Data Classification using Support Vector Machines

Kamini Nalavade[1], B.B. Meshram[2]
[1]Research Scholar, [2]professor,
[1,2]VJTI, Matunga, Mumbai
[1]knalavade@yahoo.com, [2]bbmeshram@vjti.org.in

ABSTRACT
Classifying data is a common task in machine learning. In machine learning, statistical classification is the problem of identifying the sub-population to which new observations belong on the basis of a training set of data containing observations whose sub-population is known. Therefore these classifications will show a variable behavior which can be studied by statistics. In machine learning, the classification problem is known as supervised learning, while clustering is known as unsupervised learning. SVM's analyze data and recognize patterns with support vector methods. This paper is a comprehensive study of SVM fundamentals, research and how it is used for anomaly detection. The paper discusses the mathematical modeling on which SVM is based on. We also discuss tools and algorithms which can be used to perform classification using SVM.

Keywords
Network, Machine learning, Classification, Support Vectors

1. Introduction

With the global Internet connection, network security has gained significant attention in the research and industrial communities. Currently network security components like Firewalls, Antivirus are unable to stop attackers. An intrusion detection system gathers and analyzes information from various areas within a computer or a network to identify possible security breaches. Anomaly detection systems have two major advantages over signature based intrusion detection systems. The first advantage that differentiates anomaly detection systems from signature detection systems is their ability to detect unknown attacks as well as "zero day" attacks. This advantage is because of the ability of anomaly detection systems to model the normal operation of a system/network and detect deviations from them. Machine learning techniques enable the development of anomaly detection algorithms that are non-parametric, adaptive to changes in the characteristics of normal behaviour in the relevant network and portable across applications. Researchers have begun to use machine learning techniques to detect outliers in datasets from a variety of fields. Gardener et al. use a One-Class Support Vector Machine (OCSVM) to detect anomalies in EEG data from epilepsy patients [2].

SVMs were developed by Cortes & Vapnik (1995) for binary classification. A support vector machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input, making the SVM a non-probabilistic binary linear classifier.

The paper is organized as below; the second section examines the extensive literature in the intrusion prevention domain while third section is dedicated to our proposed model for the data & network security using data mining and the fourth section concludes the results

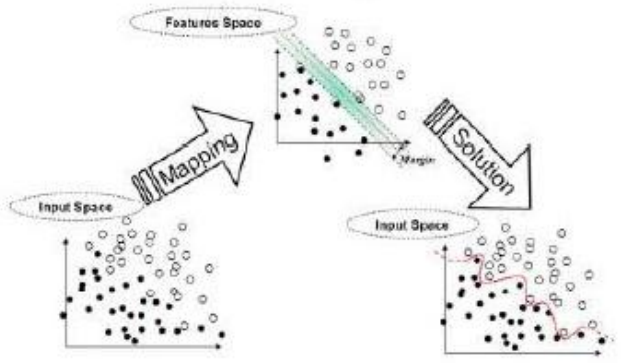
2. Support Vector Machines and Classification

Classification maps a data item into one of several predefined categories. These algorithms normally output "classifiers", for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient "normal" and "abnormal" audit data for a user or a program, then apply a classification algorithm to learn a classifier that will determine audit data as belonging to the normal class or the abnormal class. Link Analysis determines relations between fields in the database. Finding out the correlations in audit data will provide insight for selecting the right set of system features for intrusion detection. Anomaly detection is about finding the normal usage patterns from the audit data, whereas misuse detection is about encoding and matching the intrusion patterns using the audit data[1].

A support vector machine constructs a hyper plane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class, since in general the larger the margin the lower the generalization error of the classifier. In mathematics, the dimension of a space or object is informally defined as the minimum number of coordinates needed to specify any point within it. A hyperplane of an n-dimensional space is a flat subset with dimension $n - 1$. By its nature, it separates the space into two half spaces. That is, the points that are not incident to the hyperplane are partitioned into two convex sets (i.e., half-spaces), such that any subspace connecting a point in one set to a point in the other must intersect the hyperplane.

