\nu \tag{7}$$

$$0 \leq \alpha_i,\alpha^* \leq C, i=1,\ldots,l \tag{8}$$

Where e is the vector of all ones, Q is a l by l positive semidefinite matrix. The output function of SVR classifier is

$$Z(x)=\sum_{i=1}^{l}(-\alpha+\alpha^*)K(x_i,x)+b \tag{9}$$

Assuming the attribute vector of classification problems is x=[$X_1, X_2, \ldots, X_n$], where $n$is the number of attributes in sample set, and the corresponding classification label is z and $z \in R^1$, then the sample of classification problems can be represented as $< X, z >$.In order to constructing SVR classifiers as the form shown by formula (9), quan-tification and normalization of the attribute values and classification labbels will be carried out as follows.

The quantification is performed for the values of discrete attributes and classification labels. In this paper, values of discrete attributes and classification labels are quantified as integer numbers in some order, for example,0,1,2,3,....

The normalization is performed to adjust the of SVR input ranges. For a given attribute value space Ω,utilizing the following linear transformation to map the attribute value X of sample to the SVM input$x$,making every elements in the x with the same range of $[\Delta, -\Delta]$.

$$x = bX + b_0 \tag{10}$$

where $b = (b_{ij})$ is a transformation coefficient matrix.Here

$$b_{ij} = \begin{cases} \frac{2\Delta}{MaxX_i - MinX_i} : j = i \\ 0 : \qquad\qquad otherwise \end{cases} \tag{11}$$

$b_0 = (b_{0i})$ is a transformation vector.Here,

$$b_{0i} = \Delta - a_{ii} MaxX_i \tag{12}$$

The parameter Δ affects the generalization of trained SVM, in this paper we usually set $\Delta = 0.5 \sim 2$.

During the construction of classification rules, only the attribute space covered by the sample set should be taken into account. Obviously according to formula (9), when the kernel function of SVR is the radial basis function shown by formula (5), any order derivatives of network output $Z(x)$ to each SVR input x_k exist.

4 Measure for Attribute Importance Ranking

Without losing the universality, next we will discuss the classification problems with two attributes and two class labels.

For a 2-dimension classification problem, assuming the shape of classification hypersurface in the given area Ω is as shown in Fig.1, in which the perpendicular axis is attribute $Z(x)$is class label, the area A and B are the distribution area of different classes. In the cases (a) and (b),the importance level of x_1 for classification is obviously higher, so area should be divided via attribute x_1.In the case (c), attribute x_1 and attribute x_2 have the equal classification powers. Therefore, for a given attribute value space Ω , the importance level of each attribute depends on the mean perpendicular degree between each attribute axis and classification hypersurface in space Ω or its adjacent space. The higher is the mean perpendicular degree, the higher is the importance level.

For a given sample set, the attribute value space Ω is defined as follows.

$$\Omega = \{x | Minx_k \le x_k \le Maxx_k, k = 1, \ldots, n\} \tag{13}$$

Where $Minx_k$ and $Maxx_k$are the minimal and maximal value of k-th attribute in the given sample set, respectively.

For a given trained SVM and the attribute value Γ,$\Gamma \subset \Omega$, the perpendicular level between classification hypersurface and attribute axis x_k is defined as follows.

$$P_{x_k}(x) = \frac{|\frac{\partial Z(x)}{\partial x_k}|}{\sqrt{\sum_k [(\frac{\partial Z(x)}{\partial x_k})^2 + 1]}} \tag{14}$$

According to formula (14), the value of the perpendicular level $P_{x_k}(x)$, mainly depends on the value $\frac{\partial Z(x)}{\partial x_k}$. Therefore for the convenience of computing, we can use the following formula to replace formula (14).

