Optimization Approach for Feature Selection and Classification with Support Vector Machine

S. Chidambaram and K.G. Srinivasagan

Abstract The support vector machine (SVM) is a most popular tool to resolve the issues related to classification. It prepares a classifier by resolving an optimization problem to make a decision which instances of the training data set are support vectors. Feature selection is also important for selecting the optimum features. Data mining performance gets reduced by Irrelevant and redundant features. Feature selection used to choose a small quantity of related attributes to achieve good classification routine than applying all the attributes. Two major purposes are improving the classification functionalities and reducing the number of features. Moreover, the existing subset selection algorithms consider the work as a particular purpose issue. Selecting attributes are made out by the combination of attribute evaluator and search method using the WEKA Machine Learning Tool. In the proposed work, the SVM classification algorithm is applied by the classifier subset evaluator to automatically separate the standard information set.

Keywords Data mining · Kernel methods · Support vector machine · Classification

1 Introduction

A Support Vector Machine is an overture which is considered to categorize vectors of input features into single of the two categories. By constructing SVMs for dissimilar class pairs, it is possible to reach a classification into various categories. The procedures of the decision function are trained on a set of construction of

S. Chidambaram (✉)
Department of IT, National Engineering College, Kovilpatti, India
e-mail: chidambaramraj1@gmail.com

K.G. Srinivasagan
Department of CSE, National Engineering College, Kovilpatti, India
e-mail: kgsnec@rediffmail.com

H.S. Behera and D.P. Mohapatra (eds.), *Computational Intelligence in Data Mining—Volume 1*, Advances in Intelligent Systems and Computing 410, DOI 10.1007/978-81-322-2734-2_11

patterns which contains the class labels. The major learning procedures are those that guide to the SVM with the minimum classification error on feature vectors. SVMs with a Gaussian kernel and the learning parameters are interpreted as follows: (1) the kernel parameter γ, measuring the distribution of the essence in the input space; and (2) the cost parameter C, measuring the comparative meaning of the error phrase in the cost procedure being reduced during SVM training. For that, one or more training and validation cycles are required. Each cycle contains one or more assessment procedures in which an SVM is used as a part of the construction of a model called the training set. The remaining portion is applied for the validation set.

2 Related Work

Huang et al. [1] compared neural networks with genetic programming and decision tree classifiers. For some set of input features, SVM classifier has accomplished the maximum classification accuracy. In the proposed hybrid GA-SVM strategy, it is possible to execute feature selection and parameters optimization process. Yang et al. [2] expressed the method which can extract three important attributes from user's network behavior. The extracted feature is categorized into various types, such as browsing news and downloading shared resources etc. Support vector machine is suitable for performing clustering process. It also provides rapid and valid execution, particularly for the tiny datasets.

Nematzadeh Balagatabi et al. [3] described the support vector machine which depends on cognitive style factors. A significant focus of this research is to categorize the researchers with respect to decision tree and Support Vector Machine methods.

3 Feature Selection

This procedure starts with subset creation that travels along a search policy to generate feature subsets. Each subset is estimated based on an evaluation condition and make a comparison with the previous solution. If the evaluated subset produces a better result, then it should be replaced with the old one, else remains the same. The method of subset creation and evaluation gets replicated until it satisfies the certain criteria. Figure 1 shows the Feature Selection process. At last, the validation applies to the selected best subset by using the prior data or a few test data. Search method and evaluation criteria are the two important points in feature selection.

Subset creation starts from the collection of attributes in the data set. It can be of any one of the following such as an empty set, the full set or a randomly formed subset. The feature subsets are searched in special ways, such as forward, backward, and random.

