Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

A Comparative study of SVM and SSVM in Big Data Analytics

[1]Parameswari.K and [2]Keerthi.R,
[1,2]Assistant Professor, Bachelor of Computer Application, Krupanidhi Degree College, Bangalore, India

Abstract: Data analytics is the backbone of all the business today. No business can survive without analyzing the data available. Why Big Data? The Big data analytics bring speed and accuracy. The term Big data is applicable to all the information that can no longer be handled by the traditional techniques. Statistical analysis is widely used analytical tool. In that Support vector machine is most powerful and widely used machine learning technique for classification. But it is not suitable for large dataset because of its complexity in the algorithm. We can apply the SVM to large data set with smoothing technique which is called Smooth Support Vector Machine. In this paper, we have focused on a detailed survey about the different algorithm applied on the Big data analytics and also done a comparative study of SVM and SSVM.

Keywords--*Smoothing, Newton Armijo Algorithm, Statistical analysis.*

I. INTRODUCTION

As the data is increasing now-a-days, we have to deal with large set of data. When we think about large data set, Big Data comes into our mind. Big Data is the emerging technology where we deal with large database. In order to store the data, the traditional RDBMS is not sufficient. To store data, we need HDFS. The biggest challenge now is how to analyze such large data set. In data mining, plenty of algorithms are there to analyze the data. As we dealing with Big data, there is a need to find which algorithm will be best suited for the big data analytics. The best statistical learning algorithm is the SVM, which is widely used in the machine learning for both classification and regression. As SVM is supervised learning technique, it is used for the generalization of data. The problem with SVM is the it is not suited for large set of data. We prefer smoothing technique on the SVM and the term is called Smooth Support Vector Machine. In this paper, Newton Armijo algorithm is used for the smoothing on SVM, we also did a comparative study of the statistical learning algorithm and proposed that SSVM is more applicable for analyzing large data set.

II. HDFS

Hadoop framework is open-source software which encourages distributed application. It allows user application to communicate and work with several independent computer nodes and terabytes or even petabytes of data. Goggle introduced a Google File System (GFS) and Google's MapReduce white papers in the year 2003 and 2004 respectively. The most important characteristics of Hadoop framework are it partitions the data into thousands of machines and execute it in parallel manner. The Hadoop cluster can be setup by simply using commodity hardwares. These commodity servers can process large scale data efficiency. The Hadoop framework works with two main components. These two main components are Hadoop Distributed File System (HDFS) and MapReduce distributed programming model. HDFS is a distributed and scalable file system for Hadoop framework. HDFS stores all its metadata to its devoted server known as NameNode also called master node. NameNode is the first node through which the user communicates to perform any input and output to the Hadoop cluster. There is only one master node in a Hadoop cluster and it should be the most reliable node of the whole cluster because without NameNode the whole cluster becomes unserviceable. It is the single point of failure of whole system. The actual data is stored in DataNodes also called slave nodes. DataNodes are responsible to process read and write operation and also the creation, deletion and replication of data block under the guidance of NameNode.

The architecture of Hadoop MapReduce programming model shows how the input is divided into logical chunks and partitioned into various separate sets. These sets are then sorted and each sorted chunks are passed to the reducer. MapReduce model implements Mapper and Reducer interfaces to implement the map and reduce function.

A. Mapper

Here Map function takes the input in the form of <key, value> pairs and generates a set of <key, value> pairs as an intermediate result. Here the term key corresponds to the unique group number associated with each value. And the term value is the actual data related to the process. MapReduce programming model merge the intermediate results with similar key and sends the output to the reduce function.

B. Reducer

Here Reduce function is also defined by the user as per their requirement. Reduce function takes the intermediate <key,value> pair and merges this <key value > pairs to get the final set of values. Programmers are required to set only the correct set of<key,value> pairs . Mapreduce framework can correctly combine the values in one set having similar key together.

III. SUPPORT VECTOR MACHINE

Support vector machine is a statistical learning algorithm which is widely used in the machine learning for classification and regression. It is a fundamental tool for classification in machine learning and data mining. SVM is powerful because it uses supervised learning model for classification and regression. The main working of SVM is separating the data of two classes by a hyper plane. SVM can be applied for both linear and non-linear.

IV. SMOOTH SUPPORT VECTOR MACHINE

The smoothing method is used for solving mathematical problems. These techniques can be applied to reframe the SVM for classification of the patterns. This method of applying smoothing technique to svm is called smooth Support Vector Machine. What is the need of smoothing techniques. Because the normal SVM is not well suited for large data set. For this reason only, we are switching to SSVM. On larger problems, SSVM is faster compared to other statistical learning algorithm. The smoothing technique can be applied to SVM with the help of Newton Armijo Algorithm.

The Kernel function value defines the product of two training points in the feature space. The kernel function map input data from input space to feature space. There are different types of kernel, polynomial and Gaussian kernel.

The svm as an unconstrained minimization problem,

$$min\frac{c}{2}k\emptyset k_2^2+\frac{1}{2}(k\omega k_2^2+b^2)$$
$$s+D(A\omega+eb)+\emptyset$$

Smooth the plus function:

$$p(x;M)=x+\frac{1}{i}\text{lo}$$

Smooth support vetor machine is used for replacing the plus function in the non smooth svm by the smooth gives our SSVM. Newton-Armijo Method is used for the quadratic approximation of the SSVM. The sequence generated by solving a quadratic approximation of SSVM, converges to the unique solution of SSVM at a quadratic rate.