In the SVM literature, a predictor variable is called an attribute, and a transformed attribute that is used to define the hyperplane is called a feature. The task of choosing the most suitable representation is known as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. So the goal of SVM modeling is to find the optimal hyperplane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyperplane are the support vectors. SVM is a useful technique for data classification. Even though it's considered that Neural Networks are easier to use than this, however, sometimes unsatisfactory results are obtained. A classification task usually involves with training and testing data which consist of some data instances. Each instance in the training

set contains one target values and several attributes. The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes. The figure below presents an overview of the SVM process.



Kernel: If data is linear, a separating hyper plane may be used to divide the data. However it is often the case that the data is far from linear and the datasets are inseparable. To allow for this kernels are used to non-linearly map the input data to a high-dimensional space.

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 feature space is chosen suitably, pattern recognition can be easy.

$$\langle x_1 \cdot x_2 \rangle \leftarrow K(x_1, x_2) = \langle \Phi(x_1) \cdot \Phi(x_2) \rangle$$

Kernel Functions

$K(x_i, x_j) = \Phi(x_i)^T x_j$ is called the kernel function. Though new kernels are being proposed by researchers, following four basic kernels are properly used:[8]

- linear: $K(x_i, x_j) = (x_i)^T x_j$
- polynomial:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i{}^T \mathbf{x}_j + r)^d, \gamma > 0$$

- radial basis function (RBF):

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2), \gamma >$$

- sigmoid:

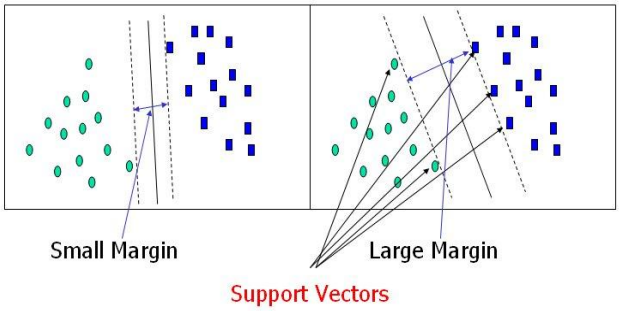
$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i{}^T \mathbf{x}_j + r)$$

Here γ, r, and d are kernel parameters.

In general, the RBF kernel is a reasonable first choice. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. Furthermore, the linear kernel is a special case of RBF Keerthi and Lin (2003) since the linear kernel with a penalty parameter C has the same performance as the RBF kernel with some parameters (C,γ). In addition, the sigmoid kernel behaves like RBF for certain parameters (Lin and Lin, 2003). The second reason is the number of hyperparameters which influences the complexity of model selection. The polynomial kernel has more hyperparameters than the RBF kernel.

Finally, the RBF kernel has fewer numerical difficulties. One key point is $0 < K_{ij} \leq 1$ in contrast to polynomial kernels of which kernel values may go to infinity ($\gamma\, xi^T xj + r > 1$) or zero ($\gamma\, xi^T xj + r < 1$) while the degree is large. Moreover, we must note that the sigmoid kernel is not valid (i.e. not the inner product of two vectors) under some parameters. There are some situations where the RBF kernel is not suitable. In particular, when the number of features is very large, one may just use the linear kernel[2].

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. We can plot the data points using the value of one predictor on the X axis and the other on the Y axis. 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.



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.

