Hybrid Classification Procedure Using SVM with LR on Two Distinctive Datasets

Jale BEKTAŞ
School of Applied Technology and Management
Computer Technology and Information Systems, Mersin University, Erdemli, Mersin, 33730, TURKEY
(+90) 324 515 60 23
bektasjale@gmail.com

Turgay IBRIKCI
Department of Electrical-Electronics Engineering
Cukurova University, Adana, 01330, TURKEY
(+90) 322 338 68 68
ibrikci@cu.edu.tr

ABSTRACT
Traditionally, Support Vector Machines (SVMs) are used in classification and pattern recognition, which is also the case in Neural Networks and other learning algorithms. However, a major limitation is that when the training sets are imbalanced, SVM cannot perform satisfying results by the chosen kernel. To overcome this limitation, we propose a hybrid method which uses SVM linear kernel with Logistic Regression (LR) in different manner. The hybrid method is tested over two datasets. The results showed that the hybrid approach enables us to develop an efficient algorithm, which solves the problem with all imbalanced dataset training at one time. At the same time, when it is compared with SVM that uses Radial Basis Functions (RBF) and linear kernel, better accuracy estimations are achieved with promising results on classification compared with SVM that uses RBF and linear kernel.

CCS Concepts
• **Computing methodologies→Boosting**

Keywords
Artificial intelligence; machine learning; support vector machine; logistic regression; hybrid systems.

1. INTRODUCTION
In recent years, the use of the kernel concept [1] in many machine learning problems, in which the SVM is the most prominent one, has become common. Traditionally, SVMs are used in classification and pattern recognition, which is also the case in Neural Networks and other learning algorithms. SVM can minimize the risk especially in feature space by selecting the maximum hyperplane. A good geometric definition [2] and a good generalization performance are important advantages here; however, the SVM model has some problematic steps during optimizing the model parameters and in selecting the related input features when it is performing calculations due to its faulty sides. When the SVM Model is being formed, it is necessary to optimize some important parameters that may influence its generalization performance [3]. For this reason, the hybrid SVM learning approach model has been designed to discover specific features and calculate high-dimensional inputs in an accurate manner. Furthermore, this model focuses on linear SVM that performs well on datasets that can be easily separated by a hyper-plane into two classes [4]. Choosing linear kernel allows the calculation of signed distances after the hyperplane has been created. It is difficult to classify some complex formed datasets using this kernel. These types of datasets must be transformed into high dimensional space where a linear decision boundary solves the classification problem.

In situations where the training dataset is imbalanced, when it has a top-closed geometrical structure, and where it is not separated with the defined kernel function in a linear manner, it may not show adequate performance. Generally, this problem is resolved by choosing another kernel function in an empirical way.

When solving the classification problem with a training dataset that has an imbalance form, if the training dataset is separated into smaller sub-sets in a proper manner, each sub-set gains a structure that is separable in an easier manner in the feature space. Then, an independent SVM classifier may be operated for each sub-set, which is far from the imbalance form of the whole dataset. In this way, better accuracy estimations may be achieved in the sub-sets when compared with performing all dataset training [5]. In further stages of this method, the thing that has to be done is converting the results of the classification, which was made with the sub-sets, into one single and collective result that may be decided on. For this purpose, the results of the local classification performed with sub-datasets may be merged with Logistic Regression (LR) technique. Here, the calculation of the hybrid model will be performed by SVM, and in case the size of the dataset is big, this situation will pose a benefit. In the study, all possible outcomes for all samples must be calculated; however, there must not be any situations that will increase the complexity of the calculation.

There are several studies on Kernelized Logistic Regression Method [6]. For Machine Learning, weighting process [7] is a popular method for the purpose of using more than one classifier together. In addition, there are several other studies on the ensemble methods that are used to train imbalanced datasets. However, when weighing the classifiers that have reached different accuracy rates, there have always been uncertainties. Using LR for integrating more than one classifier, finding the optimal weights based on statistical modeling theory will relieve us from this problem. This study has been applied to the Wisconsin Diagnostic Breast Cancer (WDBC) benchmark data from University of California (UCI) repository of machine learning. The empirical results have shown that the hybrid method may be used in coping with imbalanced training datasets

ICSCA 2017, February 26-28, 2017, Bangkok, Thailand
 ISBN 978-1-4503-4857-7/17/02…$15.00
DOI: http://dx.doi.org/10.1145/3056662.3056717

classification problem. In this study, hybrid algorithm has been tested for only binary datasets for now.