$$P_k(x) = |\frac{\partial Z(x)}{\partial x_k}| \tag{15}$$

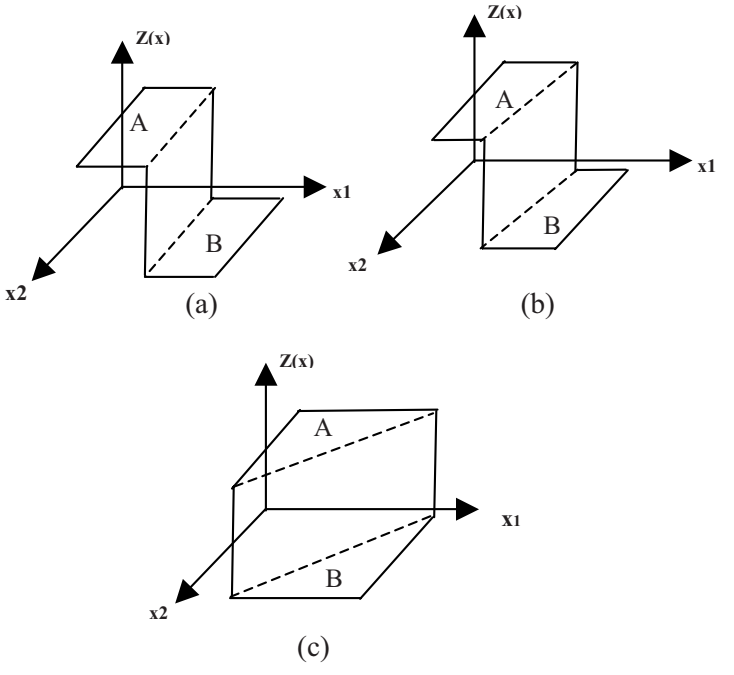

Fig. 1. Typical shapes of classification hypersurface

From the formula (9) and (5), we can get

$$\frac{\partial Z(x)}{\partial x_k} = \sum_{i=1}^{l} 2(-\alpha + \alpha^*)\gamma(x_{ik} - x_k)K(x_i, x) \tag{16}$$

Here the x_{ik} and the x_k are the k-th attribute values of the j-th Support vector and the sample point x, respectively.

For a given attribute value space Γ, $\Gamma \subset \Omega$,the measurement of classification power of attribute x_k is defined as follows.

$$JP(x_k) = \begin{cases} \frac{\int_{V_\Gamma} P_k(x)dx}{\int_{V_\Gamma} dx} : \int_{V_\Gamma} dx \neq 0 \\ 0 : \qquad\qquad otherwise \end{cases} \tag{17}$$

The importance level measure $JP(x_k)$ of the attribute x_k represents the influence degree of attribute x_k to classification. So in the process of rules extraction, the value $JP(x_k)$ is the important instruction information for selecting attributes and dividing attribute value space.

The typical classification problem of weather type for playing golf is employed to demonstrate the performance of the new measure method. The computing results are shown as table 1. The attributes and their values are as follows: Outlook

has the value of sunny, overcast and rain, quantified as 0, 1, 2. Temperature has the value of $65 \sim 96$. Humidity has the value of $65 \sim 96$. Windy has the value of true, false, quantified as 0, 1. The size of the training sample set is 14.

Table 1. Computing Results of Measurement Value $JP(x_k)$

Attributes	Whole Area	Outlook=sunny	Outlook=rain
Outlook	0.0547	–	–
Temperature	0.0292	0.0551	0.0334
Humidity	0.0353	0.0861	0.0552
Windy	0.0362	0.0457	0.0929

From table 1, in the whole space the measure value of importance level of attribute outlook is the biggest, therefore attribute outlook should be selected as the root node of the decision tree, and the attribute value space should be divided by its values. While in the subspace of outlook= rain, measure value of attribute windy is the biggest. Thus according to this information, the optimal decision tree and corresponding classification rules can be generated.

5 Rules Extraction Method

The algorithm for classification rule construction from trained SVR proposed in this paper is described as follows.

