IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 9, September 2017

Distribution Preserving Kernel Based Supervised Machine Learning Algorithms for Big Data

Sudha M[1], Saravana Kumar E[2]

P.G. Scholar, Dept of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India[1]

Associate Professor, Dept of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India[2]

Abstract: Data mining is the process of sorting through large datasets to identify patterns and establish relationships to solve problems through data analysis. Data mining is a technique which is used to separate the accurate value from the dataset. Support vector machine is a supervised machine learning algorithm used for classification and regression, SVM mainly used to classifies the datasets to improve classification accuracy, several SVM algorithm are there such as LIB-SVM,DC-SVM,CA-SVM and Dip-SVM these algorithms are used to find the accuracy and performance while performing classification in data mining. Dip-SVM also reducing the communication overhead between clusters.

Keywords: Classification, SVM, Dip-SVM, CA-SVM, DC-SVM.

I. INTRODUCTION

Data mining predicts the relevant information from the large database. Support vector machine is based on the statistical learn-in theory developed by Vapnik et al [1]. The core of training a support vector machine (SVM) involves solving a quadratic programming problem which demands more computing power for large datasets[1][20]. Data mining is the process of discovering actionable information from large sets of data. Each data mining function specifies a class of problems that can be modeled and solved. Data mining functions fall generally into two categories: supervised and unsupervised.

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data duration, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

Data sets grow rapidly - in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The work may require "massively parallel software running on tens, hundreds, or even thousands of servers". What counts as "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

Quadratic programming is a type of mathematical optimization problem which optimizes a quadratic function of several variables subject to linear constraints on these variables. The quadratic programming problem solver separates support vectors from the rest of the training data. For moderately sized datasets, support vector machine has long been used extensively for classification and regression problems in many areas like genomics, e-commerce, computer vision, cyber security, etc. due to its generalization capabilities. However, classification of high volume data in the form of text, images or videos into meaningful classes using support vector machine is quite challenging Various implementations of SVM are available such as LIBSVM , LS-SVM, SVM light and so on. LIBSVM is the most popular among them because it utilizes a highly optimized quadratic solver. However, training a kernel SVM is still difficult for large datasets where the sample size reaches more than one million instances [14] [3]. Distributed environments like a high-performance computing and cloud clusters are widely used for solving data-intensive and time-consuming problems. However, sequential minimal optimization (SMO), the most successful quadratic programming (QP) solver used extensively in SVM implementations cannot leverage the benefits of these distributed environments because there is high dependency among the parameters used for optimization. So far, there is no true parallel or distributed algorithm in literature that solves the constrained quadratic programming problem used to identify the support vectors in the training data. The local support vectors (LSVs) are the sup-port vectors resulting from individual SVM training over the smaller partitions (i.e., support vectors from local SVM models), while global support vectors (GSVs) are the support vectors of final SVM model obtained by training over assembled local support vectors[14].

DOI10.17148/IJARCCE.2017.6921

A variety of techniques to support classification in large database have been proposed in this literature. Most existing techniques adopt definitions that enable different performance metrics in an attempt to improve the classification accuracy. Literature review id classified into two forms as,

- Algorithm based survey
- Dataset based survey

TABLE I RELATED PAPER'S

Title	Author	Year	Methods	Advantages
A divide-and-conquer solver for kernel support vector machines	cho-jui hsieh cjhsieh @ cs . utexas . edu si si ssi @ cs . utexas . edu inderjit s. dhillon	2011	Partition the kernel svm problem into smaller sub problems by clustering the data	Good Performance
Support-vector networks,” mach learn	f. ahmad, s. lee, m. thottethodi, and t. vijay kumar	2012	This study concerns fine grained scheduling on MapReduce operations, with each other.	High efficiency.
A distributed svm method based on the iterative MapReduce	g. ananthanarayanan, a. gods, a. wang, d. borthakur, s. candela, s. shankar, and i. stoic	2012	Data-intensive analytics on large clusters is important for modern internet services. as machines in these clusters have large memories, in-memory caching of inputs is an elective way to speed up these analytics jobs.	Simple design and architecture
CA-svm: communication-avoiding support vector machines on distributed systems	james	2010	Implement communication-efficient versions of parallel support vector machines, a widely used classifier in statistical machine learning, for distributed memory clusters and supercomputers.	Simple implementation
Core vector machines: fast on very large data sets	p. costa, a. donnelly, a. rowstron, and g. o.shea	2014	Large companies like facebook, google, and microsoft as well as a number of small and medium enterprises daily process massive amounts of data in batch jobs and in real time applications.	Good performance