2. MATERIAL AND METHODS

2.1. Datasets

This hybrid method has been tested on two datasets. Datasets were taken from UCI Machine Learning Archives. The WDBC consists of 699 samples and 9 independent features that include continuous data, and one dependent feature includes class info. Other one is Dermatology dataset which consists of 366 samples, 32 independent features and one dependent feature. The features include categorical and continuous type data.

In many practical applications, it is difficult to reach a good estimation ratio in case the distribution of the dataset is not homogenous. In modern data collection techniques and in many classification applications, it is common to work with big and imbalanced datasets like healthcare services, text classifications, etc. The size of the negative examples may mostly double the positive examples. On the other hand, we do not have adequate idea on what the ratio of positive examples to negative examples should be.

Let us assume that the number of the negative examples is more than the positive samples, or the situation may be just the opposite. In such a situation, the class that has less samples may be defined as set1, and the class that has more samples may be separated into smaller datasets with clustering algorithm (e.g. subset1, subset2 …). In the next step, independent classifiers that consist of the integration of set1 and the subsets are formed. Then, the LR Model has been used to integrate and convert these independent classifiers into a result that may be decided collectively on.

2.2. SVM

SVM is a supervised learning model that has been developed to analyze and classify data. The training data is defined as points in the vector space [8]. The important thing in SVM is the finding of the hyperplane that may separate the training samples in R^N. Training samples are considered to be preassigned to two classes A or B.

We consider having *n* training samples. For eachx_i, *i = 1,..., n* is a member of class A or class B. We look for a separating rule of the form.

$$\omega^T x \geq \gamma, \quad (1)$$

where x ∈ R^N and denotes the feature vector while ω ∈ R^N, and $\gamma \in \{-1,1\}$ denotes the class labels of samples. To obtain ω and γ we solve the following linear support vector machine optimization problem:

$$minimize \frac{1}{2}\|\omega\|^2 + Ce^T y, \quad (2)$$

where $\omega \in IR^N$, $\gamma \in IR^N$, $y \in IR^N$

The optimal ω and γ of this problem yields the separating hyperplane. The data points that are defined as the support vector and that are the closest to the Hyperplane x_j are found. The signed distances are calculated with the following formula in the ω vector.

$$\omega^T = \left[\left(\sum_j \alpha_j x_j\right)^T b\right], \quad (3)$$

where *b* value shows bias.

2.3. Logistic Regression (LR)

LR is a regression analysis method that may estimate the outcome of a variable that exist in a limited number of classes [9]. By giving an input vector $x_i \in R^N$ and the output values $\gamma \in \{0,1\}$, LR may be fit with the likelihood principle that may estimate the probability of the outcome. This probability will be *p*, if $\gamma_i = 1$, or *1-p* if $\gamma_i = 0$.

$$L(\emptyset) = \prod_{i=1}^{n} (p_i)^{\gamma_i} (1 - p_i)^{(1-\gamma_i)} \quad (4)$$

It is considered mathematically more practical to calculate log of equation. The log-likelihood can be defines as follows:

$$lnL(\emptyset) = \sum_{i=1}^{n} \left(y_i ln p_i + (1 - \gamma_i) ln(1 - p_i)\right) \quad (5)$$

$L(\emptyset)$ is maximized by the change of ∅ which is called the maximum likelihood estimate. (DEV) or with a known name loss function is the negative log-likelihood and calculated in Eq.6.

$$DEV = -2lnL(\emptyset) \quad (6)$$

2.4. Hybrid Procedure

Labelled A and B are assumed to represent 2 classes in the dataset. In an imbalanced structure in which Class A has more samples than Class B in the dataset, Class A may be separated into n number subsets $(A_1, A_2, \ldots A_n)$by using clustering algorithm. The separation number *n* may be dependent on the $|A|/|B|$ ratio. In this study, the testing has been made with different *n* values in such a manner in which the $|A|/|B|$ ratio must not be too far from 1. Each I is merged with the samples in the *i =1, 2,..., n* subset with Class B, and presented to the SVM classifiers as $A_i \cup B$. If we define each classifier as $\aleph_i$, we will have *n* number classifiers in total. For this reason, each time when the algorithm is run, each classifier will form a hyperplane H_i for $\aleph_i$. In the next step, the signed distances d_k to these hyperplanes must be calculated for each *k* sample in the dataset, and a vector with *n* dimension $d = (d_{k,1}, \ldots d_{k,n})$ must be obtained. The *D* vector is prepared to be presented to the LR Model and will weigh the responses on SVM models in the study and then estimation probability is calculated. The whole framework in the hybrid algorithm is given in Figure 1.

