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

Transformative Approach Using Multiple Kernels SVM – Map Reduce for Sentiment Analysis Based Text Mining in Big Data

[1]M.V.Anish Kumar, [2]Dr.K.Mahalakshmi, [3]P.Sherubha and [4]K. Narmatha
[1]Student, [2]Professor, [3,4]Asst. Professor
[1,2,3,4]Department of Information Technology

Abstract**--** Sentiment Analysis is the process for determining the semantic orientation of the reviews. There are many algorithms existing for the sentiment classification. Support Vector Machines (SVM) are a specific type of machine learning algorithm used for many statistical learning problems, such as text classification, face and object recognition, handwriting analysis, spam filtering and many others. We have studied the SVM as the recent machine learning method for sentiment classification, this method later suppressed by using feature extraction method. In this paper we are extending SVM and investigating the method by adding the parallel processing methods of sentiment classification such as MapReduce and Hadoop. The combinational evaluation method of SVM with and without MapReduce is presented in this work.

Index Terms**--** Sentiment Analysis, Support Vector Machine (SVM), Text Mining, Feature Extraction, MapReduce, Hadoop.

I. INTRODUCTION

With the evolution of web technology, there is a large amount of data available in the web for the internet users. These users not only use the available resources in the web but also give their suggestions and feedbacks which are much essential to organize and analyze their views for better decision making. In the real world, organizations and businesses always want to find consumer or public opinions about their products and services. Individual consumers also want to know the opinions of existing users of a product before purchasing it and others opinions about political candidates before making a voting decision in a political election. Nowadays, if one wants to buy a consumer product, one is no longer limited to asking one's friends and family for opinions because there are many user reviews and discussions in public forums on the Web about the product. Due to a large collection of opinions on the Web, some form of summary of opinions is needed. Sentiment analysis has grown to be one of the most active research areas in natural language processing. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole. Recently many researchers found that sentiment classification accuracy is mainly affected by decision function used in machine learning methods. We simply used Support vector machine to analyze positive and negative opinions [2].SVM is a useful technique for data classification. In this paper our main aim is to investigate the kernels of SVM and improve further its performance by using the Hadoop and MapReduce. Here mainly MapReduce parallel programming model is presented along with SVM; propose a MapReduce and the Hadoop distributed classification method, and presented its practical evaluation.

II. RELATED WORKS

In this section, Support vector machine, Kernels and Map Reduce programming model are introduced briefly.

A. Support Vector Machine

Support vector machines were introduced in (Vapnik) and basically attempt to find the best possible surface to separate positive and negative training samples. In this module, a document is composed of sentences and a sentence is composed of terms, it is reasonable to determine the semantic orientation of the text from terms. SVM has been shown to be highly effective in traditional text categorization. SVM measure the complexity of hypothesis based on the margin with which they separate the data instead of the number of features. One remarkable property of SVM is that their ability to learn can be independent of the dimensionality of the feature space. To construct a feature vector of the document stop words are removed first and then each distinct word in the document is used to represent a feature [9] [4]. Support Vector Machines (SVMs) are supervised learning methods used for classification. [14] In this work, SVM is used for sentiment classification. Support vector machines perform sentiment classification task on review data. The kernel function plays a critical role in SVM and its performance. Here use RBF kernel for classification in high dimensional. Radial basis functions (RBF) have received significant attention. [16]. It supports multi-class classification by employing SVM to perform the classification.

B. Feature Extraction

Mahalakshmi et al. (2015) proposed a hybrid optimization of SVM for improved classification results.[16]. Have also used a hybrid technique namely optimization and clustering based on the Support Vector Machine and Artificial Bee Colony as well as Differential Evolution. Their proposed hybrid ABC- DE with ABC based feature selection/extraction shows high classification accuracy of 90.54%.