Steps for Classification using SVM

1. Prepare the pattern matrix
2. Select the kernel function to use

3. Select the parameter of the kernel function and the value of C
    a. Use the values suggested by the SVM software or set apart a validation set to determine the values of the parameters
4. Execute the training algorithm and obtain the α_i
5. Unseen data can be classified using the α_i and the support vectors

3. Tools for SVM

SVMs (Support Vector Machines) are a useful technique for data classification. Many open source as well as commercial tools is available for performing classification of data set along with different kernel functions. Some of the tools are explained here in detail

- **LIBSVM** (Library for Support Vector Machines), is developed by Chang and Lin and contains *C*-classification, ν-classification, ε-regression, and ν-regression. Developed in C++ and Java, it supports multi-class classification, weighted SVM for unbalanced data, cross-validation and automatic model selection. It has interfaces for Python, R, Splus, MATLAB, Perl, Ruby, and LabVIEW. The Kernels available are linear, polynomial, radial basis function, and neural (tanh).
- SVM^{light}, by Joachims, is one of the most widely used SVM classification and regression package. It has a fast optimization algorithm which can be applied to very large datasets. It has a very efficient implementation of the leave-one-out cross-validation. It is distributed as C++ source and binaries for Linux, Windows, Cygwin, and Solaris. The Kernels available are polynomial, radial basis function, and neural (tanh).
- LS-SVMlab, by Suykens, is a MATLAB implementation of least squares support vector machines (LS-SVM) which reformulates the standard SVM leading to solving linear KKT systems. LS-SVM alike primal-dual formulations have been given to kernel PCA, kernel CCA and kernel PLS, thereby extending the class of primal-dual kernel machines. Links between kernel versions of classical pattern recognition algorithms such as kernel Fisher discriminant analysis and extensions to unsupervised learning, recurrent networks and control are available.
- LSVM (Lagrangian Support Vector Machine) is a very fast SVM implementation in MATLAB by Mangasarian and Musicant. It can classify datasets with several millions patterns.

4. Applications of SVM

SVM has been found to be successful when used for pattern classification problems. Applying the Support Vector approach to a particular practical problem involves resolving a number of questions based on the problem definition and the design involved with it. One of the major challenges is that of choosing an appropriate kernel for the given application [4]. There are standard choices such as a Gaussian or polynomial kernel that are the default options, but if these prove ineffective or if the inputs are discrete structures more elaborate kernels will be needed. By implicitly defining a feature space, the kernel provides the description language used by the machine for viewing the data. Once the choice of kernel and optimization criterion has been made the key components of the system are in place [8]. Let's look at some examples.
The task of text categorization is the classification of natural text documents into a fixed number of predefined categories based on their content. Since a document can be assigned to more than one category this is not a multi-class classification problem, but can be viewed as a series of binary classification problems, one for each category. One of the standard representations of text for the purposes of information retrieval provides an ideal feature mapping for constructing a Mercer kernel [25]. Indeed, the kernels somehow incorporate a similarity measure between instances, and it is reasonable to assume that experts working in the specific application domain have already identified valid similarity measures, particularly in areas such as information retrieval and generative models [25] [27]. Traditional classification approaches perform poorly when working directly because of the high dimensionality of the data, but Support Vector Machines can avoid the pitfalls of very high dimensional representations [12]. A very similar approach to the techniques described for text categorization can also be used for the task of image classification, and as in that case linear hard margin machines are frequently able to generalize well [8]. The first real-world task on which Support Vector Machines were tested was the problem of hand-written character recognition. Furthermore, multi-class SVMs have been tested on these data. It is interesting not only to compare SVMs with other classifiers, but also to compare different SVMs amongst themselves [23]. They turn out to have approximately the same performance, and furthermore to share most of their support vectors, independently of the chosen kernel. The fact that SVM can perform as well as these systems without including any detailed prior knowledge is certainly remarkable [25].

Strength and Weakness of SVM:

The major strengths of SVM are the training is relatively easy. No local optimal, unlike in neural networks. It scales relatively well to high dimensional data and the trade-off between classifier complexity and error can be controlled explicitly. The weakness includes the need for a good kernel function.