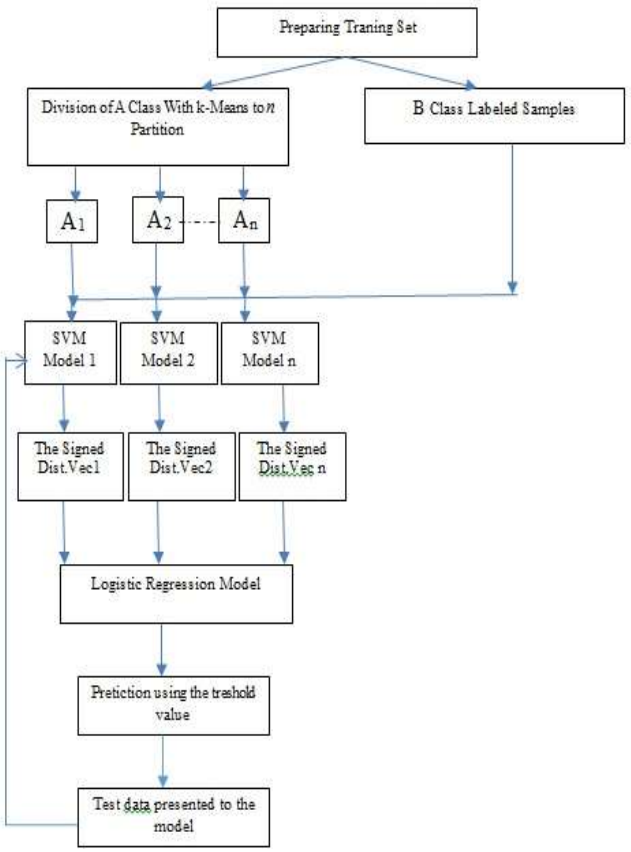

Figure 1. Hybrid algorithm procedure

The procedure starts with the separation of class samples, which are many in number, with the k-means clustering method. The k parameter has been given as 2, 3 and 4, respectively; and the results are evaluated separately. The SVM classifiers have been structured by using the merged sub-sets with the same number of *k* defined during clustering. Linear Kernel has been used for the SVM models that are trained and tested with cross validation, and the classifiers have been merged with the LR analysis. The selection of statistical model for LR will influence the performance of end-classifier at a significant level. During the model selection, in order to facilitate the interpretation and to obtain a structure that is far from being over-fit, the estimated coefficient values are examined. Then, the selected variables are fit to LR by using the binomial analysis method. Possible connections are created between each pair of the variable. In the hybrid model, the outcome of the SVMs is $d = (d_{k,1}, \ldots d_{k,n})$ vector. The linear equation obtained by the LR with this vector will be $i,\ i=1,2,\ldots,n$ *and* $b_0+b_1d_1+b_2d_2+b_3d_3+ \ldots + b_nd_n$. Here, B is the parameter vector.

3. RESULTS

The WDBC dataset has been presented to the classifiers as raw without any pre-processing steps. However, a transformation process was performed on the Dermatology dataset. Class distribution of dermatology dataset consists of 5 different values. It is not intervened to class number #1 that corresponds to 112 samples and other 254 samples are associated with unique class that is labelled with class number #2.

Clustering results are presented for each k = 2, 3, 4 and for negative samples. WDBC dataset is clustered and input data distribution for two features versus cluster number #3 is given in Figure 2.

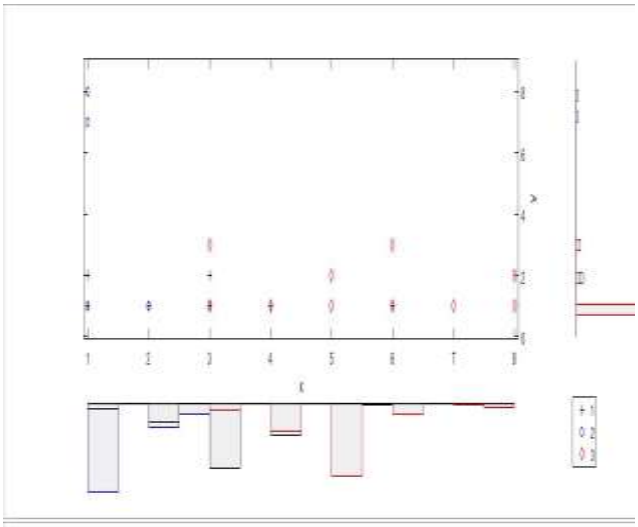

Figure 2 Typical data map in the feature space of negative samples of WDBC for k=3

