ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 10, December 2012

A Review on Support Vector Machine for Data Classification

Himani Bhavsar, Mahesh H. Panchal

Abstract-- **With increasing amounts of data being generated by businesses and researchers there is a need for fast, accurate and robust algorithms for data analysis. Improvements in databases technology, computing performance and artificial intelligence have contributed to the development of intelligent data analysis. Support vector machines are a specific type of machine learning algorithm that are among the most widely-used for many statistical learning problems, such as spam filtering, text classification, handwriting analysis, face and object recognition, and countless others. Support vector machines have also come into widespread use in practically every area of bioinformatics within the last ten years, and their area of influence continues to expand today. The support vector machine has been developed as robust tool for classification and regression in noisy, complex domains. The two key features of support vector machines are generalization theory, which leads to a principled way to choose an hypothesis; and, kernel functions, which introduce non-linearity in the hypothesis space without explicitly requiring a non-linear algorithm.**

This paper highlight the advantages of SVM over existing data analysis techniques, also are noted some important points for the data mining practitioner who wishes to use support vector machines.

Index Terms **Data classification, Support Vector Machine, Kernel functions.**

I. INTRODUCTION

The Data mining is the process of extracting patterns from data. Data mining is the process of discovering knowledge from large amounts of data stored either in databases or warehouses[14]. Data mining is becoming an increasingly important tool to transform these data into information. Data mining can also be referred as knowledge mining or knowledge discovery from data. Classification is a data mining (machine learning) technique used to predict group membership for data instances and Support Vector Machine is used for Linear and Non Linear classification.

Support Vector Machines (SVM) is a powerful; state-of-the-art algorithm with strong theoretical foundations based on the Vapnik-Chervonenkis theory.SVM has strong regularization properties. Regularization refers to the generalization of the model to new data. Support vector machines in particular were designed as a tool to solve supervised learning classification problems[3]. Su[1]pervised learning is an area of machine learning in which we are given some "training set" of data for which we know a priori the appropriate classifications.

The geometrical interpretation of support vector classification (SVC) is that the algorithm searches for the optimal separating surface, i.e. the hyperplane that is, in a sense, equidistant
from the two classes.SVC is outlined first for the linearly separable case. Kernel functions are then introduced in order to construct non-linear decision surfaces. Finally, for noisy data, when complete separation of the two classes may not be desirable, slack variables are introduced to allow for training errors.

This paper is organized as follow. Section 2 ,contain introduction of this paper and related background of Data classification techniques. In section 3,4 and 5, it shows some basic concept of SVM along with Linear and Non Linear Classification .The section 6 will show importance of kernel selection when you are using SVM for classification. Finally we have some conclusion in section 7.

II. CLASSIFICATION WITH DATA MINING

Data mining is the discovery of knowledge and useful information from the large amounts of data stored in databases. It is referred to as knowledge discovery from databases (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve.
Classification techniques are widely used in data mining to classify data among various classes. Classification techniques are being used in different industry to easily identify the type and group to which a particular tuple belongs. Classification is a data mining (machine learning) technique used to predict group membership for data instances. Classification is a two step process. 1st step is Model Construction and 2nd step is Model Usage. There are many algorithms which are used for classification in data mining.
Following are some classification techniques:

1) Decision tree induction
   - Decision tree classification is the learning of decision trees from class labeled training tuples [14]. A decision tree is a flowchart like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label[14].
2) Rule based classifier
   - Represent the knowledge in the form of IF-THEN rules and One rule is created for each

Himani Bhavsar-*Department of Information Technology, KITRC, Kalol. Ahmedabad, India.*
Mahesh H. Panchal-*Head, Department of Computer Engineering, KITRC, Kalol, Ahmedabad, India.*

path from the root to a leaf. Rules are easier to understand than large trees.

3) Bayesian classifier
   - A statistical classifier: performs probabilistic prediction, *i.e.*, predicts class membership probabilities
4) Artificial neural network
   - Neural Networks are analytic techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called *learning* from existing data.
5) Nearest neighbor Classifier
   - The k-nearest neighbor's algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning. It can also be used for regression.
6) Support vector machine
   - A new classification method for both linear and nonlinear data and SVMs are a set of related supervised learning methods used for classification and regression[14].
7) Ensemble classifier
   - Use a combination of models to increase accuracy and Combine a series of k learned models, M_1, M_2, …, M_k, with the aim of creating an improved model M*

III. SUPPORT VECTOR MACHINE:

Support Vector Machines were first introduced to solve the pattern classification and regression problems by Vapnik and his colleagues.