II. SUPPORT VECTOR MACHINE (SVM)

A Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression purposes. SVMs are more commonly used in classification problems. In SVM the idea of finding a hyper plane that best divides a dataset into two classes. Hyper plane is a line that linearly separates and classifies a set of data's. Data points nearest to the hyper plane is called support vectors The distance between the hyper plane and the nearest data point in set is known as the margin. SVM used in the following areas:

- Text mining
- Nested data problems e.g. transaction data or gene expression data analysis.
- Pattern recognition

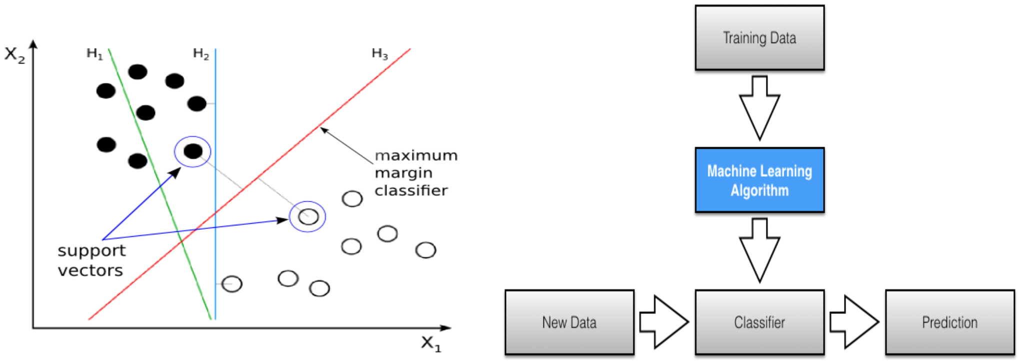

Fig. 1 SVM Classification

III. LIBSVM: A LIBRARY FOR SUPPORT VECTOR MACHINES

LIBSVM is a library for Support Vector Machine. It helps the users to apply SVM to their applications. LIBSVM has popularity in machine learning and many other areas. Support Vector Machines (SVMs) are a popular machine learning method for classification, regression. LIBSVM supports the following learning tasks, some domains that have successfully used LIBSVM [2] [2].
1. SVC: support vector classification (two-class and multi-class).
2. SVR: support vector regression
3. One-class SVM

LIBSVM involves two steps:
1. Training a data set to obtain a model
2. Using the model to predict information of a testing data set.

SVM formulations supported in LIBSVM:
- C-support vector classification (C-SVC)
- v-support vector classification (v-SVC)

LIBSVM supports SVM formulations for classification, regression, and distribution.

$$\text{classification Accuracy} = \frac{\#\ \text{correctly predicted data}}{\#\ \text{total testing data}}$$

Two implementation techniques
- Shrinking
- Caching

Shrinking: To save the training time, the shrinking technique tries to identify and remove some bounded elements to solve smaller optimization problem.
Caching: This is an effective technique for reducing the computational time of the decomposition method.

IV. DIVIDE AND CONQUER SUPPORT VECTOR MACHINE (DC-SVM)

DC-SVM partitions the SVM problems in to smaller sub-problems by clustering the data. Each sub-problem can be solved separately with efficiency. In early prediction strategy, DC-SVM achieves about 96% accuracy. The DC-SVM has two steps
- DIVIDE
- CONQUER

Divide step specifies how to divide the problems in to sub-problems. DC-SVM identifies support vectors much faster than shrinking strategy in LIB-SVM. We can use any SVM solver in Divide and Conquer framework, we are using coordinate decent method to solve whole problem. The benefit of this coordinate decent method is to avoid a lot of unnecessary access to the kernel [3]. Multiple levels used in DC-SVM: DC_SVM is used to reduce the time for solving the sub-problems using multiple levels. To speed up our procedure using [3],
- Adaptive clustering
- Early identification of support vector
- Early prediction based on ‘L’ level solution
- Refine solution before solving the whole problem

4.1 Competing methods:

The competing methods, LIB-SVM, Cascade SVM, approximate SVM solver (SP-SVM, LLS-SVM, LTPU) and On-line SVM (La SVM).In DC_SVM we use the modified LIB-SVM to solve sub-problems.
Advantages: Efficiency High
Disadvantages: Improper Security.

