World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 2, No. 5, 174-178, 2012

Multi-Level Support Vector Machine

Milad Aghamohammadi
Department of Computer Engineering
Iran University of Science and Technology
Tehran, Iran

Morteza Analoui
Department of Computer Engineering
Iran University of Science and Technology
Tehran, Iran

Abstract—Many type of classifiers have been presented in machine learning and large margin classifier is one of them. Support Vector Machine (SVM) is the most famous of large margin classifiers. SVM is a very useful classifier, but has some limitations. Only patterns near the decision boundary are used as support vectors and decision making in SVM is done locally. In this paper we propose a method that also uses information in other patterns for classification which is called "Multi-Level Support Vector Machine" (MLSVM). We compare our method with the original SVM and an artificial neural network using some UCI datasets and the final results shows that our method is better or equal to other methods.

Keywords- Pattern Recognition; Classification; Large Margin Classifier; Multi-Level Support Vector Machine.

I. INTRODUCTION

Classification is an important part of machine learning and large margin classifier is one of them. Support Vector Machine (SVM) [1]-[2] [3][4][5][6] is the most famous of large margin classifiers. SVM finds a hyperplane which separates data into two classes. The hyperplane is calculated in a way that it will have the maximal margin possible from patterns of each class. SVM is a very useful method but only uses some patterns that are near the decision boundary that are called “Support Vectors” and define the separating hyperplane due to their positions and ignores the information in other patterns.

In this paper, we present a method that not only makes decision using patterns near the decision boundary, but takes advantage of other patterns as well. This method is called “Multi-Level Support Vector Machine” (MLSVM).

Cheng et al. [7] proposed a method called “Localized Support Vector Machine” (LSVM) by changing the SVM objective function and adding a similarity term to the original SVM. For each unlabeled pattern, a new local SVM is trained and By using Euclidian distance as similarity function, only the patterns near the unlabeled pattern have the ability to become support vector. This method has a high order of Time complexity when in comes to testing phase and this makes it unpractical for real-time problems.

Blanzieri et al. [8] proposed a method called “k-Nearest Neighbor Support Vector Machine” (kNNSVM) that first finds k-Nearest Neighbors of unlabeled pattern and then creates a SVM using only the neighbors to classify the unlabeled pattern.

Huang et al. [9] proposed a new method that used local and global information of data to define the decision boundary and called it "M^4". This method uses SVM to define the decision boundary but also uses covariance matrix of each class to make the final decision. Having the global information of each class in Covariance matrix gave this method the benefit of injecting global information in defining the final hyperplane.

This paper is organized as followed:

Section II recalls the SVM method and describes the new MLSVM method. Experimental results are presented in Section III and conclusion is in Section IV.

II. PROPOSED METHOD

A. Support Vector Machines

SVM is a useful method for classification in high dimensional problems. In this method, two class of data separate with a hyperplane with has maximal margin from patterns of each class. If two class of data can not separate in linearly, SVM uses the kernel trick [10] which maps input data to a higher dimensional space where may be find a hyperplane that can separate two class of data linearly. Using the kernel function allow to SVM avoids the costly computation in new high dimensional space.

Objective function of none linear SVM model can be written as

$$\min_{W,\xi,b} \quad L_P = \frac{1}{2}\|W\|^2 + C(\sum_{i=1}^{n}\xi_i)^k \qquad (1)$$

$$s.t. \quad y_i(W^T.\Phi(x_i)+b) \geq 1-\xi_i,$$
$$\xi_i \geq 0, i = 1,...,n$$

Where x_i is *i*th pattern of training set, $y_i \in \{+1,-1\}$ is class label of x_i, $\Phi : R^d \mapsto H$ is a mapping function, ξ_i is slack variable of x_i that let to x_i for misclassification, $W \in H$ is normal vector of the separating hyperplane, $b \in \Re$ is distance of hyperplane from the origin and C is a parameter that define by user and control the trade-off between width of margin and misclassification risk.. With Lagrangian multiplier method we can insert constraint of (1) to objective function. For all positive value of k, (1) will be a convex problem and can solve dual of it. For converting L_P to its dual form, must be vanish of gradient of L_P with respect to W and b and ξ. This type of dual in [11] is called the Wolf dual. For k=1, dual of (1) has been changed to

$$\max_{\alpha} \quad L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum\alpha_i\alpha_j y_i y_j K(x_i,x_j) \qquad (2)$$

$$s.t. \quad 0 \leq \alpha_i \leq C,$$
$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