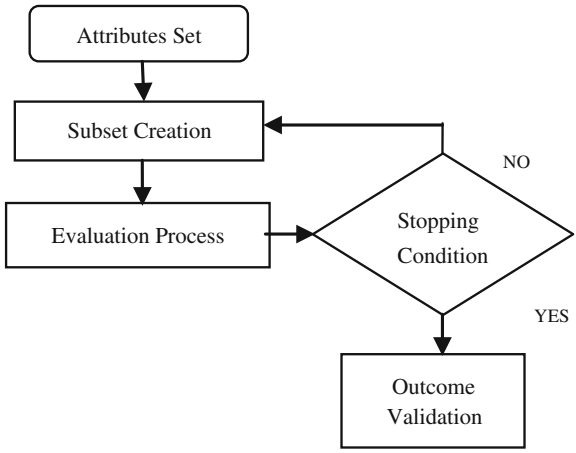

Fig. 1 Feature selection process

During evolutionary process, after the generation of feature subsets, the results are measured by some criteria to determine their performance. Normally, the main functionality of feature subsets is the discriminating capability of subsets that differentiate between or within various classes. With respect to the dependency of learning algorithms, it can be mostly classified into Wrapper Method, Filter Method. Wrapper method is used to determine attribute subsets with maximum accuracy because the features are coordinated exactly with the learning algorithms.

An attribute selection approach gets completed in any one of the observing conditions:

(i) If the search process is finished.
(ii) If a predefined size of feature subsets are finalized.
(iii) If a predefined number of iterations are completed.
(iv) If an optimal feature subset along with the evaluation process is accomplished.
(v) If the modification (addition or deletion) of feature subsets does not create a better subset.

In the outcome validation, related features are known in advance. Then it is possible to validate the feature selection outcome by comparing the previous facts. Though, in the real time applications, it is very hard to find which features are mostly related. Feature selection engages searching activity through all potential combinations of attributes in the data to discover which subset of attribute work well for improving performance of classification.

4 Proposed Method

In the proposed method, the classifier subset evaluator is combined with the search method which produces the maximum classification accuracy by compared to performing an individual function. In this way optimization can be achieved.

4.1 Attribute Evaluator

Using this method, attributes subset can be assessed. Another way is, creating a model and measuring the accuracy of the model. Classifier subset evaluator is used as an attribute evaluator to access the attributes of giving data sets. It Evaluates attribute subsets on training data.

Option in Classifier Subset Evaluator

- Classifier = Classifier to apply for evaluating the accuracy of subsets
- Hold out File = File containing hold out/test instances.
- Use Training = Use training data instead of hold out/test instances.

4.2 Search Method

It is the structured representation in which the possible attributes subsets are navigated with respect to the subset evaluation. Best first search may begin with the empty set of attributes and proceed forward, or begin with the complete set of attributes and proceed in the reverse direction, or begin at any position and proceed in both the direction. The combination of best first search method and Classifier Subset Evaluator will give the best feature selection in result which is further used for improving classification accuracy.

Some of the options available in Best First Search method are,

- Direction = Set the direction in which the search process carried out.
- Lookup Cache Size = Position the highest size of the lookup cache of estimated subsets. This can be represented as multiplying the given number of features in the specified dataset.
- Search Termination = Set the quantity of backtracking.

5 C4.5 Classification

C4.5 algorithm is an evolution of ID3. It uses Gain Ratio as splitting criteria. But in ID3 considers, gain as splitting criteria in tree growth phase. This algorithm considers both continuous and discrete attributes. In order to work with the continuous attributes, C4.5 sets a threshold for an attribute value and then splits the list of the given attribute value with respect to threshold. C4.5 can operate with the case of attributes with continuous domains by discretization. Missing attribute values are not used for calculating gains and entropy. In the tree pruning those misclassification errors are decreased. The steps of C4.5 algorithm in decision tree construction are specified below:

- Select the attribute of the root node
- Construct branch based on each attribute value and split conditions.
- Replicate the steps for each branch until all instances in the branch come under the single class.

The root node is selected based on the feature whose gain ratio is greater. Gain ratio is measured by Eq. (1).

$$GainRatio(D,S) = \frac{Gain(D,S)}{H\left[\frac{|D_1|}{|D|},\ldots,\frac{|D_s|}{|D|}\right]} \tag{1}$$

where S = {D_1, D_2…D_s} denotes the new states.