V. COMMUNICATION AVOIDING SVM(CA-SVM)

Most popular Kernel SVM training algorithm is sequential Minimal optimization (SMO), SMO has low arithmetic intensity SMO has poor scaling the reason is that SMO is an iterative algorithm[5] .
CA-SVM achieves significant speedup over the original algorithm with small losses in accuracy on our test sets .by this way we can manage the speedup and accuracy

DOI10.17148/IJARCCE.2017.6921

VI. DISTRIBUTION PRESERVING KERNEL SVM(DIP-SVM)

This approach has two distinct phases

- Distribution preserving partioning phase(DPP)
- Disrtribution Learning phase (DL)

6.1 Distribution preserving partioning phase(DPP)

In this phase the dataset is divided in first and second order stategic(mean and variance)of the datasets are preserved in every parttition. This phase revolves around balanced partitioning while preserving the statistical properties of entire datasets. The partition is approximately close to the given dataset.

6.2 Disrtribution Learning phase (DL)

The learning phase uses the modified cascade SVM for distributed support vectors. The partitioning the dataset instead of sequential/random partition using cascade SVM or clustering partition, it helps to maintain the classification accuracy. Relevant training vectors are using to reduce the communication overhed.

VII. RELATED WORKS

7.1 Data Mining

Data Mining is the process of extraction of useful information and patterns frome huge volume of data[11]. The other names are knowledge discovery process, knowledge extraction.After find the useful data used to make ertain decision. The steps are cleaning data, form the pattern identification, deployed for desired outcomes. The classification process involves learning trainin data by using classification algorithm[11]. To find the accuracy the classification can be used. The set of parameters are determined by using the classifier training algorithm.

Types of classification models: • Classification by decision tree induction • Bayesian Classification • Neural Networks • Support Vector Machines (SVM) • Classification Based on Associations

7.2 Cloud computing

Cloud computing is an information technology (IT) paradigm, a model for enabling ubiquitous access to shared pools of configurable resources (such as **computer** networks, servers, storage, applications and services), which can be rapidly provisioned with minimal management effort, often over the Internet. In a cloud computing environment, the traditional role of service provider is divided into two: the infrastructure providers who manage cloud platforms and lease resources according to a usage-based pricing model, and service providers, who rent resources from one or many infrastructure providers to serve the end users (eg: Amazon EC2)[12]. EC2 provides the ability to place instances in multiple locations. Amazon Virtual Private Cloud (VPC) is a secure and seamless bridge between a company's existing IT infrastructure and the AWS cloud.

7.3 Map Reduce

Map Reduce jobs are controlled by the client node through a multi-step process. During configuration, the client assigns Map Reduce methods to the job, prepares Key Value pairs and prepares static data for Map Reduce tasks through the partition file if required. The message communicate between job is realized with message brokers, i.e. Narada Brokering or Active MQ. Map daemons operate on computation nodes, loading the Map classes and starting them as Map workers. Reduce daemons operate on computation nodes. The number of reducers is prescribed in client configuration step. The reduce jobs depend on the computation results of Map jobs. The communication between daemons is through messages. Combine job is to collect Map Reduce results. It operates on client node [13].

VIII. PROPOSED WORK

Datasets

The most of the researches commonly used datasets in text based classification are described as follows, gisette, adult, web spam, cifart [14]. This research accordingly the sensor related accidental datasets are used for finding classification accuracy in data mining. In this dataset several attributes are their (eg: state, year, male, female, total).

List of Modules

- Preprocessing
- Partitioning
- K-Means
- Dip-SVM

Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data intoan understandable format. Eliminating redundant data to save memory. Eliminating NULL data. Making the datasets in Structured format for ease use.

DOI10.17148/IJARCCE.2017.6921

Partitioning

Dataset is splitted in to manageable partitions, each partitions are trained by a svm model. This is one of the mining Method here we apply some logic and in this module and spit our file randomly. Partition has done based on the size of file.

Clustering

A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. In proposed system we are using K-MEANS algorithm for clustering. In our dataset year based clustering to find out overall accident status.

Dip-SVM

Data classification is the process of organizing data into categories for its most effective and efficient use. The goal of classification is to accurately predict the target class for each case in the data. Dip-SVM –algorithm is used to obtain the minimal loss in classification accuracy and low communication overhead, two phases:

1. Distribution preserving partitioning (DPP)
2. Distributed learning phase(DL)

In our dataset the state based classification is going to perform by using Dip-SVM algorithm.
Dataset (state accidental details in year) compared graphs

