Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings 5 (2018) 1679–1685

materialstoday: PROCEEDINGS

www.materialstoday.com/proceedings

Efficient Decision Tree Based Data Selection and Support Vector Machine Classification

Arumugam. P[a], Jose. P[b]

[a]*Associate Professor, Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, 628012, Tamilnadu, India*
[b]*Research Scholar,Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, 628012, Tamilnadu, India*

Abstract

Today' real world data bases witnessed significant increase in the amount of data in digital format, due to the widespread use of datasets and storage system. There is a need to developing fast and highly accurate algorithms to automatically classify large data. It becomes a vital part of the machine learning and knowledge discovery. The main intention of this paper is however data sizes increases, our proposed method make faster computation and scalable machine learning algorithm is used to learn faster from the labelled training data. Due to its strong mathematical background and theoretical foundation and good generalization performance, Support Vector Machine (SVM) Classification becomes more feasible options for large datasets. A major research goal of SVM is to improve the speed in training and testing phase. In this paper We introduce a proposed algorithm to speed up the training time of SVM is presented. It is highly accurate classification method. However SVM classifiers suffer from slow processing, when training with a large set of data tuples. Our novel approach selects a small representative amount of data from large datasets to enhance training time of SVM. This method uses an induction tree to reduce the training dataset for SVM classification, it generate faster results with improving accuracy rates than the current SVM implementations.

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

Keywords: Microarray; large datasets; Classification; Decision Tree, SVM

1. Introduction

SVM grew out of early effort by Vladimer Vapnik and Alexei Chervonenkis on statistical learning theory[1] . And the first paper on SVMs was presented by Boser, Guyon, and Vapnik[2]. The SVM is a highly accurate classification method. However SVM classifiers suffer from slow processing, when training with a large set of data tuples. Although the training time of even the fastest SVMs can be extremely slow. Conversely, it is known that the major drawback of SVM occurs in its trainingphase[8, 9]. To investigate this problem the efficient Decision Tree (DT) based Data selection and Support Vector Classification technique is needed for training large data sets, for data reduction DT s have been used in several works .In each disjoint region exposed by a DT to train a SVM[11]. Extracting

2214-7853 © 2017 Elsevier Ltd. All rights reserved.
Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

bioinformatics information from DNA micro array technology is the popular method in biological data. In bioinformatics, SVM classification is one of the popular tool for identify correlation among various clause and the features of various objects. In recent years, determining the informative genes that cause cancer has its great impact in microarray technology. The foremost snag that exists in microarray data analysis is the curse of dimensionality, this issue carried out in many ways. According to their strong mathematical background and generalization capability, SVM produced better result. Meanwhile in its expensive computation, high dependency on the size of input dataset. We proposed a method to reduce the dimensionality of training set for SVM based decision tree. The novel method produces better accuracy and in a faster way that the traditional SVM implementations.

Classification has been an indispensable premise in statistics, data mining, machine learning, bioinformatics and Medical science. Recent advances in data mining research have led to the progress of abundant and scalable methods for mining fascinating patterns and knowledge in large databases. In the past decades spectator changes in biomedical research an explosive progress of biomedical data. In particular cancer therapy investigations find whether it is cancerous cell or not. Our novel approach has training dataset with two classes. SVM is usually regard as the most accurate classification tool for many bioinformatics applications .however the complexity of training an SVM is $O(N^2)$,where N is the number of objects/points. In this progress find how to scale up SVMs for large data novel method uses. Therefore the novel method uses in its training phase with decision tree to reduce the training dataset for SVM. In this paper to reduce the size of datasets based on a data selection method Decision Tree approach proposed. That is ability to deal with redundant attributes and robustness to noise, able to partition the input space into regions with low entropy and liable human interpretation. The experimental results show that proposed method reduces the training time with higher accuracy. Decision Tree (DT) have been used a dimensionality reduction approach in some previous work .It reduces the number of data dimensions and introduce the problems of feature selection and feature construction. In each disjoint region exposed by a decision tree is used to train SVM. Thus the region found by small datasets is less sophisticated than the region obtained by the entire training set. Even though small learning datasets reduce decision tree complexity through decision rule. A SVM is a good classification method was obtainable in[4].It reveal in reducing the number of instances to train a SVM. The key idea was to fairly accurate decision boundary of SVM by using DT that is to confine the objects near the decision boundary. The novel approach presented in this paper is very efficient with biological datasets, first it eliminates dispensable data, and then it recovers the data points that are near to the decision boundaries. The training dataset of SVM is done on those remained valuable data. This effective way to reduce SVM training time approach is necessary a rational tradeoff between accuracy and training time.

2. Related Work

Our novel method is based on that, the region originate on undersized data sets are less elaborated than the region obtained by the full training set[13,14].A related approach to our proposed technique was presented in[16], it consists to reduce the training instances of SVM. The idea was to fairly accurate the decision boundary of SVM by using DT. It also used to choose an example which is near to decision boundary[17], but the consideration of sigma parameters optimization is not proper. The novel method presented in this paper is efficient with large data sets .It eradicates nonessential data, then it recuperate the data points that are near to the decision boundaries. After that the training of SVM is made on those remained valuable data. The way presented in this paper yet prominent and effective mode to reduce the SVM training time. It is very necessary applications on the real world data base accuracy and training time.

3. Proposed Work

The proposed method for classification of microarray input data is processed by statistical analysis and database analysis. To mine microarray and large dataset might require data normalization (data transformation) with esteem to the same control gene and a selection of a subset of treatments (data reduction).In this paper introduce a preprocess step to rapidly remove data that do not contribute to classify the decision boundary of SVM. While preserving Support vector candidate's .The novel approach applies SVM on a tiny subset of original data set in order to attain a draft of the optimal separating hyperplane, then it labels objects that are remote from sketched hyperplane and objects that are close to it. In this to identify objects that have akin characteristics to the computed Support