6 SVM Classification

Support vector machine is a very efficient and suitable approach for regression, classification and pattern recognition. The main idea of SVM is to calculate the exact classification procedure to differentiate the instances of the two different classes in the training data. The classification parameters can be calculated geometrically.

The margin is the quantity of space or act as a separator between the two different classes by using the hyper plane. To substantiate the maximum margin hyper planes, an SVM classifier attempt to improve the procedure based on W and b.

$$L_p = \frac{1}{2}||W|| - \sum_{i=1}^{t} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{t} \alpha_i \tag{2}$$

In Eq. (2), 't' is the number of training samples and α_i, i = 1,…, t are positive numbers such that the derivatives of Lp based on α_i are zero. α_i is the Lagrange multipliers and Lp is called the Lagrangian. In this equation, 'w' is the vectors and constant 'b' used to identify the hyper plane. A machine learning approach such as the SVM can be designed as a specific task class based on few parameters 'α'.

6.1 Linear Kernel

This type of kernel function is represented by the inner product <x, y> plus an optional constant 'c'. A kernel algorithm uses a linear kernel which is same as the non kernel counterparts. It is calculated by Eq. (3).

$$k(x, y) = x^T y + c \tag{3}$$

6.2 Polynomial Kernel

This type of kernel is a non-stationary kernel. Polynomial kernels are appropriate for the issues where all the training data are normalized.

$$k(x, y) = (\alpha x^T y + c)^d \tag{4}$$

In Eq. (4), 'α' is the slope and constant term 'c', polynomial degree 'd' and 'x' 'y' are the feature vectors.

6.3 Radial Basis Kernel Function

By this type of kernel function, the transformation is denoted by Eq. (5),

$$k(x, y) = \exp\left\{\frac{\|x - y\|^2}{2\sigma^2}\right\} \tag{5}$$

where 'σ' is the variance to measure the kernel output. If it is overestimated, the exponential will perform linearly. If it is underestimated, then the procedure will require regularization and decision boundary. It is extremely sensitive to noise during training stage.

6.4 Sigmoid Kernel Function

This type, also called as a hyperbolic Tangent kernel method. It is similar to a two layer perceptron neural network. It is denoted by Eq. (6),

$$k(x, y) = \tanh(\alpha x^T y + c) \tag{6}$$

where 'α' and constant 'c' are the adjustable parameters in this kernel and 'x' and 'y' are the feature vectors. The general value for alpha is 1/N, where N is the data dimension.

7 Experimental Results

The proposed method is implemented to the diabetes data set. It consists of eight numbers of numerical attributes. And ten fold crossover validation is applied to validate the results of classifiers.

The following performance metrics such as accuracy, error rate, precision, recall, are measured to find the efficiency of the proposed method. Accuracy measures the proportion of correct predictions considering the positive and the negative inputs. The error rate is the average loss over the test set. Precision is the positive predictions proportions that are correct, and recall is the positive sample proportion that is correctly positively predicted. From the Table 1, the proposed method produced the maximum accuracy (96.5 %) by compared to other classifiers performance. Figure 2 shows the performance comparison of various classifiers.

Table 1 Performance comparison of various classifiers with the proposed method

Algorithms	Acc (%)	ER (%)	TP	FP	TN	FN	Pre	Rec	Sen	Spe	TT (s)
C4.5	68.1	11.5	0.3	0.21	0.65	0.79	0.5	0.4	0.5	0.5	1698.8
Linear SVM	70.6	29.3	0.7	0.53	0.30	0.47	0.6	0.7	0.7	0.6	1655.2
Polynomial SVM	69.9	30.6	0.6	0.47	0.31	0.53	0.4	0.7	0.5	0.7	1856.6
RBF SVM	71.6	28.3	0.7	0.58	0.29	0.42	0.6	0.7	0.6	0.7	1620.6
Sigmoidal SVM	69.9	30.6	0.6	0.69	0.31	0.31	0.5	0.6	0.6	0.7	1900.5
Proposed method (RBF SVM+CSE)	96.5	3.49	0.9	0.04	0.51	0.55	0.9	0.8	0.9	0.8	1422.8

