INTERNATIONAL JOURNAL FOR TRENDS IN ENGINEERING & TECHNOLOGY
VOLUME 4 ISSUE 1 – APRIL 2015 - ISSN: 2349 - 9303

# Optimization Technique for Feature Selection and Classification Using Support Vector Machine

**M. Kumar**[1]
[1]M.E. (student)
National Engineering College
kumarec92@gmail.com

**S. Chidambaram M.E**[2]
[2]Asst.Professor, Dept of IT
National Engineering College
chidambaramraj1@gmail.com

**Dr. K G. Srinivasagan**[3]
[3] Professor & Head, Dept of CSE (PG)
National Engineering College
kgsnec@rediffmail.com

**Abstract**— Classification problems often have a large number of features in the data sets, but only some of them are useful for classification. Data Mining Performance gets reduced by Irrelevant and redundant features. Feature selection aims to choose a small number of relevant features to achieve similar or even better classification performance than using all features. It has two main objectives are maximizing the classification performance and minimizing the number of features. Moreover, the existing feature selection algorithms treat the task as a single objective problem. Selecting attribute is done by the combination of attribute evaluator and search method using WEKA Machine Learning Tool. We compare SVM classification algorithm to automatically classify the data using selected features with different standard dataset.

**Index Terms— Data Mining, Kernel methods, Support Vector Machine, WEKA, Classification.**
.

## 1 INTRODUCTION

Support Vector Machine (SVM) was first described in 1992, introduced by Boser, Guyon, and Vapnik. Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression [1]. They belong to a family of linear classifiers. In other words, Support Vector Machine (SVM) is a tool of classification and regression prediction that make use of the machine learning theory to increase the predictive accuracy by automatically avoiding the data over-fit. Support Vector machines can also be defined as systems that uses the hypothesis space of linear functions in a multi dimensional feature space, trained with a learning protocols from optimization theory that implements a learning bias derived from the statistical learning . SVM becomes famous when, using pixel maps as input; it gives accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task [2]. It is also being used for many applications, such as hand writing, face, texture analysis and so on, especially for classification and regression based applications. The fundamentals of Support Vector Machines (SVM) has been developed by Vapnik [3] and gained popularity due to many promising features such as better empirical performance. SVM attempts to minimize a bound on the generalization error (i.e., the error made by the learning machine on test data not used during training).

## 2 FEATURE SELECTION

Feature selection (also known as subset selection or variable selection) is a process commonly employed in machine learning to solve the high dimensionality problem. It select a subset of important features and removes irrelevant and redundant features for simpler and more concise data. The advantages of feature selection are multi-fold. First, feature selection greatly saves the running time of a learning process by removing irrelevant, redundant and noisy features. Second, without the interaction of irrelevant and redundant features, learning algorithms can focus on the most important perspective of the data and build accurate data models. Thus, the classification performance can be improved. Third, feature selection can help us to build a simpler and more generic model and provide a better insight into the underlying concept of the task.

### 2.1 Feature Selection Objectives

Different feature selection algorithms may have various objectives to achieve. The following is a list of common objectives used by researchers:

1. Find the minimally sized feature subset that is necessary and sufficient to the target concept.

2. Select a subset in a set of N features from a set of M features,

N < M, such that the value of a criterion function is optimized over all subsets of size N.

3. Choose a subset of features for improving prediction accuracy or decreasing the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features.

4. Select a small subset such that the resulting class distribution, which provides the values of the selected features, is as close as possible to the class distribution for all given feature values.

### 2.2 Feature Selection Procedure

The process begins with subset generation that employs a certain search strategy to produce feature subsets. Then each subset is evaluated according to an evaluation criterion and compared with the previous best result. If the evaluated subset is better, then it replaces the previous one, else remains the same. The process of subset generation and evaluation gets repeated until it satisfies the given criteria. Figure 1 show the Feature Selection procedure.

Finally the selected best subset is validated by using the prior knowledge or some test data. Search strategy and evaluation criteria are two key points in the study of feature selection.

