A Review of Adaptation SVM(A-SVM) in Grid Environment

Manjiri V. Kotpalliwar
PG Student: Department of computer science and Engineering
YCCE
Nagpur, India
manjirikotpalliwar@yahoo.com

Prof. Rakhi Wajgi
Professor: Department of computer science and Engineering
YCCE
Nagpur, India
wajgi.rakhi@gmail.com

Abstract: Distribution, collection, sharing, controlling and manipulation of data allows for solving computational problems and executing the applications that are distributed in nature. Data mining algorithms are data intensive; therefore the Grid can offer a computing and data management infrastructure for parallel data analysis. SVM (Support Vector Machine) is a method which is used in data mining to extract predicted data. But building a distinctive SVM model for each class of large multiclass database is both laborious and time consuming in terms of labeling and training the data. In these paper, we address these difficulties by using a regularization-based algorithm called adaptation SVM (A-SVM), through which existing prediction model is adapted to a new domain, results in decreasing amount of labeled data and training cost keeping performance into consideration. This paper discussed about various applications of A-SVM in Grid environment.

Keywords*- SVM Algorithm, A-SVM (Adaptation SVM), Grid Environment.*

*****

I. INTRODUCTION

The Grid is a collection, distribution and controlling of services in which resources is shared individually within dynamic organizations. A grid is a distributed system that allows sharing, selection and aggregation of distributed "autonomous" resources dynamically at runtime depending on their availability, performance, cost, and user's Quality–Of-Services (QoS) requirements [11]. To solve large-scale computation problems, Grid uses the resources of many separate computers which are connected by a network. Grids provide the facility to perform computations on large datasets, by breaking them down into many smaller parts, or give the ability to perform many more computations at once that would be possible on a single computer, by modeling a parallel computing between processes [11]. The advantage of the Grid is high efficiency of using technological ability. The grid is high effective in using associated technological capacities of creative users potential, the safety, the reliability and high level of transportability for computational applications. Network is the main building block of the Grid. The main aim of grid computing is to give organizations and application developers the ability to create distributed computing environments that can utilize computing resources on demand [11].

A support vector machine (SVMs) was first proposed by Vapnik [23] which is used for binary classification. SVM is a classifier which can be store the result in two ways: partial or true. SVM is a simplest linear form which is the hyperplane for separating the positive and negative examples.

II. LITERATURE REVIEW

A. Support Vector Machine (SVM)

Ali Meligy et al. [11] presented the grid-based distributed Support Vector Machine (SVM) Algorithm. They gives the fundamental concept of SVM which is defined over the vector space in which the problem is to find decision surface that separates the best data vectors into two classes [11]. SVM is a simplest linear form, generally it is a hyperplane that separates the positive examples from the negative example. From figure 1, it showed hyperplane that separates the training data by a maximal margin between two classes [25]. All vectors lies on one side of the hyperplane labeled as .1 and other vectors lies on another side of the hyperplane labeled as 1. The training objects that lie closest to the hyperplane are called the support vectors [20].

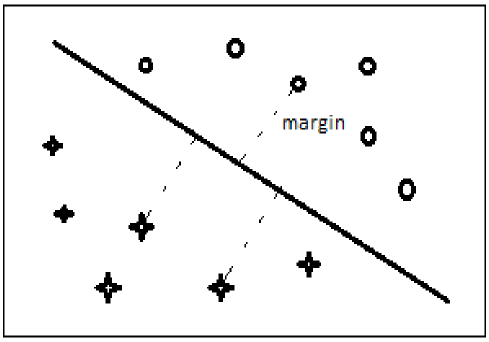

Fig.ure1. Simple Linear Support Vector Machine [11]

An advantage of the method is that the modeling of that hyperplane is only deals with these support vectors, preferably than the whole training dataset, and so the size of the training dataset is not usually an issue [11],[25]. A disadvantage is that the algorithm is time consuming and also it is sensitive to choice of the parameters, making it harder to use.