Where, α_i is a positive Lagrangian multiplier for *i*th inequality constraint in (1) and $K(x_i,x_j)$ is kernel function used for computing $\Phi(x_i).\Phi(x_j)$. After the training SVM and finding values of α_i, class label of unknown pattern x' has been defined as follow:

$$SVM(x') = sign(\sum_{i=1}^{n}\alpha_i y_i \Phi(x_i).\Phi(x') + b)$$
$$= sign(\sum_{i=1}^{n}\alpha_i y_i K(x_i,x') + b) \qquad (3)$$

Note that the class label of unknown pattern depends only on training points that have nonzero α_i. Such training points are called "support vectors".

With solving (2), normal vector (W) of hyperplane will be

$$W = \sum_{i=1}^{n}\alpha_i y_i x_i \qquad (3)$$

B. *Multi-Level Support Vector Machine (MLSVM)*

As it have been mentioned before, in finding the margins, the only samples that matches our pattern are the one's that are close to the hyperplane and other samples have been found irrelevant to our pattern elsewhere.

"Fig. 1" shows a synthetic dataset which consists of two classes of +1 and -1. Samples that selected with SVM as support vectors are marked with a circle around them. As seen, only the places that are marked are specified to match favored results the other conclusions in this case have no value in defining vectors. But with, carefully observing it's noticeable that one of SVM results that are in +1 are the one which are encircled by do bold lining and this shows the existence of data in an area which had given similar results in other cases causes deduction in the width of margin (in the mentioned area), thus, the most favored outcomes and so it ruins the chances of separating hyperplanes.

Meanwhile, the rest of the data are located in different places and the majority of data are pointing elsewhere. It is obvious that this discussion with the existence of the kernel brought to this conclusion that when this result is sought a different environment it could have separable linear outcomes.

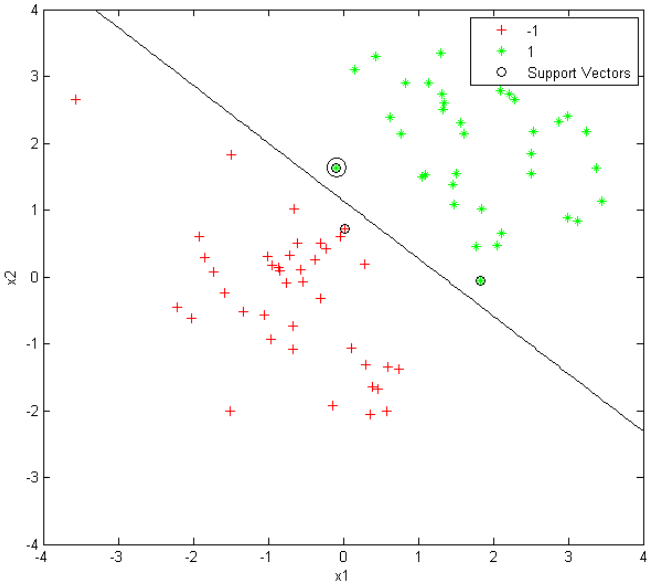

Figure 1. Artificial dataset and final decision of the SVM

The omittance of outlier brings up this crucial point that one of the pre-processes happens before classifications. Although,

the outcomes that follow a different pattern of repetition is omitted, but the possibility of cases such as "Fig. 1" is always possible. As it have been pointed out in "Fig. 1" the only results that are close to favored margin are considered relevant –regardless of the data and the places of their positioning in confirming our favored results.

The outcome although vary but they contain useful details that helps in classification of so called data. To get resolve hidden information embedded in the data we use MLSVM.