For each run of simulation of Hybrid algorithm with different cluster numbers, 210 samples are tested for WDBC dataset and also 110 samples for Dermatology dataset. Each negative subset is merged with positive samples and presented to the SVM classifiers as $Xn_i \cup X_p$. If each classifier is defined as $\aleph_i$, then n number classifiers in total are obtained. The margin distance around the hyper plane is calculated as the mean distance of the support vectors. The mean distance is accepted as threshold value.

In the test dataset, prediction probabilities for all testing samples are calculated using logistic regression model. Testing sample is accepted as positive, when probability value is larger than the threshold value, otherwise the sample is labelled as negative.

To evaluate Hybrid method, false positive and false negative rates are taken into consideration as well as accuracy. Classifier performances are given in Table 1.

Table 1 Hyb.Proc. = SVM Linear + Logistic regression, SVM L = SVM classifier with linear kernel, SVM R = SVM classifier with radial base kernel function for different k's..

Method		k	TP	FN	FP	TN	ACC	Sen	Spe
Hyb. Proc	**WDBC**	2	66	1	8	135	0.96	0.89	0.99
		3	74	5	0	131	0.98	1	0.96
		4	69	1	5	135	0.97	0.93	0.99
SVM L.			66	1	7	136	0.96	0.95	0.98
SVM RBF			62	32	5	111	0.84	0.75	1
Hyb. Proc	**Dermatology**	2	23	1	12	74	0.88	0.66	0.98
		3	28	8	8	66	0.85	0.79	0.89
		4	34	2	4	70	0.94	0.89	0.97

The results are compared between Hybrid, SVM Linear and SVM RBF for WDBC dataset. It is found that Hybrid has more balanced false positive and false negative rates. It is observed to have an impact for different clusters numbers *(k = 2, 3, 4)* of the Hybrid procedure, the best balanced results were taken for *k=3*. However, Hybrid procedure evaluation metrics are better than using SVM alone with linear or RBF kernel.

Hybrid procedure is executed for the Dermatology dataset and different *k* values were evaluated as well as WDBC. ROC curves are integrated for different k's of Hybrid procedure and given in Figure 3 and Figure 4 respectively.

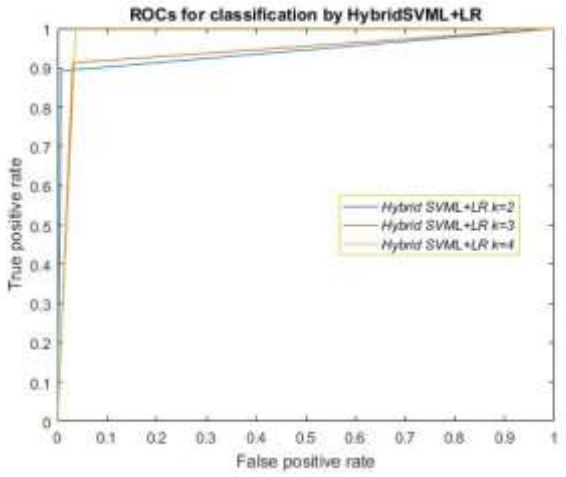

Figure 3 ROC curves for k = 2, 3, 4 of Hybrid procedure for WDBC dataset

According to the Table 1 for Dermatology dataset, the most effective results are taken for k=4. Depending on the natural structure of this dataset it is known that class number #2 has the knowledge to carry information for 4 different sample groups. It is likely that the best evaluation performance is obtained for k = 4. This suggests that subset separation value, which delivers the best result at different values for both datasets, can play an important role in the results of Hybrid procedure.