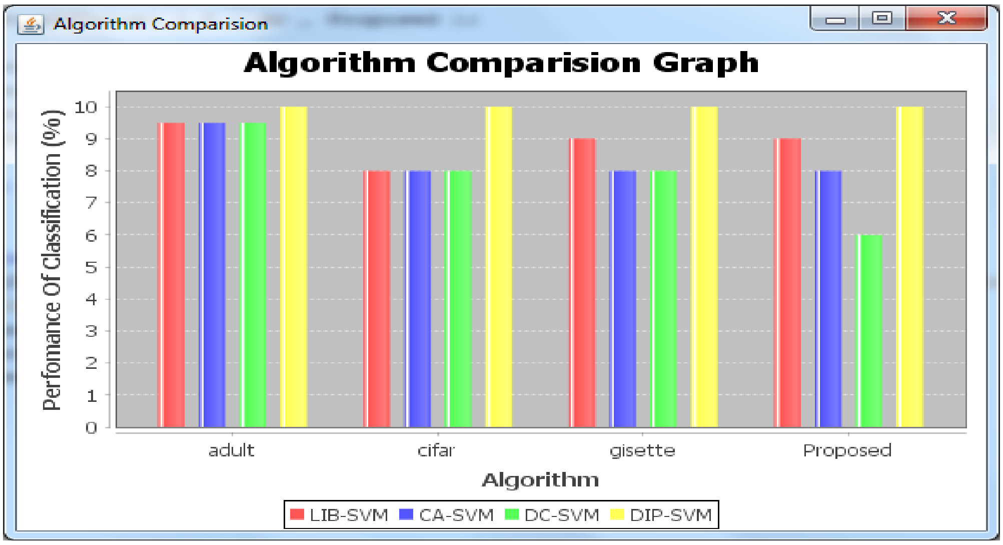

Fig. 2 Algorithm comparison graph

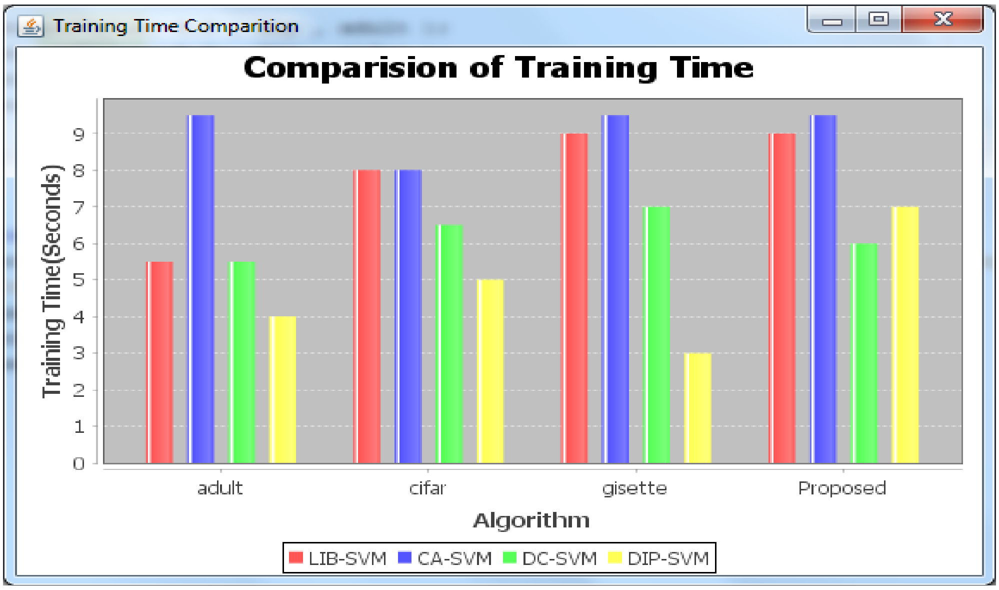

Fig. 3 Training time graph

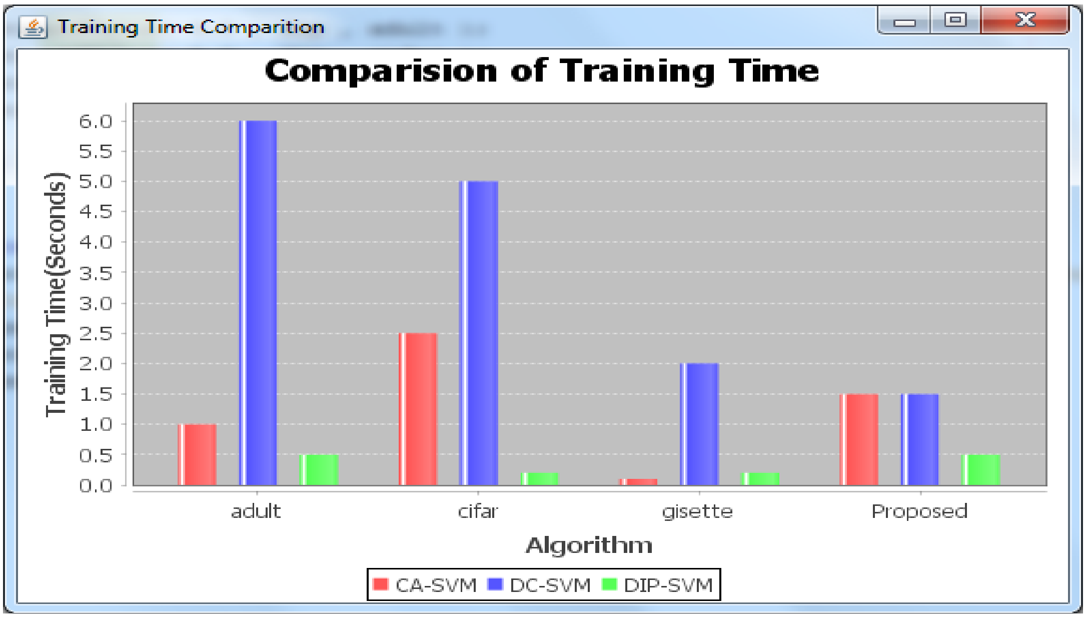

Fig. 4 Comparison of Training time

IX. CONCLUSION

Due to the requirement of accurate classififcation we need to compare the supervised and unsupervised algoritms to obtain the expecting classified output. After comparing all these algorithms Dip-svm give more acuuracy. Most of the implementation was carried out in Java or Matlab with commonly used datasets. This approach we may try on image dataset also if needed.

REFERENCES

1. C. Cortes and V. Vapid, "Support-vector networks," Mach. Learn. , vol. 20, no. 3, pp. 273–297, 1995.
2. C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intel. Syst. Technol., vol. 2, pp. 1–27, 2011, [Online]. Available: http://www.csie.ntu.edu.tw/ clan/libsvm
3. C.-J. Hsieh, S. Si, and I. Dhillon, "A divide-and-conquer solver for kernel support vector machines," in Proc. Int. Conf. Mach. Learn., Jun. 21–26, 2014, vol. 32, no. 1, pp. 566–574.
4. N. K. Alham, M. Li, S. Hammoud, Y. Liu, and M. Ponraj, "A dis-tributed SVM for image annotation," in Proc. Int. Conf. Fuzzy Syst. Knowl. Discovery, Aug. 10–12, 2010, pp. 2983–2987
5. Y. You, J. Demmel, K. Czechowski, L. Song, and R. Vuduc, "CA- SVM: Communication-avoiding support vector machines on dis- tributed systems," in Proc. IEEE Int. Parallel Distrib. Process. Symp., May 25–29, 2015, pp. 847–859