Newton-Armijo Algorithm is globally and quadratically converge to unique solution in a finite number of steps.

$$D_i(\omega;b)=\frac{c}{2}kp\left((ea\ D(Aw+eb));i\right)k_2^2+\frac{1}{2}(kwk_2^2+b^2)$$

Start with any. Having =0; else

i) Newton Direction:

$$r^2D_i(w^i;b_i)d^i=arD_i(w^i;b_i)^c$$

ii) Armijo Stepsize:

$$(w^{i+1};b_{i+1})=(w^i;b_i)+X_id^i$$

Such that Armijo's rule is satisfied

V. RESULT AND ANALYSIS

For the experiment purpose, we have taken the the kaggle dataset for Heart Disease prediction, the experiment is carried out in Anaconda python. The dataset is collected from UCI machine learning repository having 48842 instances having 14 attributes and 2 class labeled. It is a binary class problem having two class label denoted by +1 and -1. The whole dataset is divided into two parts. The training data file contains 32542 instances and 16300 instances. Here the experiment is carried out by taking RBF kernel function, penalty parameter C=1 and σ =0.01.

No. of nodes	Training time(sec)	Accuracy (in %)
1	499.5	84.2
2	294.56	83.4
3	237.1	88.4
4	235.65	84.10

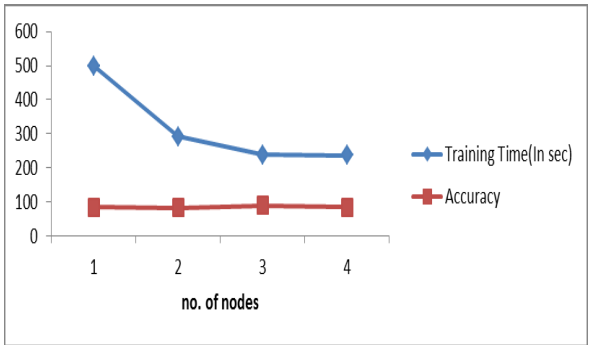

Fig 3: Comparison between Training and Accuracy

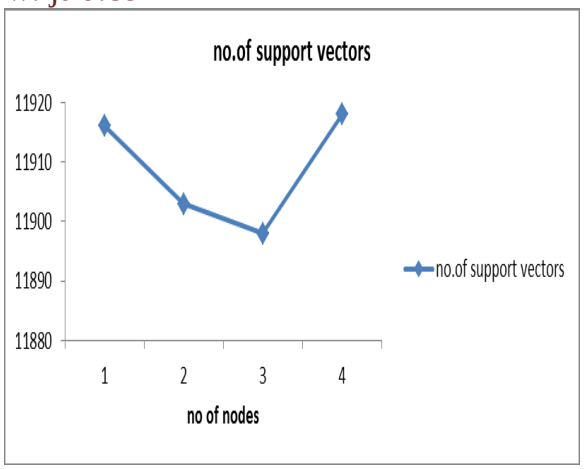

Fig 4: SVM after Smoothing

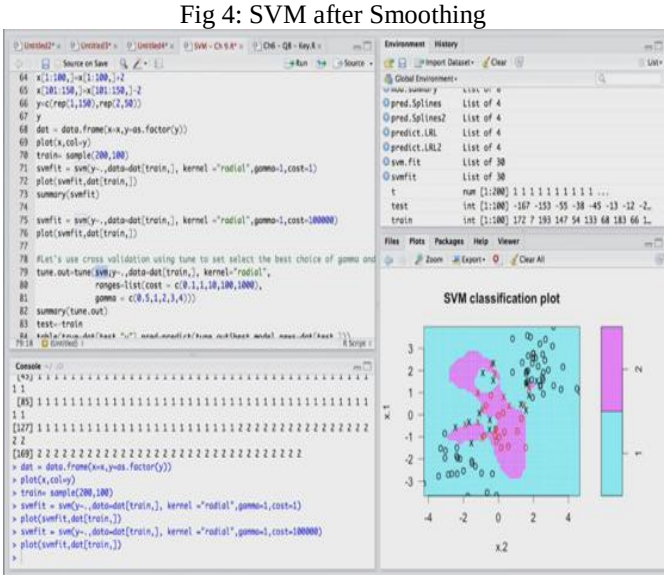

Fig 5: python code for SSVM

CONCLUSION

Data mining is still a big research area for large scaled data. Support Vector Machine is considered as the most effective classifier. SVM classification model depends on the count of support vectors generated by the support vector classifier. The number of support vectors is directly proportional to the required memory which is used to store the support vectors. Most commonly used sequential SVM is difficult to work with large scale data set. As the SVM is not supporting large scale dataset, the smoothing technique is applied to the dataset which is better compared to the normal SVM. In this paper, we have taken large dataset which is trained by both SVM and SSVM, then the result is shown in the paper.

References

[1] L. Povoda, R. Burget, and M. K. Dutta, "Sentiment analysis based on Support Vector Machine and Big Data," *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, 2016.

[2] J.-D. Shen, "New smooth support vector machine for regression," *2012 International Conference on Machine Learning and Cybernetics*, 2012.

[3] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," *Twenty-first international conference on Machine learning - ICML '04*, 2004.

[4] X.-W. Chen, "Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines," *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003.*

[5] Z. Camlica, H. Tizhoosh, and F. Khalvati, "Medical Image Classification via SVM Using LBP Features from Saliency-Based Folded Data," *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015.