Inderjit S. Dhillon et al. [20] proposed a new information theoretic-divisive algorithm which is used for word clustering applied to text classification [20]. At lower number of features, this divisive algorithm achieves higher classification accuracy. Also, they have presented the detailed experimental result using Naïve Bayes and Support Vector Machine on 20 Newsgroup datasets [20].

Support Vector Machines (SVMs) [23] are generally inductive-learning schemes which are used for solving two class pattern recognition problems. Nowadays SVM also give good results in text categorization [16]. SVM is defined over the vector space in which the classification problem is to find decision surface that then "best" separates the data points of two classes. This decision surface is called as hyperplane which is maximizes the "margin" between two classes in case of linearly separable data [20].

Thanh-Nghi Do et al. [9] proposed the incremental, parallel and distributed Support Vector machine (SVM) using linear and non-linear kernels. There aims to classifying very large size of datasets on standard personal computers (PC's) [9]. They have proposed the Least Square SVM (LS-SVM) [21] for building the incremental, parallel and distributed algorithm [17].

The LS-SVM is used to construct new algorithm which is very fast to build incremental, parallel and distributed SVM for classification task. This new algorithm has ability to classify one billion data points in 20-dimensional input space into two classes in some minutes on ten machines [9].

The incremental LS-SVM algorithm is very fast to train the data and efficient to classify the large datasets, it needs to load the whole dataset in memory first [9]. The incremental LS-SVM can only runs on a single machine. It deals with large dataset to classify.

Simon Tong et al. [12]introduced an algorithm which is used for performing active learning with support vector machine (SVM) which means that algorithm is used for choosing which instances to request next [12]. Their experimental results showing that employing the active learning with SVM can reduce the need of labeled training instances of both inductive and transductive settings [12]. In SVM for induction, a labeled training set of data or a task is used to create a classifier which has a good performance on unseen test data. SVM also used for transduction. Here SVM can perform transduction by finding the hyperplane which maximizes the margin related to both labeled and unlabeled data. This transductive SVM is used for text classification [12].

Chih-Wei Hsu et al. [2] proposed several methods in which multiclass classifier are constructed by combining several binary classifiers. To solve the multiclass SVM [22] in one step, that has variables proportional to the number of classes. Therefore, for multiclass SVMs methods, there is need of binary classifiers to be constructed and the optimization problem. That's why this method is more computationally expensive to solve the multiclass problem than the binary problem with the same number of data [2]. They have shown three methods such as one-against-all [8], one-against-one and directed acyclic graph SVM (DAGSVM) [18] are based on solving the problem of binary classification.

Classification is the most important task specially used for the different applications such as text categorization, image classification, data classification etc. Durgesh K. Srivastava et al. [10] applied Support Vector Machine (SVM) on different data like Diabetes data, Heart data which have more than one or multi class in terms of their features, classes, number of training data and number of testing data. They have shown the comparative results by using the different kernels for all data samples [10].

They introduced the concept of SVM with kernel function selection and model selection which are used to solve the classification problem. Out of all kernel function, the RBF is the most popular kernel function because it has less numerical difficulties. SVMs are related to the supervised learning method which is used for the classification and regression [23].SVM is generally used for linear classification. SVM has a property which is simultaneously minimize the empirical classification error and maximize the geometric margin. That's why SVM called Maximum Margin Classifiers.

B. Grid-Based SVM

The Grid is collection, distribution and sharing of service within dynamic organizations consisting of resources individuals. The Grid-based SVM is generally focused on the Grid data services, i.e. data access or metadata access. It can process the large data sets. For multiclass classification, Grid data services are used to access the metadata i.e. the data about data [11]. Some of the methods are time consuming and use of the Grid infrastructure can gives the important benefits. Implementation of text mining techniques allows us to access different geographically distributed data collection and perform text mining task in parallel fashion [11].

They have proposed some applications on grid-based data mining infrastructure and these are Knowledge Grid (K-Grid) and NASA's Information Power Grid (IPG). Both applications are the input to the data mining algorithm to extracting the new knowledge. They used the Open Source Grid Toolkit as Grid middleware in the Grid architecture as shown in the figure. 2 [11]. In that the XML and HTTP services are provided by the web service.