RBF Radial Basis Function, *CSE* Classifier Subset Evaluation, *Acc* Accuracy, *ER* Error Rate, *TP* True Positive, *FP* False Positive, *TN* True Negative, *FN* False Negative, *Pre* Precision, *Rec* Recall, *Sen* Sensitivity, *Spc* Specificity, *TT* Time Taken

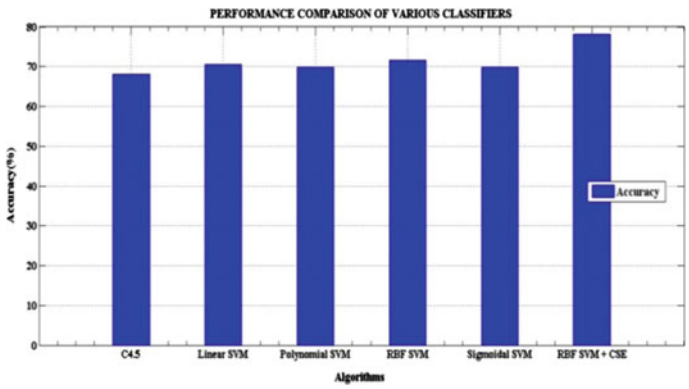

Fig. 2 Performance measures of various classifiers

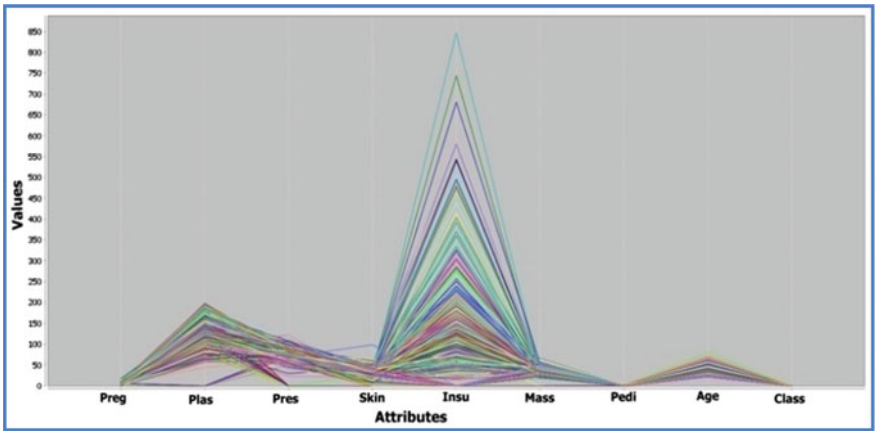

Fig. 3 Parallel coordinates plot for the diabetes dataset

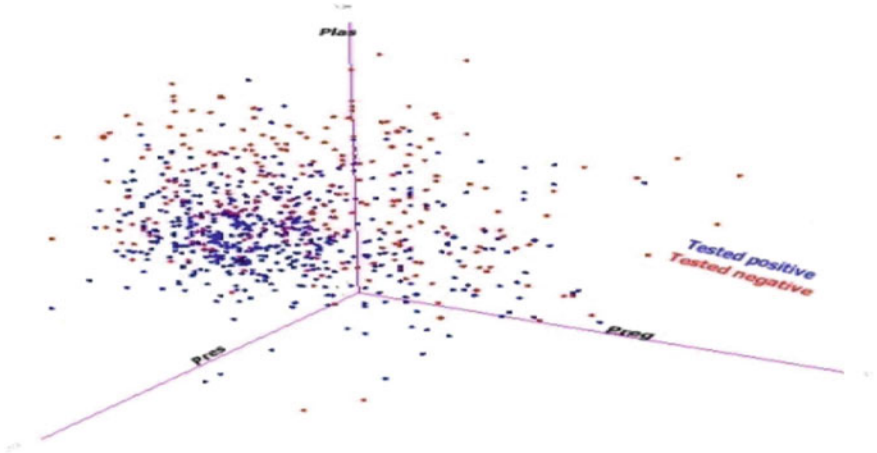

Fig. 4 3D scatter plot for the diabetes dataset

Parallel coordinates are an ordinary approach of visualizing high dimensional geometry and evaluating multivariate data. There are three important visualization parameters are considered, such as the order, the rotation, and the scaling of the axes. Figure 3 shows the parallel coordinate graph for the diabetes dataset. Here the data exploration is given for each attribute in the dataset.