In the follow method and our first step the where about of possible answer is found by support vector indicates the primary resolution which we refer to it as step 1. In the current step some data are selected as support vector.

The new data is derived from old ones with the difference that certain support vectors have been removed from the equation. In second step, by using SVM on the new data in order to find new separating hyperplane. In future steps same as this step the support vectors are removed and by using a SVM on these data, we could ascertain new separating hyperplanes.

The results of 3timings of this method on the data in "Fig. 1", had been brought up on "Fig. 2". In this figure the hyperplane that we have in step 2 and 3 are different then the one presented in step 1 but their differences compare to each other is tolerable and by comparison their hyperplanes are in accordance to step 1 and it shows more general to the main pattern.

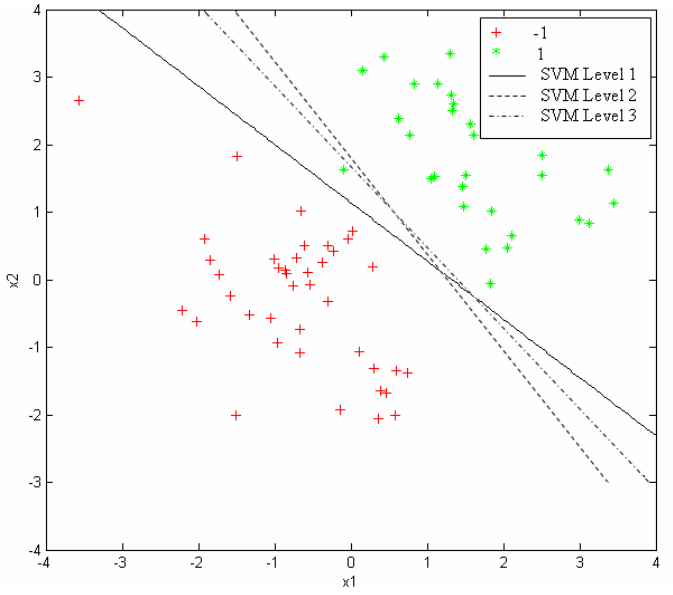

Figure 2. The 3 descending levels of MLSVM that have processed on the data of "Fig. 1"

C. MLSVM Classification

In order to reach the more generalized decision, it's more acceptable that the results of these steps are combined together. So, in this article we are using the idea of "Stack Generalization"[12] method to proceed.

The following method gives concept to voting the level of outcome this means that and no longer is a linear result approached. But, it is a more expanded than basic routines, by using a classifier for making a final decision this classifier needs training of its own .This concept is more conspicuous in "Fig. 3".

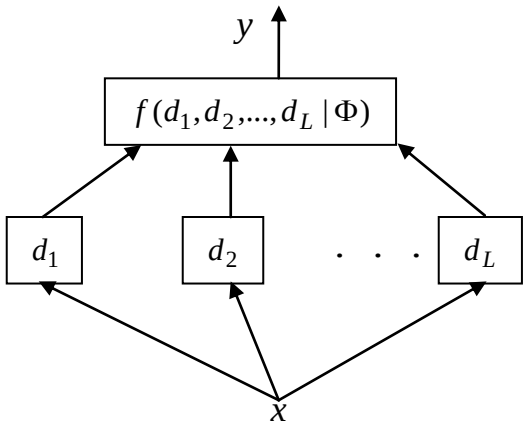

Figure 3. in stack generalization the combiner itself is classifier.[12]

In "Fig. 3", d_i 's are basis classifiers and $f(.|\Phi)$ itself is a classifier with Φ as its' parameter which the amount of it is determined by the training. In this part with accordance to the idea of stack-generalization it had been proposed the methods which it's combiner by using decision of levels and the unknown pattern itself defines the final class of the mentioned pattern.

The conspicuous process is reveled in "Fig 4" the data d_i, takes the class-label of unknown pattern, which is indicated by *i*th levels, the level (L) of processed data and the sample by itself can interfere in the process of as the entry data $f(.|\Phi)$ and this classifier goes back up the decision which relates with the type of the data.