6. L. Zanni, T. Serafini, and G. Zanghirati, "Parallel software for training large scale support vector machines on multiprocessor systems," J. Mach. Learn. Res., vol. 7, pp. 1467–1492, 2006.
7. K. Xu, C. Wen, Q. Yuan, X. He, and J. Tie, "A Mapreduce based parallel SVM for email classification," J. Newt. , vol. 9, no. 6, pp. 1640–1647, 2014.
8. [19] N. K. Balham, M. Li, Y. Liu, S. Hammond, and M. Pankaj, "A dis- tribute SVM for scalable image annotation," in Proc. Int. Conf. Fuzzy Syst. Know. Discovery, Jul. 26–28, 2011, pp. 2655–2658.
9. X. Zeng and T. R. Martinez, "Distribution-balanced stratified cross-validation for accuracy estimation," J. Exp. Theoretical Art if. Intel. , vol. 12, no. 1, pp. 1–12, 2000.
10. H. Yu, J. Yang, J. Han, and X. Li, "Making SVMs scalable to large data sets using hierarchical cluster indexing, "Data Mining Know. Discovery, vol. 11, no. 3, pp. 295–321, 2005.
11. Bharati M. Ramageri / Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305 "Data mining techniques and applications".
12. Qi Zhang · Lu Cheng '"Raouf Boutaba " Cloud computing: state-of-the-art and research challenges'', 20 April 2010,
13. Zhanquan Sun1, Geoffrey Fox2,' Study on Parallel SVM Based on MapReduce', 26 November 2014.
14. Dinesh Singh, Student Member, IEEE, Debaditya Roy, Student Member, IEEE, and C. Krishna Mohan, Member, IEEE" DiP-SVM : Distribution Preserving Kernel Support Vector Machine for Big Data" ieee transactions on big data, vol. 3, no. 1, January-march 2017