Step 1: Initializing.

a) Divide the given sample set into two parts, the training sample set and the test set. According to the training sample set, generate the attribute value space Ω by formula (13).

b) Set the interval number of attributes and the predefined value of error rate.

Step 2: Rule generating.

a) Generate a queue R for finished rules and a queue U for unfinished rules.

b) Select attribute x_k with the biggest value of $JP(x_k)$ computed by formula (17) as the extending attribute out of the present attributes. Divide the attribute x_k into intervals according to the chosen interval number. Then for each interval, pick attribute x_j with the biggest $JP(x_k)$ as the extending attribute for each interval. Merge the pairs of adjacent intervals with the same extending attribute and same class label with the largest proportion in all of the class labels. A rule is generated for each merged interval. If the class error of the generated rule is less than the predefined value, put it into the queue R, otherwise put it into the queue U.

c) If U is empty, the extraction process terminates, otherwise go to d).

d) Pick an unfinished rule from the queue U by a certain order, and perform division and mergence. A rule is generated for each merged interval. If the class error of the generated rule is less than the predefined value, then put it into the queue R, otherwise put it into the queue U. Go to c).

Step 3: Rule Processing.

Check the rule number of each class label. Let the rules of the class label with the largest number of rules be default rules.

6 Experiment and Analysis

The spiral problem [13] and congressional voting records(voting for short), hepatitis, iris plant(iris for short), statlog australian credit approval(credit-a for short) in UCI data sets [14] are employed as computing cases, shown in table 2. The attribute value distribution of spiral problem is shown as Fig.2, in which solid points are of Class C0, empty points are of Class C1.

Table 2. Computing Cases

	Spiral	Voting	Hepatitis	Iris	Credit-A
Total Samples	168	232	80	150	690
Training Samples	84	78	53	50	173
Testing Samples	84	154	27	100	517
Classification Numbers	2	2	2	3	2
Total Attributes	2	16	19	4	15
Discrete Attributes	0	16	13	0	9
Continuous Attributes	2	0	6	4	6

Table 3. Experimental Results Comparison between New Approach(NA) and C4.5R

	#Rules(NA: C4.5R)	Err.Train(NA: C4.5R)	Err.Test(NA: C4.5R)
Spiral	8: 3	0: 38.1%	1.1: 40.5%
Voting	3: 4	2.5%: 2.6%	2.5%: 3.2%
Hepatitis	3: 5	7.5%: 3.8%	11.1%: 29.6%
Iris	4: 4	0%: 0%	4%: 10%
Credit-A	5: 3	12.1%: 13.9%	14.5%: 14.9%

Since no other approaches extracting rules from SVR are available, we include a popular rule learning approach i.e.C4.5R for comparison. The experimental results are tabulated in Table 3. For the spiral problem and the Iris plant problem, the rules set extracted by the new approach are shown in Table 4 and Table 5, respectively. Table 3 shows that the rules extraction results of the new approach are obviously better than that of C4.5R, especially for spiral problem. For the case of spiral problem, C4.5R is difficult to extract effective rules, but the new approach has so impressive results that are beyond our anticipation. This means that the new approach proposed can improve the validity of the extracted rules for complicated classification problems remarkably. Moreover, the generalization ability of those rules extracted by the new approach is also better than that of rules extracted by the C4.5R.

Table 4. Rules Set of Spiral Problem Gemerated by the Algorithm Proposed

R1 $x_0 \geq 2.22 \longrightarrow C1$
R2 $x_0[-2.22, -1.33) \wedge x_1 < 1.68 \longrightarrow C1$
R3 $x_0[-1.33, 0) \wedge x_1 < -1.75 \longrightarrow C1$
R4 $x_0[-1.33, 0) \wedge x_1[0.76, 1.2) \longrightarrow C1$
R5 $x_0[0, 1.33) \wedge x_1 < -2.2 \longrightarrow C1$
R6 $x_0[0, 1.33) \wedge x_1[-0.76, 1.75] \longrightarrow C1$
R7 $x_0[1.33, 2.22) \wedge x_1 < -1.68 \longrightarrow C1$
R8 $Default \longrightarrow C0$