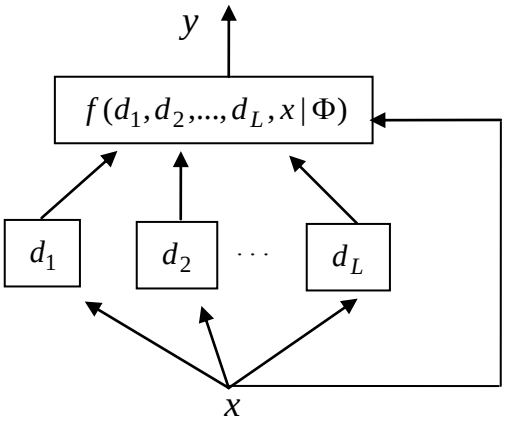

Figure 4. the suggested method for combining the results in different levels with regards to the stack- generalization persona

In this article, by using Multi Layered Perceptron (MLP) and also 2 levels of mentioned methods as combining classifier,

MLP tries to find a relation between features and label of class, by using algorithm to minimize the errors [13]. with reference to above mentioned notes and the high level of co-relations with the class label, the substantiality were determined by the levels and actual class label of data that was received by correctional pattern in our MLP process and the main focus is than turn on these features.

III. EXPERIMENTAL RESULTS

A. Parameter Selection

Because the suggested method is a combination of SVM's and MLP'S in order to see it's efficiency led to the comparison of these 3 methods. We use 2 levels of MLSVM and all of our SVM with Gaussian kernel. The structures of MLP are defined by use of different topology which was known to have been learned by a set of trail and error methods. The important point is that the same σ parameter is applied on both SVM and in level 1of MLSVM.

For the second level again we need to set the right amount of σ parameter and use it. The topology of MLP is the same, regardless of the both methods.

In order to, determine the σ parameter different amount $\{10^{-3},10^{-2},10^{-1},10^{0},10^{1},10^{2}\}$ have all been tested and the best amount with the most accuracy is selected 10-fold cross validation .also, the amount of C in SVM is an constant and it had been set on 1. The data have been converted to normal standard distribution.

B. Dataset Description

out of 7 dataset of UCI had been used for comparison of the results. The list of these datasets and their description had been presented on Table I.

C. Experimental Results

In Table II the average percentages of accuracy of 30 runs have been mentioned. In Table II in all 6 out of 7 cases of given groups MLSVM the proposed method is superior to the others. But, the higher average accuracy doesn't approve the fact that the method is better than the other. With applying statistical hypothesis testing can be measured to see whether there is a significant difference between the methods or not.

In this research, we use k-hold-out paired t- Test to see the difference among the mentioned methods .in this test we have presumed that, k=30 and the significant level is $\alpha = 0.05$. The amount of t in t-Table is 2.045.

TABLE II. THE CLASSIFICATION ACCURACY OF THE 3 METHODS ON DATA SETS IN TABLE I

Dataset name	MLP	SVM	MLSVM
BREAST	95.28	96.62	**96.84**
LETTER_MN	97.02	98.83	**98.86**
IRIS	93.22	95.44	**98.11**
LIVER	66.73	70.19	**71.42**
DIABETES	73.06	**76.60**	76.55
VEHICLE	75.10	74.92	**78.45**
AUSTRALIAN	84.13	86.52	**86.57**

Table III shows the statistic amount for SVM and MLP and proposed method in every dataset.