To find the suitable resources, user can be submit the instructions to the SVM application and through Globus. The architecture also showed the Parallel of Support Vector Machine (PSVM) algorithm and Sequential SVM algorithm [11].

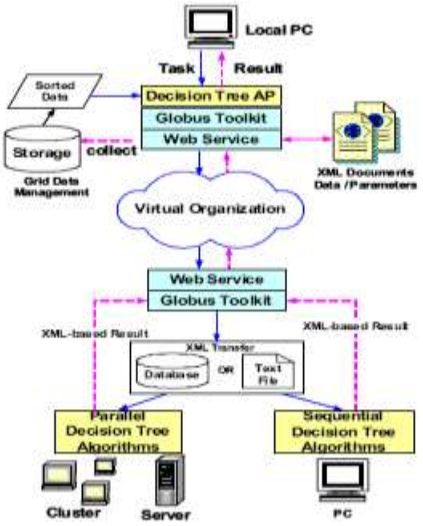

Figure2. Grid-Based SVM [11]

They employed a PSVM algorithm for its parallelism. Here parallel applications can be classified into some programming paradigms which are used to develop parallel programs. For the parallelism, they have shown the task farming paradigm which is also called as Master / Slave consists of two entities

Master and Multiple Slaves. The communication takes place only between the master and slaves [11]. The task farming paradigm can be used either static load balancing or dynamic load balancing. In static load balancing, the distribution of all tasks is performed at the beginning of computation [11].

The grid can play an important role in providing an effective computational support for distributed knowledge discovery applications. For the development of data mining applications on grids, they designed a system called KNOWLEDGE GRID [14], [4]. Mario Cannataro et al. [3]described the Knowledge Grid framework and also presented the toolset which is provided by the Knowledge Grid for implementing distributed knowledge discovery. They have discussed about Knowledge Grid system which is designed for the development of data mining applications on grid [3]. Knowledge Grid is a parallel and distributed software architecture which integrates data mining techniques and grid technologies [3].

Chao-Tung Yang et al.[5]presented the decision tree architecture for applying it to both parallel and sequential algorithm. Decision tree is one common method which is used in data mining to extract the predicted information. They have been proposed some applications on Grid-based data mining infrastructure, such as the Knowledge Grid (K-Grid) [19], NASA's information Power Grid (IPG) [25]. Both are the input on the data mining algorithms used to extract new knowledge. Their goal is to explore the relationship between the Grid architecture and data mining. Here grid is classified into three types: Top-Down, Reciprocal type and Broadcast type. These three types can be used to identify the data mining application model and user equality in Grid Environment [5].

C. *Adaptation SVM*

Bo Geng et al. [1]proposed an algorithm called Ranking adaptation SVM (RA-SVM) which is used to address the difficulties while building an unique ranking model for each domain is both laborious and time consuming for labeling data and training data [1]. Using this algorithm we can adopt ranking model to a new domain, so that the amount of labeled data and training cost is reduced that keeping performance into consideration [1]. Ranking adaptation is closely related to the classifier adaptation which is effective for the many learning problems [6], [7], [13].

In simplest linear form, an SVM is a hyper plane that separates a set of positive examples from a set of negative examples with maximum margin. Thus for different classes they select different distributed nodes, and adapt single trained SVM of such class for specific classification task.