3.1 Overview of SVM

SVMs are set of related supervised learning methods used for classification and regression[2]. They belong to a family of generalized linear classification. A special property of SVM is , SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM called Maximum Margin Classifiers. SVM is based on the Structural risk Minimization (SRM). SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be [2].We consider data points of the form {(x1,y1),(x2,y2),(x3,y3),….,(xn, yn)}.
Where yn=1 / -1 , a constant denoting the class to which that point xn belongs. n = number of sample. Each x n is p-dimensional real vector. The scaling is important to guard against variable
(attributes) with larger variance. To view this Training data , by means of the dividing (or
separating) hyperplane , which takes
w . x + b = o ----- (1)
Where b is scalar and w is p-dimensional Vector. The vector w points perpendicular to the separating hyperplane . Adding the offset parameter b allows us to increase the margin. Absent of b, the hyperplane is forced to pass through the origin , restricting the solution. As with the interest in the maximum margin , we are interested in SVM and the parallel hyperplanes. Parallel hyperplanes can be described by equation
w.x + b = 1
w.x + b = -1
If the training data are linearly separable, we can select these hyperplanes so that there are no points between them and then try to maximize their distance. By geometry, We find the distance between the hyperplane is $2 / |w|$. So we want to minimize $|w|$. To excite data points, we need to ensure that for all I either
$w. xi – b \geq 1$ or $w. xi – b \leq -1$
This can be written as
$yi (w. xi – b) \geq 1 , 1 \leq i \leq n$ ------(2)

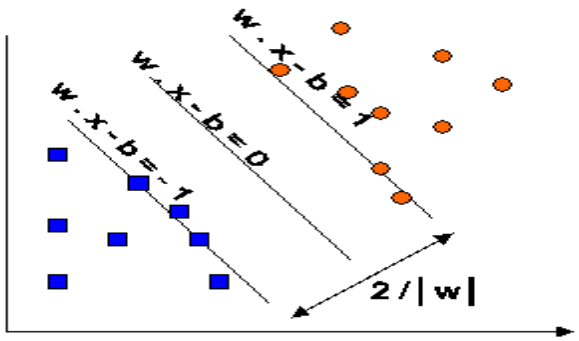

Figure 1 Maximum margin hyperplanes for a SVM trained with samples from two classes

SVMs fall into the intersection of two research areas: kernel methods, and large margin classifiers. SVM has been applied to feature selection, time series analysis, reconstruction of a chaotic system, and non-linear principal components. Further advances in these areas are to be expected in the near future. SVMs and related methods are also being increasingly applied to real world data mining.

3.2 Classification in SVM

We can classify linearly separable and non-linear separable data using Support Vector Machine.

3.2.1 Linear Classification

Before considering *N*-dimensional hyperplanes, let's look at a simple 2-dimensional example. Assume we wish to perform a classification, and our data has a categorical target variable with two categories. Also assume that there are two

predictor variables with continuous values. If we plot the data points using the value of one predictor on the X axis and the other on the Y axis we might end up with an image such as shown below. One category of the target variable is represented by rectangles while the other category is represented by ovals.

In this idealized example, the cases with one category are in the lower left corner and the cases with the other category are in the upper right corner; the cases are completely separated. The SVM analysis attempts to find a 1-dimensional hyperplane (i.e. a line) that separates the cases based on their target categories. There are an infinite number of possible lines; two candidate lines are shown above. The question is which line is better, and how do we define the optimal line.

The dashed lines drawn parallel to the separating line mark the distance between the dividing line and the closest vectors to the line. The distance between the dashed lines is called the *margin*. The vectors (points) that constrain the width of the margin are the *support vectors*. The following figure illustrates this.

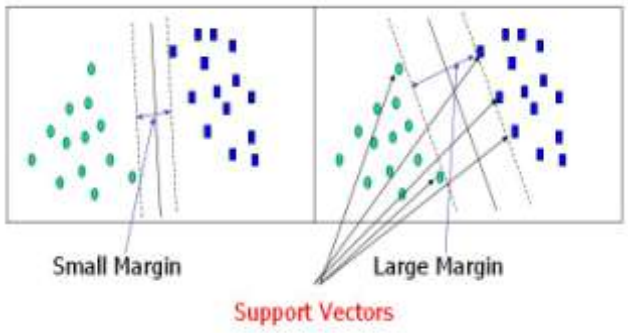

Figure 2 Two Dimensional Classification[16]

An SVM analysis finds the line (or, in general, hyperplane) that is oriented so that the margin between the support vectors is maximized. In the figure above, the line in the right panel is superior to the line in the left panel.