TABLE I. DATA SETS AND THEIR BREAF DESCRIPTION

Dataset name	Brief description	Training set cardinality	Testing set cardinality	Number of features
BREAST	Wisconsin breast cancer data	546	137	10
LETTER_MN	Letter recognition data (M and N)	946	629	16
IRIS	Iris data	120	30	4
LIVER	Liver disorder data	242	103	6
DIABETES	Pima Indians diabetes data	461	307	8
VEHICLE	Vehicle silhouettes data, from Statlog	507	339	18
AUSTRALIAN	Australian credit approval, from Statlog	414	276	14

TABLE III. THE STATISTICAL AMOUNT HAS BEEN EVALUATED BY PROPOSED METHOD AND SVM AND MLP FOR 30 RUN SO ANY AMOUNT THAT SURPASS 2.045 HAD BEEN BOLD. THE ZERO HYPOTHESES ARE REJECTED THAT MEANS THE 95% OF ACCURACY, DATASET IN SUGGESTED METHOD HAS SIGNIFICANT DIFFERENCE WITH THE RELATED METHOD.

Dataset name	MLP	SVM
BREAST	**4.0605**	**2.7572**
LETTER_MN	1.1403	0.3622
IRIS	**2.9891**	**6.1336**
LIVER	**4.8552**	**3.1368**
DIABETES	**5.8970**	0.3226
VEHICLE	**5.1097**	**8.0184**
AUSTRALIAN	**6.5500**	0.3759

In Table III the suggested method in all given dataset show better results compare to the MLP method and this differences are conspicuously significance in all cases except the LETTER_MN.

It should be reminded that the proposed method in comparison to SVM in all cases is accurate except the Diabetes and this difference is not a significant one.

Previously it was mentioned, the amount of t in t-Table with the degree of freedom 29 and the significance level $\alpha = 0.05$ the co-related measurement should be 2.045 but the evaluated statistical measurement in Table III in such place shows the difference is significant and is far beyond the caped number. The amount of t with the degree of freedom 29 and significance level of $\alpha = 0.01$ is equal with 2.756 so in reality all the given data in the suggested method shows the differences are significant with great accuracy of 99%.

IV. CONCLUSION

In this research, the suggested method resolves the informational data other than support vectors. This method is called MLSVM and with inspired by stack generalization idea we presented the complete classifier.

The accuracy of MLSVM have been tested over and over again on 7 UCI dataset by, in comparison to MLP and SVM had shown better accurate results and our suggested method in over all had been far superior than that of MLP and with the expectation of 1 case MLSVM had been better than SVM.

REFERENCES

[1] Vladimir Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[2] Vladimir Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[3] V. Magdin and Hui Jiang, "Large margin estimation of n-gram language models for speech recognition via linear programming," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, Dallas, TX, 2010, pp. 5398 - 5401.

[4] A.M.S. Muniza et al., "Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait," *Journal of Biomechanics*, vol. 43, no. 4, pp. 720-726, March 2010.

[5] J.D. Qiu, S.H. Luo, J.H. Huang, X.Y Sun, and R.P. Liang, "Predicting subcellular location of apoptosis proteins based on wavelet transform and support vector machine," *AMINO ACIDS*, vol. 38, no. 4, pp. 1201-1208, April 2010.

[6] J. Luts et al., "Nosologic imaging of the brain: segmentation and classification using MRI and MRSI," *NMR in Biomedicine*, vol. 22, no. 4, pp. 374-390, May 2009.

[7] Haibin Cheng, Pang-Ning Tan, and Rong Jin, "Efficient Algorithm for Localized Support Vector Machine," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 4, pp. 537-549, April 2010.

[8] E. Blanzieri and F. Melgani, "Nearest neighbor classification of remote sensing images with the maximal margin principle," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 6, pp. 1804-1811, June 2008.

[9] Kaizhu Huang, Haiqin Yang, I. King, and M.R. Lyu, "Maxi–Min Margin Machine: Learning Large Margin Classifiers Locally and Globally," *IEEE Transactions on Neural Networks*, vol. 19, no. 2, pp. 260-272, February 2008.

[10] M.A. Aizerman, E.M. Braverman, and L.I. Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821‑837, 1964

[11] R. Fletcher, *Practical Methods of Optimization*, 2nd ed.: John Wiley and Sons, Inc., 1987.

[12] David H Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, 1992.

[13] Simon Haykin, *Neural networks; A Comprehensive Foundation*, 2nd ed., Alice Dworkin, Ed.: Prentice Hall International Inc., 1999.