TABLE 1: REVIEW SUMMARY

Reference No.	Applications	Summary	Database Used
[11]	Text mining, Task Farming Paradigm, Static load balancing	SVM is used for the binary classification.SVM is a simplest linear form, in that it is a hyperplane that separates the positive examples from the negative example. The Grid is collection, distribution and sharing of service within dynamic organizations consisting of resources. For multiclass classification, Grid data services are used to access the metadata.	The data sets from hundreds of Tera Bytes to Peta Bytes
[20]	Text classification, Hierarchical classification, Distributional clustering of words	Information theoretic-divisive algorithm is used for word clustering applied to text classification. Support Vector Machines (SVMs) are generally inductive-learning schemes which are used for solving two class pattern recognition problems. Nowadays SVM also give good results in text categorization.	20 Newsgroups and Dmoz data sets
[9]	Classify one billion datapoints in 20-dimensional input space into two classes in some minutes on ten machines	The incremental, parallel and distributed Support Vector machine (SVM) using linear and non-linear kernels. To classifying very large size of datasets on standard personal computers (PC's), they have proposed the Least Square SVM (LS-SVM) for building the incremental, parallel and distributed algorithm which is used to construct new algorithm which is very fast to build incremental, parallel and distributed SVM for classification task.	Ring Norm dataset
[12]	Web searching, Email filtering, text classification	An algorithm which is used for performing active learning with support vector machine (SVM) which means that algorithm is used for choosing which instances to request next. In SVM for induction, a labeled training set of data or a task is used to create a classifier which has a good performance on unseen test data. SVM also used for transduction. Here SVM can perform transduction by finding the hyperplane which maximizes the margin related to both labeled and unlabeled data.	Reuters-21578 data set and the Newsgroups data set
[2]	One-Against-all, One-Against-one and Directed Acyclic Graph (DAG) method	To solve the multiclass SVM in one step, that has variables proportional to the number of classes. Therefore, for multiclass SVMs methods, there is need of binary classifiers to be constructed and the optimization problem. That's why this method is more computationally expensive to solve the multiclass problem than the binary problem with the same number of data.	DNA, Satimage, Letter, and Shuttle datasets
[10]	Kernel function selection and model selection of SVM	Classification is the most important task specially used for the different applications such as text	RSES

		categorization, image classification, data classification etc. The concept of SVM with kernel function selection and model selection which are used to solve the classification problem. Out of all kernel function, the RBF is the most popular kernel function because it has less numerical difficulties. SVMs are related to the supervised learning method which is used for the classification and regression.	data sets
[3]	Knowledge Grid	The grid can play an important role in providing an effective computational support for distributed knowledge discovery applications. For the development of data mining applications on grids They designed a system called KNOWLEDGE GRID. The KNOWLEDGE GRID architecture uses basic grid mechanisms are used to build specific knowledge discovery services on top of grid toolkits and services.	Intrusion Detection of network data
[5]	Knowledge Grid, NASA's Information Power Grid (IPG), PC cluster	Decision tree is one common method which is used in data mining to extract the predicted information. Here grid is classified into three types: Top-Down, Reciprocal type and Broadcast type. These three types can be used to identify the data mining application model and user equality in Grid Environment.	Synthetic datasets with text file
[1]	Margin rescaling and Slack rescaling	Ranking adaptation SVM (RA-SVM) is used to address the difficulties while building a unique ranking model for each domain is both laborious and time consuming for labeling data and training data. Using this algorithm we can adopt ranking model to a new domain, so that the amount of labeled data and training cost is reduced that keeping performance into consideration.	Letor Benchmark datasets

III. Conclusion

In this paper we discussed the Adaptation SVM (A-SVM) algorithm in which we can address the difficulties arising during labeling and training the data of multiclass database. Support Vector Machines (SVMs) are used for binary classification. SVM is one common method used in data mining to extract the predicted information. Sometimes SVM is laborious and time consuming in case of training and labeling the multiclass database that's why we are using A-SVM algorithm. Multiclass database is complex in nature. Accessing this complex database leads to computational problem while doing data analysis. Therefore it is necessary to classify this large multiclass database.

Grid is the distribution, controlling and sharing of data used to solve the computational problem and executing the application in parallel nature. This could be used in attack classification in future.

References

[1] Bo Geng, Linjun Yang, Chao Xu, and Xian-Sheng Hua, "Ranking Model Adaptation for Domain-Specific Search", IEEE Transaction on Knowledge and Data Engineering, 2012.