If all analyses consisted of two-category target variables with two predictor variables, and the cluster of points could be divided by a straight line, life would be easy. Unfortunately, this is not generally the case, so SVM must deal with (a) more than two predictor variables, (b) separating the points with non-linear curves, (c) handling the cases where clusters cannot be completely separated, and (d) handling classifications with more than two categories.

3.2.2 Non Linear Classification

The simplest way to divide two groups is with a straight line, flat plane or an N-dimensional hyperplane. But what if the points are separated by a nonlinear region such as shown below:

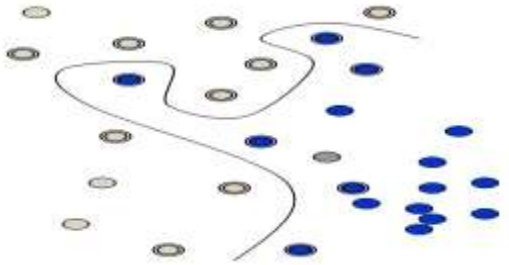

Figure 3 Multi Dimensional Classification[16]

In this case we need a nonlinear dividing line. Rather than fitting nonlinear curves to the data, SVM handles this by using a *kernel function* to map the data into a different space where a hyperplane can be used to do the separation.

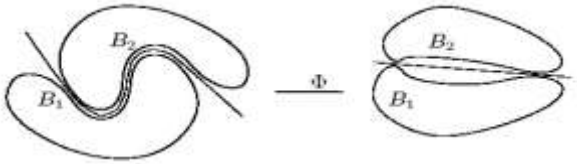

Figure 4 Kernel Function Mapping[16]

The kernel function may transform the data into a higher dimensional space to make it possible to perform the separation.

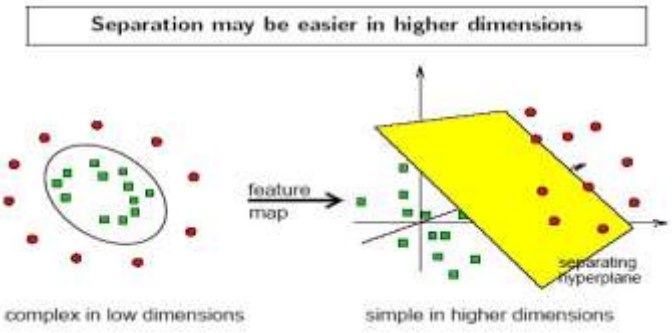

Figure 5 Higher Dimensional Mapping[16]

IV. KERNELS IN SVM

The kernel function may transform the data into a higher dimensional space to make it possible to perform the separation. Kernel functions are a class of algorithms for pattern analysis or recognition, whose best known element is the support vector machine (SVM).Training vectors xi are mapped into a higher (may be infinite) dimensional space by the function Φ. Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimension space .C > 0 is the penalty parameter of the error term.

Furthermore, K(xi , xj) ≡ Φ(xi)T Φ(xj) is called the kernel function. There are many kernel functions in SVM, so how to select a good kernel function is also a research issue[1].

The choice of a Kernel depends on the problem at hand because it depends on what we are trying to model. A polynomial kernel, for example, allows us to model feature conjunctions up to the order of the polynomial. Radial basis functions allows to pick out circles - in contrast with the Linear kernel, which allows only to pick out lines (or hyperplanes).

Many kernel mapping functions can be used – probably an infinite number. We can use Normalized Polynomial, RBF, linear, Sigmoid, GussianRBF and String Kernels etc based on application requirement. But a few kernel functions have been found to work well in for a wide variety of applications. The default and recommended kernel function is the Radial Basis Function (RBF).

4.1 Linear kernel:

$$K(xi, xj) = xi^T xj + C$$

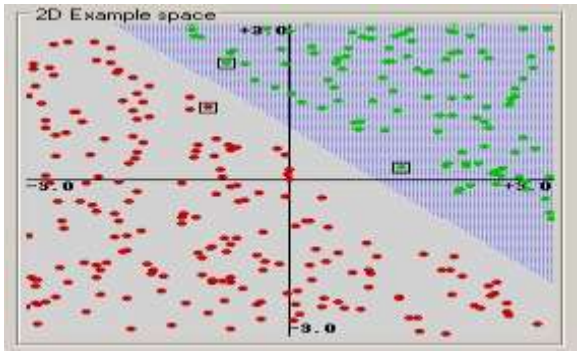

Figure 6 Linear Kernel Classification[16]