Table 5. Rules Set for the Iris Plant Problem Generated by the Algorithm Proposed

R1 $petalwidth < 0.72 \longrightarrow Iris - setosa$
R2 $petalwidth \geq 1.66 \longrightarrow Iris - virginica$
R3 $petalwidth[1.34, 1.66] \wedge petallength \geq 4.91 \longrightarrow Iris - virginica$
R4 $Default \longrightarrow Iris - versicolor$

7 Conclusions

In this paper, based on the analysis of the relations among the characteristics of position and shape of classification hypersurface, the partial derivative distribution of the trained SVR output to the corresponding inputs, a new measure for determining the importance level of the attributes based on the differential information of trained SVR is proposed, which is suitable for both continuous attributes and discrete attributes, and can overcome the shortcomings of the measure method based on information entropy. On the basis of this new measure, a new approach for rules extraction from trained SVR is presented, which is also suitable for classification problems with continuous attributes. The performance of the new approach is demonstrated by several typical examples, the computing results prove that the new approach can improve the validity of the extracted rules remarkably compared with other rule extracting approaches, especially for complicated classification problems.

References

1. Fu, L M.: Rule generation from neural network. IEEE Trans on Sys, Man and Cybernetics 8, 1114–1124 (1994)
2. Towell, G., Shavlik, J.A.: The extraction of refined rules from knowledge-based neural networks. Machine Learning 1, 71–101 (1993)
3. Lu, H.J., Setiono, R., Liu, H.: NeuroRule: a connectionist approach to data mining. In: Proceedings of 21th International Conference on Very Large Data Bases, Zurich, Switzerland, pp. 81–106 (1995)
4. Zhou, Z.H., Jiang, Y., Chen, S.F.: Extracting symbolic rules from trained neural network ensembles. AI Communications 6, 3–15 (2003)

5. Sestito, S., Dillon, T.: Knowledge acquisition of conjunctive rules using multilayered neural networks. International Journal of Intelligent Systems 7, 779–805 (1993)
6. Craven, M.W., Shavlik, J.W.: Using sampling and queries to extract rules from trained neural networks. In: Proceedings of the 11th International Conference on Machine Learning, New Brunswick, NJ, pp. 37–45 (1994)
7. Maire, F.: Rule-extraction by backpropagation of polyhedra. Neural Networks 12, 717–725 (1999)
8. Setiono, R., Leow, W.K.: On mapping decision trees and neural networks. Knowledge Based Systems 12, 95–99 (1999)
9. Battiti, R.A.: Using mutual information for selecting featuring in supervised net neural learning. IEEE Trans on Neural Networks 5, 537–550 (1994)
10. Bollacker, K.D., Ghosh, J.C.: Mutual information feature extractors for neural classifiers. In: Proceedings of 1996 IEEE international Conference on Neural Networks, Washington, pp. 1528–1533 (1996)
11. Dash, M., Liu, H., Yao, J.C.: Dimensionality reduction of unsupervised data. In: Proceedings of the 9th International Conference on Tools with Artificial Intelligence, Newport Beach, pp. 532–539 (1997)
12. Fu, X.J., Wang, L.P.: Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. IEEE Transactions on Systems, Man and Cybernetics, Part B - Cybernetics. 33, 399–409 (2003)
13. Kamarthi, S.V., Pittner, S.: Accelerating neural network training using weight extrapolation. Neural Networks 12, 1285–1299 (1999)
14. Blake, C., Keogh, E., Merz, C.J.: UCI repository of machine learning databases, Department of Information and Computer Science, University of California, Irvine, CA(1998), [`http://www.ics.uci.edu/~meearn/MLRepository.htm`]