[2] Chih-Wei Hsu and Chih-Jen Lin, "A Comparison of Methods for Multiclass Support Vector Machines", IEEE Transaction on Neural Networks, 2002.

[3] Mario Cannataro, Antonio Congiusta, Andrea Pugliese, Domenico Talia, and Paolo Trunfio, "Distributed Data Mining on Grids: Services, Tools, and Applications", IEEE Transaction on Systems, Man, and Cybernetics, 2004.

[4] M. Cannataro, A. Congiusta, D. Talia, and P. Trunfio, "A data mining toolset for distributed high-performance platforms," in Proceedings of Conference on Data Mining, Bologna, Italy, 2002.

[5] Chao-Tung ,Yang Shu-Tzu Tsai and Kuan-Ching Li , "Decision Tree Construction for Data Mining on Grid Computing Environments", Proceedings of the 19th International Conference on Advanced Information Networking and Applications ,IEEE, 2005.

[6] J. Blitzer, R. Mcdonald, and F. Pereira, "Domain Adaptation with Structural Correspondence Learning," in Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP '06), pp. 120-128, July 2006.

[7] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for Transfer Learning," in Proceedings of 24th International Conference of Machine Learning (ICML '07), pp. 193-200, 2007.

[8] L. Bottou, C. Cortes, J. Denker, H. Drucker, I. Guyon, L. Jackel, Y. LeCun, U. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of classifier methods: A case study in handwriting digit recognition," in Proceedings of international conference of Pattern Recognition, pp. 77–87, 1994.

[9] Thanh-Nghi Do and François Poulet, "Classifying one billion data with a new distributed SVM algorithm", IEEE, 2006.

[10] Durgesh K. Srivastava, Lekha Bhambhu, "Data Classification Using Support Vector Machine", Journal of Theoretical and Applied Information Technology, 2005 – 2009.

[11] Ali Meligy and Manar Al-Khatib, "A Grid-Based Distributed SVM Data Mining Algorithm", European Journal of Scientific Research, 2009.

[12] Simon Tong and Daphne Kolle, "Support Vector Machine Active Learning with Applications to Text Classification", Journal of Machine Learning Research, 2001.

[13] H. Daume III and D. Marcu, "Domain Adaptation for Statistical Classifiers," Journal of Artificial Intelligence Research, vol. 26, pp. 101-126, 2006.

[14] M. Cannataro and D. Talia, "The knowledge grid," Communi. ACM, vol. 46, no. 1, pp. 89–93, 2003.

[15] Isabelle Moulinier, "Feature Selection: A Useful Preprocessing Step", in Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research, pp. 140-158, 1997.

[16] T. Joachims, "Text categorization with support vector machines: learning with many relevant features", in Proceedings of ECML-98, pages 137-142, 1998.

[17] F. Poulet and T-N. Do, "Mining Very Large Datasets with Support Vector Machine Algorithms", in Enterprise Information Systems V, Camp O., Filipe J., Hammoudi S. et Piattini M. Eds., Kluwer Academic Publishers, pp. 177-184, 2004.

[18] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAG's for multiclass classification," in Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000, vol. 12, pp. 547–553.

[19] S. Orlando, P. Palmerini, R. Perego, F. Silverstri, "Scheduling High Performance Data Mining Tasks on a Data Grid Environment", in Proceedings of Europar, 2002.

[20] Inderjit S. Dhillon, Subramanyam Mallela and Rahul Kumar, "Enhanced Word Clustering for Hierarchical Text Classification", University of Texas at Austin, 2002.

[21] J. Suykens and J. Vandewalle, "Least Squares Support Vector Machines Classifiers", Neural Processing Letters, 9(3): pp. 293-300, 1999.

[22] C. Cortes and V. Vapnik, "Support-vector network," Machine Learning, vol. 20, pp. 273–297, 1995.

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

[24] Thomas H. Hinke, Jason Novotny, "Data Mining on NASA's Information Power Grid", HPDC, 2000.

[25] Pang-Ning Tan, M. Steinbach, and V. Kumar, "Data Mining Addison Wesley", London, 2006.