ISSN 2395-1621

Survey on Support Vector Machine for Data Mining

[#1]Vidya Patil, [#2]Prof. Vandana Navale

[1]patilvidya169@gmail.com

[#12]Department of Computer Engineering

Dhole Patil Collage of Engineering, Pune, India

ABSTRACT

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. 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.

ARTICLE INFO

Article History

Received: 2nd May 2018

Received in revised form :

2nd May 2018

Accepted: 8th May 2018

Published online :

8th May 2018

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. 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.

The Support Vector Machine (SVM) was first proposed by Vapnik and has since attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. Sims have been employed in a wide range of real world problems such as text categorization, hand-written digit recognition, tone recognition, image classification and object detection, micro-array gene expression data analysis, data classification.

It has been shown that Sims is consistently superior to other supervised learning methods. However, for some datasets, the performance of SVM is very sensitive to how the cost parameter and kernel parameters are set. As a result, the user normally needs to conduct extensive cross validation in order to figure out the optimal parameter setting. This process is commonly referred to as model selection. One practical issue with model selection is that this process is very time consuming. We have experimented with a number of parameters associated with the use of the SVM algorithm that can impact the results.

II. LITERATURE SURVEY

The support vector algorithm is a nonlinear generalization of the generalized portrait algorithm developed in Russia in the sixties [1][2]. However, a similar approach using linear instead of quadratic programming was taken at the same time in the US, mainly by Mangasarian [3][4][5]. As such, it is firmly grounded in the framework of statistical learning theory, which has been developed over the last three decades by Vapnik himself [6][7][8]. In its present

© 2018, IERJ All Rights Reserved

form, the support vector machine (SVM) was largely developed at AT&T Bell Laboratories by Vapnik and co-workers. SVM have been recently proposed as a very effective method for general purpose classification and pattern recognition [8]. Intuitively, given a set of points which belong to either of two classes, a SVM finds the hyperplane leaving the largest possible fraction of points of the same class on the same side, while maximizing the distance of either class from the hyperplane. According to [7] [9], this hyperplane minimizes the risk of misclassifying examples of the test set. Prior related work includes that of Baek et al. [10], who presented a vehicle color classification based on the SVM. The implementation results showed 94.92 of success rate for 500 outdoor vehicle with 5 colors. Ambardekar et al. [11] used optical flow and knowledge of camera parameters to detect the pose of a vehicle in the 3D world. This information is used in a model-based vehicle detection and classification technique employed by their traffic surveillance application. Ma et al. [12] proposed an approach to vehicle classification under a mid-field surveillance framework. They discriminate feature based on edge points and modified SIFT descriptors. Eigenvehicle and PCA-SVM were proposed and implemented to classify vehicle into trucks, passenger cars, van and pick-ups in paper [13].

III. PROPOSED METHODOLOGY

Classification is a data mining (machine learning) technique used to predict group membership for data instances.
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. 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.

2) Rule based classifier: Represent the knowledge in the form of IFTHEN rules and One rule is created for each 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.

IV. 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.

V. 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 overfitting. 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.

REFERENCES

[1] V. Vapnik and A. Lerner, "Pattern recognition using generalized portrait method," Automation and Remote Control, 24, pp. 774-780, 1963.

[2] V. Vapnik and A. Chervonenkis, "A note on one class of perceptrons," Automation and Remote Control, 25, 1964.

© 2018, IERJ All Rights Reserved

[3] O.L. Mangasarian, "Linear and nonlinear separation of patterns by linear programming," Operations Research, 13:pp. 444-452, 1965.

[4] O.L. Mangasarian, "Muti-surface method of pattern separation," IEEE Transactions on Information Theory IT-14, pp. 801-807, 1968.

[5] O.L. Mangasarian, "Nonlinear Programming," McGraw-Hill, New York, 1969.

[6] V.Vapnik, "Estimation of Dependences Based on Empirical Data," Springer, Berlin. 1982.

[7] V. Vapnik, "The Nature of Statistical Learning Theory," Springer, New York. 1995.

[8] C. Cortes and V. Vapnik, "Support vector network," Machine Learning, vol. 20, pp. 1-25, 1995.

[9] V. Vapnik, "An overview of statistical learning theory," IEEE Transaction on neural networks, vol. 10, No. 5, pp. 988-999, 1999.

[10] N. Baek, S.-M. Park, K.-J. Kim, S.-B. Park, "Vehicle Color Classification Based on the Support Vector Machine Method", ICIC 2007, CCIS 2, pp. 1133–1139, 2007.

[11] A. Ambardekar, M. Nicolescu, G. Bebis, "Efficient Vehicle Tracking and Classification for an Automated Traffic Surveillance System," Signal and Image Processing, August 2008.

[12] Ma, W. E. L. Grimson, "Edge-based rich representation for vehicle classification", International Conference on Computer Vision, 2006, pp. 1185-1192.

[13] C. Zhang, X. Chen, W.-B. Chen, "A PCA-based Vehicle Classification Framework", International Conference on Data Engineering Workshops (ICDEW'06), 2006.

© 2018, IERJ All Rights Reserved