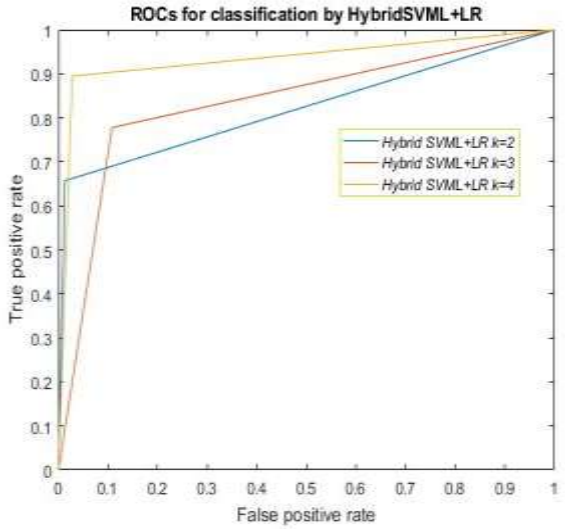

Figure 4 ROC curves for k = 2, 3, 4 of Hybrid procedure for Dermatology dataset

4. CONCLUSION

In this paper, a hybrid method which utilizes the SVM classifiers up to the number of partitions selected during clustering is used. Imbalanced datasets were preferred and the effect of the model for imbalanced datasets was measured.

For imbalanced dataset analysis, as the proportion of the samples between the classes increase, a decline in the success of the SVM classification is observed. To tackle this problem, we propose to use the procedure which provides the use of SVM with LR. The procedure uses a clustering algorithm such as k-means to separate class samples into smaller subsets which is moreover. The testing is made with different partition values in such a manner, in which the ratio of classes to each other must not be too far from 1. In this study, clusters numbers *(k = 2, 3, 4)* are chosen respectively. In the next step, independent classifiers which use the integration of the other class samples and one of the subsets are utilized. Then, the LR Model is used to integrate and convert these independent classifiers into a result that may be decided collectively on. The hybrid procedure is used to measure the evaluation metrics to display the success of the model. Linear kernel for SVM and binomial LR are used in order not to increase the complexity of the model. This approach enables us to develop an efficient algorithm, which solves the problem with all dataset training at one time. The hybrid procedure is compared with SVM that uses RBF and linear kernel. In this way, better accuracy estimations are achieved with promising results on classification when compared with SVM that uses RBF and linear kernel.

Every time the procedure is run, the probability outputs of all samples are calculated. This knowledge leads us to apply this procedure to multiclass problems. Moreover, depending on the natural structure of the datasets the class which has more samples may be separated by using different clustering algorithms rather than k-means method. The number of partition number can be found automatically. In that way, the discovery process to find the optimal partition number is shortened.

5. REFERENCES

[1] Maji, S., Berg, A. C., and Malik, J. 2013. Efficient classification for additive kernel SVMs. *IEEE transactions on pattern analysis and machine intelligence*. 35, 1 (Mar. 2012), 66-77. DOI= http://doi.org/10.1109/TPAMI.2012.62 .

[2] Barbero, A., Takeda, A., López, J. Geometric Intuition and Algorithms for Ev–SVM. *J. Mach. Learn. Res.* 16 (2015), 323-369.

[3] Chapelle, O. Choosing multiple parameters for support vector machines. *Machine learning.* 46.1 (2002), 131-159.

[4] Schmidt, B. Algorithms for Bioinformatics. Bioinformatics: High Performance Parallel Computer Architectures. CRC Press. (2010), 1-27.

[5] Chang, Y.I. Boosting SVM classifiers with logistic regression. *See "www. stat. sinica. edu. tw/library/c_tec_rep/2003-03. Pdf".* 46.1 (2003).

[6] Elbashir, M.K. Predicting beta-turns in proteins using support vector machines with fractional polynomials. *Proteome science.* 11.1 (2013), 1.

[7] Wang, Fan, W., Yu, P. S., and Han, J. 2003. Mining concept-drifting data streams using ensemble classifiers. *In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (ACM, 226-235.DOI=* http://doi.acm.org/10.1145/956750.956778.*)*

[8] Wang, L., Zhu, J., and Zou, H. 2007. Hybrid huberized support vector machines for microarray classification. *In Proceedings of the 24th international conference on Machine learning (*Corvalis, Oregon, USA — (June.2007), 20 – 24. DOI= http://dl.acm.org/citation.cfm?doid=1273496.1273620*)*

[9] Ergun, U., Serhatlioglu, S., Hardalac, F., and Guler, I. Classification of carotid artery stenosis of patients with diabetes by neural network and logistic regression. *Computers in biology and medicine. 34. 5 (2004), 389-405.*