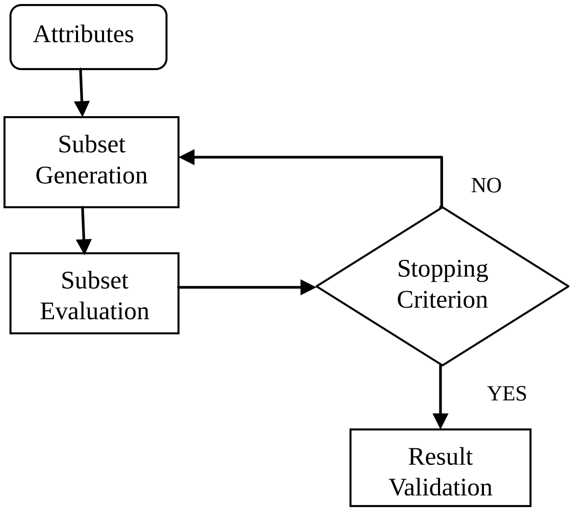


Fig 1. Feature Selection Procedure

1 SUBSET GENERATION

Subset generation begins with a search start point, which can be of any one of the following an empty set, the full set, or a randomly generated subset. From the starting point, it can search feature subsets in different directions, such as forward, backward, and random. In forward search, features gets added at a time, while in backward search the least important feature is removed based on evaluation criterion. Random search adds or deletes features at random to avoid being trapped into local maxima. There are various search strategies for finding an optimal or suboptimal feature subset.

2 EVALUATION CRITERIA

After the generation of feature subsets, the results are evaluated by certain criteria to measure their goodness. Generally, the goodness of feature subsets means that the discriminating ability of subsets that distinguish among or within different classes. Based on their dependent in the learning algorithms, this algorithms can be broadly classified into Wrapper Method, Filter Method. Wrapper method is used to find feature subsets with high accuracy because the features match well with the learning algorithms.

3. STOPPING CRITERIA

A feature selection process gets terminated under any one of the following criteria:

i. Whether the search is complete.

ii. Whether a predefined size of feature subsets is selected.

iii. Whether a predefined number of iterations are executed.

iv. Whether an optimal or sufficiently good feature subset according to the evaluation function has been obtained.

v. Whether the change (addition or deletion of features) of feature subsets does not produce a better subset.

4. RESULT VALIDATION

In some applications, the relevant features are known beforehand. Then we can able to validate the feature selection results by comparing the prior knowledge. However, in the real-world applications we do not know which features are relevant.

We have to use the classification performance on test data as an indicator of the goodness of the selected feature subsets.

Once the feature selection is done the selected features are used for classification. In Attribute Selection module irrelevant features are discard and the selected feature is used for classification in order to improve better performance. Feature Selection procedure is shown in the figure 2. WEKA Machine learning tool is used for feature selection and classification for the given datasets.

Attribute selection involves searching through all possible combination of attributes in the data to find which subset of attribute work best for improving performance of classification. WEKA has a very flexible combination of search and evaluation methods for the dataset's attributes. It consists of Search methods and Attribute Evaluator.

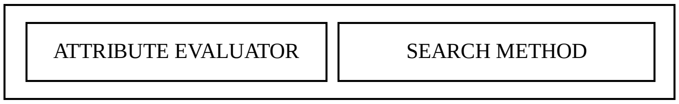


Fig 2. Attribute Selection

### 2.3 ATTRIBUTE EVALUATOR

The Attribute Evaluator is the method by which a subset of attributes is assessed. For example, they may be assessed by building a model and evaluating the accuracy of the model. ClassifierSubsetEval is used as an attribute evaluator to access the attributes of given datasets. It Evaluates attribute subsets on training data or a separate hold testing set.

Option in ClassifierSubsetEval

•classifier -- Classifier to use for estimating the accuracy of subsets

•holdOutFile -- File containing hold out/test instances.

•useTraining -- Use training data instead of hold out/test instances.

### 2.4 SEARCH METHOD

The Search Method is the is the structured way in which the search space of possible attribute subsets is navigated based on the subset evaluation. Search method traverses the attribute space to find a good subset. Quality of the subset is measured by the attribute subset evaluator. BestFirst search method is used to find the possible search space of the attributes.

BestFirst searches the space of attribute subsets by greedy hill climbing augmented with a backtracking ability. Setting the number of consecutive non-improving nodes will allow the level of backtracking. Best first may start with the empty set of attributes and proceed forward, or start with the full set of attributes and proceed backward, or start at any point and proceed in both the direction. The combination of best first search method and ClassifierSubsetEval will give the best feature selection in result which is further used for improving classification accuracy.

Option in BestFirst

•Direction -- Set the direction of the search.

•LookupCacheSize -- Set the maximum size of the lookup cache of evaluated subsets. This can be expressed as the multiplier of the given number of attributes in the given data set. (default = 1).

•Search Termination -- Set the amount of backtracking.

## 3 LIBSVM CLASSIFICATION

A wrapper class is one of the libsvm tools (the libsvm classes that is the jar file need to be in the path of the class to use this classifier).LibSVM executes faster than SMO since it uses LibSVM to develop the SVM classifier. LibSVM allows users to do experiment with One-class, Regressing , and nu-SVM supported by LibSVM tool. LibSVM reports many useful information about LibSVM classifier (e.g., confusion matrix, precision, recall, ROC, etc.).