The Linear kernel is the simplest kernel function. It is given by the inner product <x,y> plus an optional constant **c.**

4.2 Polynomial:

$$K(xi, xj) = (\gamma\, xi^T xj + r)^d, \gamma > 0$$

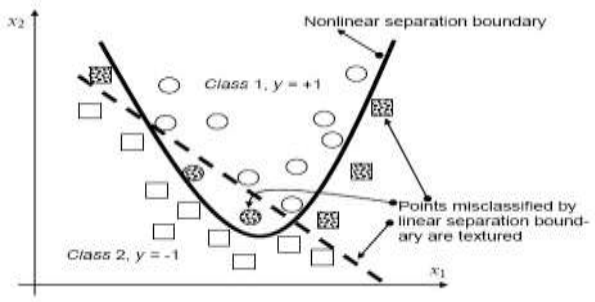

Figure 7 Polynomial kernel Classification[16]

The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized.

4.3 Radial basis function:

$$K(xi, xj) = \exp(-\gamma \|xi - xj\|^2), \gamma > 0$$

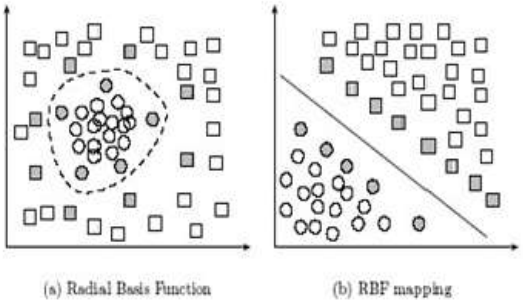

Figure 8 Separable classification with RBF kernel[16]

Here, γ, r and d are kernel parameters. In these popular kernel functions, RBF is the main kernel function because of following reasons[2]:

1. The RBF kernel nonlinearly maps samples into a higher dimensional space unlike to linear kernel.
2. The RBF kernel has less hyper parameters than the polynomial kernel.
3. The RBF kernel has less numerical difficulties.

V. Limitations of SVM:

- The biggest limitation of SVM lies in the choice of the kernel (the best choice of kernel for a given problem is still a research problem).
- A second limitation is speed and size (mostly in training - for large training sets, it typically selects a small number of support vectors, thereby minimizing the computational requirements during testing).
- The optimal design for multiclass SVM classifiers is also a research area.

VI. Conclusion:

The support vector machine has been introduced as a robust tool for many aspects of data mining including classification, regression and outlier detection. The SVM uses statistical learning theory to search for a regularized hypothesis that fits the available data well without over-fitting. The SVM has very few free parameters, and these can be optimized using generalization theory without the need for a separate validation set during training. It can be seen that the choice of kernel function and best value of parameters for particular kernel is critical for a given amount of data.

VII. REFERENCES:

[1]. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. "A Practical Guide to Support Vector Classification" . Dept. of Computer Sci.National Taiwan Uni, Taipei, 106, Taiwan http://www.csie.ntu.edu.tw/~cjlin 2007.
[2]. V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.
[3]. N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge University Press, 2000.
[4]. T. Joachims. Making large-scale support vector machine learning practical. In Advances in Kernel Methods: Support Vector Learning. B. Schölkopf, C.J.C. Burges, and A.J. Smola (Eds.), MIT Press, 1998.
[5]. O. Chapelle and V. Vapnik. Model selection for support vector machines. In Proceedings of the Twelfth Conference on Neural Information Processing Systems. S.A. Solla, T. K. Leen, and K.-R. Müller (Eds.), MIT Press, 1999.
[6]. T. Mitchell. Machine Learning. McGraw-Hill International, 1997.
[7]. E. Osuna, R. Freund, and F. Girosi. Support vector machines: training and applications. AI Memo 1602, MIT, May 1997.
[8]. Byvatov, E. and Schneider, G. (2003). Support vector machine applications in
bioinformatics. Appl Bioinformatics, 2(2):67-77.
[9]. J. W. Shavlik and T. G. Dietterich. Readings in Machine Learning. Morgan Kaufmann, 1990.
[10]. I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques, 2ed. Morgan Kaufmann, 2005
[11]. T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, 2001.
[12]. H. Yu, J. Yang, and J. Han. Classifying large data sets using SVM with hierarchical clusters. KDD'03
[13]. S. M. Weiss and N. Indurkhya. Predictive Data Mining. Morgan Kaufmann, 1997.
[14] Lyman, Peter; Hal R. Varian (2003). "How Much Information"
[15] Jiawei Han, Michelin Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, USA, 2001, 70-181.
[16]. Source: www.dtreg.com/svm.htm