5. CONCLUSION

Support Vector Machines acts as one of the best approach to data modeling and classification. They combine generalization control as a technique to control dimensionality. The kernel mapping provides a common base for most of the commonly employed model architectures, enabling comparisons to be performed. In classification problems generalization control is obtained by maximizing the margin, which corresponds to minimization of the weight vector in a canonical framework. The solution is obtained as a set of support vectors that can be sparse. The minimization of the weight vector can be used as a criterion in regression problems, with a modified loss function. We hope that this survey will help the researchers in understanding basics of SVMs. Future directions include developing a technique for choosing the kernel function and enhancing the anomaly detection process by effective classification.

6. REFERENCES

[1] Steve R Gunn, "Support Vector Machines for Classification and Regression", Technical report, University of Southampton.

[2] Tarem Ahmed, Boris Oreshkin and Mark Coates, "Machine Learning Approaches to Network Anomaly Detection", Department of Electrical and Computer Engineering, McGill University, http://www.usenix.org/event/sysml07/tech/full_papers/ahmed/ahmed.pdf

[3] B. Meshram and Alok K. Kumar , " HyIDS: Hybrid Intrusion Detection System", Proceedings of National Conference on Research & Practices in Current Areas of IT, March26-27, 2004, Department of Computer Science & Engineering, Sant Harchand Sing Longowal Central Institute of Engineering & Technology, Longowal, Dist Sangar(Punjab)-148106

[4] Nello Cristianini and John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.

[5] Image found on the web search for learning and generalization in svm following links given in the book above.

[6] David M Skapura, Building Neural Networks, ACM press, 1996.

[7] Tom Mitchell, Machine Learning, McGraw-Hill Computer science series, 1997.

[8] J.P.Lewis, Tutorial on SVM, CGIT Lab, USC, 2004.

[9] Vapnik V., "Statistical Learning Theory", Wiley, New York, 1998.

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

[11] N. Aronszajn. Theory of reproducing kernels. Trans. Amer. Math. Soc., 686:337–404, 1950.

[12] C. Cortes and V. Vapnik. Support vector networks. Machine Learning, 20:273 – 297,1995

[13] A. J. Smola. Regression estimation with support vector learning machines. Master's thesis, Technische Universit¨at M¨unchen, 1996.

[14] N. Heckman. The theory and application of penalized least squares methods or reproducing kernel hilbert spaces made easy, 1997.

[15] Vapnik, V., Estimation of Dependencies Based on Empirical Data. Empirical Inference Science: Afterword of 2006, Springer, 2006

[16] http://www.enm.bris.ac.uk/teaching/projects/2004_05/dm1654/kernel.htm

[17] Duda R. and Hart P., "Pattern Classification and Scene Analysis", Wiley, New York 1973.

[18] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, Neural Networks for Signal Processing VII — Proceedings of the 1997 IEEE Workshop, pages 276 – 285, New York, 1997. IEEE.

[19] M. O. Stitson and J. A. E. Weston. Implementational issues of support vector machines. Technical Report CSD-TR-96-18, Computational Intelligence Group, Royal Holloway, University of London, 1996.

[20] Burges B.~Scholkopf, editor, "Advances in Kernel Methods--Support Vector Learning". MIT press, 1998.

[21] Osuna E., Freund R., and Girosi F., "Support Vector Machines: Training and Applications", A.I. Memo No. 1602, Artificial Intelligence Laboratory, MIT, 1997.

[22] Trafalis T., "Primal-dual optimization methods in neural networks and support vector machines training", ACAI99.

[23] Veropoulos K., Cristianini N., and Campbell C., "The Application of Support Vector Machines to Medical Decision Support: A Case Study", ACAI99

[24] B.B.Meshram, S.S..Karvande. "Design And Implementation Of Application Layer Firewall For Secure Internet Access" at International Conferences on Soft Computing, Department of Computer Applications, Computer Science & Engineering, Information Technology, Bharath Institute of Higher Education & Research, Chennai, Tamilnadu. May 28th and 29th, 2004.

[25] Peng Ning, North Carolina State University, Sushil Jajodia, "Intrusion Detection Techniques", George Mason University.