The goals of SVM are to separate the data along with its hyper plane and extend ed this to non-linear boundaries using kernel trick [8] [5]. For calculating the SVM we see that the goal is to correctly classify all the data. For mathematical calculations we have,

[a] If Zi= +1; $vx_i + b \geq 1$ (1)

[b] If Zi= -1; $vx_i + b \leq 1$ (2)

[c] For all Zi; $vx_i + b = 1$ (3)

In this equation x is a vector point and w is weight and is also a vector. So to separate the data [a] should always be greater than that of zero. Among all possible hyper planes, SVM selects the one plane where the distance of that hyper plane should be larger as possible. If the training data is good then the every test vector will locate in radius r from training vector. Hence the chosen hyper plane is located at the farthest distance from the data. This desired hyper plane that maximizes the margin also bisects the lines between closest points on convex hull of the two datasets.

Distance from closest point on hyper plane to origin can be found by maximizing the value of x where x is on that hyper plane. Similarly for the other side points we have a same scenario. Thus by solving and subtracting the two distances we get the summed distance from the separated hyper plane to the nearest points. Maximum Margin is

$M = 2 / \|v\|$ (4)

### 3.1 Kernel Trick

If data is linear, then to divide the data the separating hyper plane is used. However it shows that the data is far away from linear and the datasets are mostly inseparable. Thus the kernels are used to map the non-linear input data to a high-dimensional space. It shows that new mapping is then linearly separable.

$K(\mathbf{x_i},\mathbf{x_j}) = \varphi(\mathbf{x_i})^{T}\varphi(\mathbf{x_j})$ (5)

**Feature Space:** Transforming the data into feature space makes it possible to define a similarity measure on the basis of the dot product. If the space of the feature is chosen suitably,then pattern recognition can be easy [1].

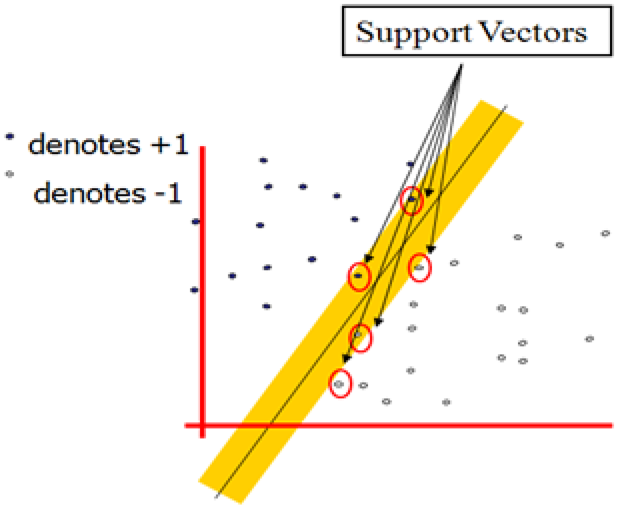
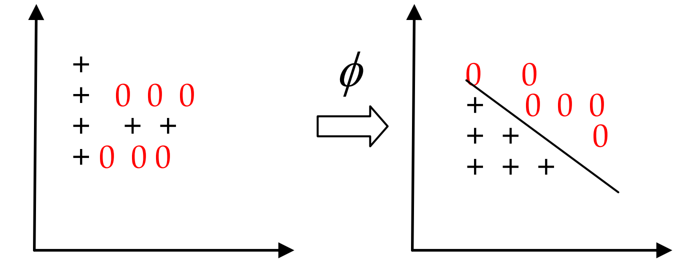


Fig 3. Representation of Support Vectors



(4a) Linearly inseparable (4b) Linearly separable

The Kernel trick allows SVM's to form boundaries of nonlinear. Steps followed in kernel trick are given below.

[a] The algorithm is expressed using the data sets inner products. This is also knows as dual problem.

[b] Original data are traveled through nonlinear maps to get new data with respect to new dimensions by adding a pair wise product of some of the original data dimension to each data vector.

[c] Rather than an inner product on these new, large vectors, and placed in tables and later do a lookup table, we can provide a dot product after doing nonlinear mapping on them of the data. This function is called as kernel function. More kernel functions are as fallows.

The kernel function plays a critical role in SVM and also on its performance. It is based upon reproduction Spaces of Kernel Hilbert [8].Then such kernel represents a legitimate product of inner in the space of feature. The training set is not separable linearly in a space of input. The training set is separable linearly in the space of feature. This is called as the "Kernel trick".