In the same way, 3D scatter plot for the diabetes data set also shown in Fig. 4. This representation is used to plot data points on three axes which show the relationship among three selected attributes.

8 Conclusion

In this paper, proposed a SVM based optimization algorithm, which can used to optimize the parameter of SVM values. It is also used for retrieving the features of optimal subset by applying the SVM method to remove insignificant features and effectively find best parameter values. The main purpose of this work is to combine the SVM classifiers along with the attribute evaluators for achieving maximum accuracy in classification. And also from the experimental results, observed that the classifier subset evaluator is best suitable for optimization.

References

1. Huang, C.-L., Chen, M.-C., Wang, C.-J.: Credit scoring with a data mining approach based on support vector machines. Expert Syst. Appl. **33**, 847–856 (2007)
2. Yang, Z., Su, X.: Customer behavior clustering using SVM. Int. Conf. Med. Phys. Biomed. Eng. Elsevier Physics Proc. **33**, 1489–1496 (2012)
3. Nematzadeh Balagatabi, Z.: Comparison of decision tree and SVM methods in classification of researcher's cognitive styles in academic environment. Int. J. Autom. Artif. Intell. **1**(1) (2013). ISSN:2320–4001
4. Qi, Z., Tian, Y., Shi, Y.: Structural twin support vector machine for classification. Knowl.-Based Syst. **43**, 74–81 (2013). doi:10.1016/j.knosys.2013.01.008
5. Maldonado, S., Weber, R., Basak, J.: Simultaneous feature selection and classification using Kernel-penalized SVM for feature selection. Inform. Sci. **181**(1), 115–128 (2011). doi:10.1016/j.ins.2010.08.047
6. Song, L., Smola, A., Gretton, A., Bedo, J., Borgwardt, K.: Feature selection via dependence maximization. J. Mach. Learn. Res. **13**(1), 1393–1434 (2012)
7. Yu, H., Kim, J., Kim, Y., Hwang, S., Lee, Y.H.: An efficient method for learning nonlinear ranking SVM functions. Inform. Sci. **209**, 37–48 (2012). doi:10.1016/j.ins.2012.03.022
8. Maldonado, S., Lopez, J.: Imbalanced data classification using second-order cone programming support vector machines. Pattern Recogn. **47**(5), 2070–2079 (2014). doi:10.1016/j.patcog.2013.11.021
9. Carrizosa, E., Martín-Barragán, B., Romero-Morales, D.: Detecting relevant variables and interactions in supervised classification. Eur. J. Oper. Res. **213**(16), 260–269 (2014). doi:10.1016/j.ejor.2010.03.020
10. Hassan, R., Othman, R.M., Saad, P., Kasim, S.: A compact hybrid feature vector for an accurate secondary structure prediction. Inform. Sci. **181**, 5267–5277 (2011). doi:10.1016/j.ins.2011.07.019
11. Patel, K., Vala, J., Pandya, J.: Comparison of various classification algorithms on iris datasets using WEKA. Int. J. Adv. Eng. Res. Dev. (IJAERD) **1**(1) (2014). ISSN:2348–4470
12. Tian, Y., Shi, Y., Liu, X.: Recent advances on support vector machines research. Technol. Econ. Dev. Econ. **18**(1), 5–33 (2012)
13. Victo Sudha George, G., Cyril Raj, V.: Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile. Int. J. Comput. Sci. Eng. Surv. **2**(3), 16–27 (2011)
14. Qi, Z.Q., Tian, Y.J., Shi, Y.: Robust twin support vector machine for pattern classification. Pattern Recogn. 305–316 (2013)