The different kernel functions are listed below where r --coef0 value coef0 for polynomial kernel -r --coef0 value coef0 for sigmoid kernel. d --degree value degree for polynomial kernel. g --gamma value gamma for polynomial kernel. g --gamma value gamma for rbf-kernel. g --gamma value gamma for sigmoid kernel default 0 is set for all gamma values. u & v are kernel parameter user defined.

TABLE I. KERNEL METHODS

| Kernel name | Representation |
|---|---|
| Linear | u'*v |
| Polynomial | (gamma*u'*v+ coef0)^degree |
| Gaussian(RBF) | exp(-gamma*\|u-v\|^2) |
| Sigmoid | tanh(gamma*u'*v + coef0) |

## 4 EXPERIMENT AND RESULT

This section describes the experimental results obtained by applying the proposed algorithms to a variety of data sets. For experimentation, five benchmark datasets are taken from the UCI machine learning repository [9] as shown in Table 2. In order to validate the prediction results of the comparison of the classification techniques and the 10-fold crossover validation is used. The present study divided the data into 10 folds where 1 fold was for testing and 9 folds were for training for the 10-fold crossover validation.

TABLE II. PROVIDES THE ATTRIBUTE INFORMATION OF FIVE DATASETS

| Datasets | Features | Instances | Class |
|---|---|---|---|
| **Iris** | 5 | 150 | 3 |
| **Ecoli** | 8 | 336 | 8 |
| **Mushroom** | 23 | 8124 | 2 |
| **Diabetes** | 9 | 768 | 2 |
| **Breast cancer** | 10 | 286 | 2 |

WEKA is a collection of machine learning algorithms used for data mining tasks. The algorithms can either be directly applied to a dataset or called into the java code.

WEKA contains tolls for data pre-processing the data, classification of data, regression, data clustering, association rules and visualization.

WEKA (Waikato Environment for Knowledge Analysis) is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. WEKA is the art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. The algorithms are applied directly to a dataset.UCI data sets are used for both feature selection and classification using LibSVM in WEKA and their performance is compared.

ROC curve is plot by selecting the visualize threshold cure in WEKA classifier result buffer. It is one of the performance metrics for the classification algorithm to know its efficient prediction.

TP rate, FP rate, Precision, recall, F-measure and ROC area are the other performance metrics to know the LibSVM performance. These metrics are used to form the confusion matrix.

### 4.1Performance Metrics

Three classical evaluation metrics of Precision data, Recall and F-score are used to validate the efficiency of the proposed method. The three metrics are traditionally defined as a classification of binary task with both positive and negative classes. Precision is the positive predictions proportion that are correct, and recall is the positive samples proportion that are correctly positively predicted. That is:

**Precision** = $TP/\ (TP+FP)$
**Recall** = $TP/\ (TP+FN)$
**F = score** = $2 * Precision * RecallPrecision + Recall$

- True positive (TP) = number of positive samples correctly predicted.
- False negative (FN) = number of positive samples wrongly predicted.
- False positive (FP) = number of negative samples wrongly predicted as positive.
- True negative (TN) = number of negative samples correctly predicted.

TABLE III. ACCURACY VALUES AND ERROR VALUES FOR ORIGINAL DATASETS USING LIBSVM CLASSIFIER

| Dataset | Correctly classified Instances (Accuracy %) | Incorrectly classified Instances (Error %) |
|---|---|---|
| **Iris** | 96.66 | 3.333 |
| **Ecoli** | 75.5 | 24.40 |
| **Mushroom** | 99.803 | 0.1669 |
| **Diabetes** | 65 | 34 |
| **Breast cancer** | 70 | 29 |

Table3 shows the various datasets with correctly and incorrectly classified instances after applying LibSVM algorithm to the data.

Firstly, SVM has nonlinear dividing hyperplane, which give high discrimination. Secondly, SVM provides good generalization ability for unseen data classification these properties of SVM used to solve the non-linear problem more efficiently.

TABLE IV. Result for Datasets Using Libsvm Classifier

| **Dataset** | **TP rate** | **FP rate** | **Precision** | **Recall** | **F-measure** | **ROC Area** |
|---|---|---|---|---|---|---|
| **Iris** | 0.967 | 0.017 | 0.967 | 0.967 | 0.967 | 0.975 |
| **Ecoli** | 0.756 | 0.087 | 0.622 | 0.756 | 0.681 | 0.613 |
| **Mushroom** | 0.998 | 0.002 | 0.996 | 0.998 | 0.998 | 0.998 |
| **Diabetes** | 0.651 | 0.651 | 0.424 | 0.651 | 0.513 | 0.500 |
| **Breast cancer** | 0.706 | 0.688 | 0.695 | 0.706 | 0.594 | 0.509 |

Table4 shows the LibSVM classification result for different data sets.

## 5 Conclusion

This paper proposed a SVM based optimization algorithm, which can used to optimize the parameter SVM values and also used to obtain the features of optimal subset by applying LIBSVM method to remove insignificant features and effectively find best parameter values.

The goal of this paper is to design Support Vector Machine Algorithm to find the classification accuracy. Feature Selection algorithm is used to improve the classification accuracy of classifier with respect to datasets.

## References

[1] Support-vector networks. Corinna Cortes and Vladimir Vapnik, Machine Learning, 1995, Volume 20, Number 3, Pages 273-297.

[2] Yuan-HaiShao, Wei-JieChen , Jing-JingZhang “An efficient weighted Lagrangian twin support vector machine for imbalanced data classification” journal homepage :www.elsevier.com/locate/pr.

[3] Winda Astuti, Rini Akmeliawati, Wahju “Hybrid Technique Using Singular Value Decomposition (SVD) and Support Vector Machine(SVM) Approach for Earthquake Prediction” IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing, Vol. 7, No. 5, May 2014.

[4] Kanu Patel, Jay Vala, Jaymit Pandya “Comparison of various classification algorithms on iris datasets using WEKA” International journal of Advance Engineering and Research Development (IJAERD) Volume 1 Issue 1, February 2014, ISSN: 2348 – 4470.

[5] J.Platt “Fast training of SVMs using sequential minimal optimization”, In B. Sch¨olkopf, C.Burges and A.Smola (ed.), Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA, 1999, 185-208.

[6] O.L.Mangasarian and D.R.Musicant “Lagrangian support vector machines”, Journal of Machine Learning Research, 1, 2008, 161-177.

[7] Z.Nematzadeh Balagatabi1, H.Nematzadeh Balagatabi “Comparison of Decision Tree and SVM Methods in Classification of Researcher’s Cognitive Styles in Academic Environment” Indian Journal of Automation and Artificial Intelligence Vol: 1 Issue: 1 January 2013 ISSN 2320 – 4001.

[8] Y.Tang,Y.-Q.Zhang,N.V.Chawla,S.Krasser,”SVMS modeling for highly imbalanced classification”, IEEE Trans.Syst.ManCybern.B:Cybern.39(1) (2009) 281–288.

[9] G. Fung and L.Mangasarian “Proximal support vector machine classifiers”, Proceedings of the 7th ACM conference on knowledge discovery and data mining, *ACM, 2001, 77-86.*

[10] Y.Lee, O.Mangasarian, W.Wolberg, “Survival-time classification of breast cancer patients”, Computational Optimization and Applications, 25(2003), 151-166.

[11] Z.Q. Qi, Y.J. Tian, Y. Shi,” Robust Twin Support Vector Machine for Pattern Classification”, Pattern Recognition, 2013, 46(1): 305-316.

[12] Y. Tian, Y. Shi, X. Liu, “Recent advances on support vector machines research”, Technological and Economic Development of Economy 18(1) (2012), 5-33.

[13] Z. Qi, Y. Tian, and Y. Shi, “Structural Twin Support Vector Machine for Classification, Knowledge-Based Systems, 2013, DOI:10.1016/j.knosys.2013.01.008.

[14] S. Maldonado, R. Weber, J. Basak,” Kernel-penalized SVM for feature selection”, Inform. Sci. 181 (2011) 115–128.

[15] L. Song, A. Smola, A. Gretton, J. Bedo, K. Borgwardt, “Feature selection via dependence maximization”, J. Mach. Learn. Res. 13 (2012) 1393–1434.

[16] G. Victo Sudha George, V. Cyril Raj, “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) (2011) 16–27.

[17] H. Yu, J. Kim, Y. Kim, S. Hwang, Y.H. Lee, “An efficient method for learning nonlinear ranking SVM functions”, Inform. Sci. 209 (2012) 37–48.

[18] S. Maldonado, J. López, “Imbalanced data classification using second-order cone programming support vector machines”, Pattern Recogn. 47 (2014).

[19] E. Carrizosa, B. Martín-Barragán, D. Romero-Morales, “Detecting relevant variables and interactions in supervised classification”, Euro. J. Oper. Res. 213 (2011) 260–269.

[20] R. Hassan, R.M. Othman, P. Saad, S. Kasim, “A compact hybrid feature vector for an accurate secondary structure prediction”, Inform. Sci. 181 (2011)5267–5277.