C. Map Reduce Programming

MapReduce programming model was proposed in 2004 by the Google, which is used in processing and generating data sets implementation. This framework solves many problems, such as data distribution, job scheduling, fault tolerance, machine to machine communication, etc. Hadoop Map Reduce is a programming paradigm and software framework which is used for writing applications that rap-idly process data in parallel on large clusters of compute nodes [15]. Map Reduce is a programming model for data processing and is used to write programs that run in the Hadoop environment. Combine Hadoop Map Reduce with SVMs to develop a methodology for managing data sets. It is also highly scalable and also improves the accuracy in categorizing [1] [6].

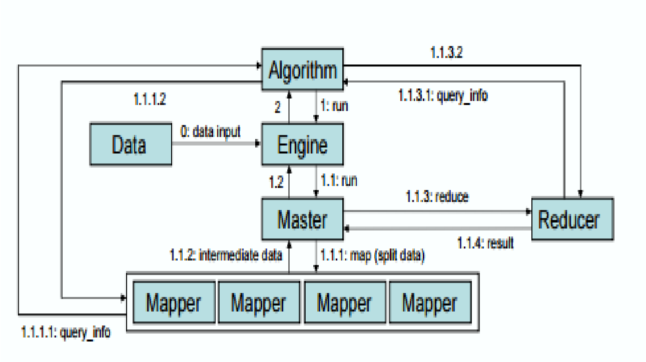

Figure 1: Multi core Map-Reduce framework

III. PARALLELIZED SVM LEARNING ALGORITHM

In this section, parameters of SVM will be analyzed firstly and then based on the related study, a parallelized SVM learning algorithm based on Map Reduce is proposed. For SVM, the selection of kernel function has a significant impact on the performance. RBFSVM, which Gaussian function is taken as a kernel function, shows a strong learning ability and is used in this paper. Performance analysis based on Cross-Validation accuracy where the accuracy rate gives the measure for classification performance.

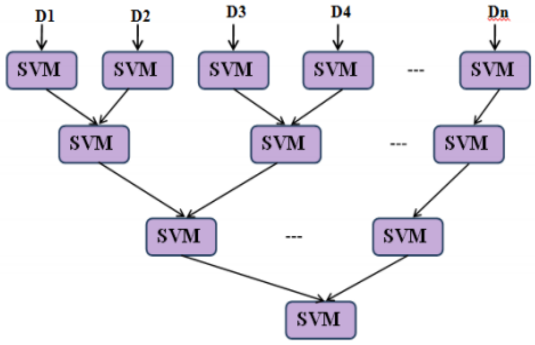

Figure 2: Architecture of Parallel SVM

The Map Reduce framework is inspired by map and reduce functions commonly used in functional programming. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Figure 3 shows the data flow of the various stages of Map Reduce. Hadoop is an open source platform based on the Map Reduce framework, which can be applied in huge data mining well.

Map Reduce programming model, by map and reduce function realize the Mapper and Reducer interfaces. They form the core of task.

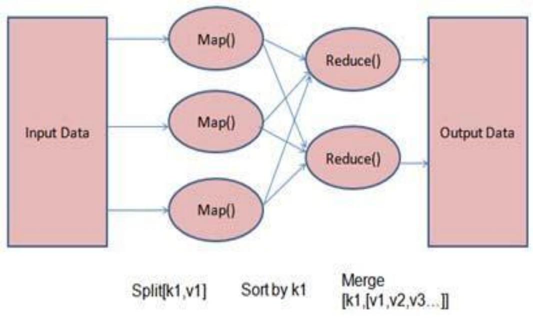

Figure 3: The Map Reduce programming model

A. Mapper

Map function requires the user to handle the input of a pair of key value and produces a group of intermediate key and value pairs. <Key, value> consists of two parts, value stands for the data related to the task, key stands for the "group number” of the value. Map Reduce combine the intermediate values with same key and then send them to reduce function.

B. Reducer

Reduce function is also provided by the user, which handles the intermediate key pairs and the value set relevant to the intermediate key value. Reduce function mergers these values, to get a small set of values. The process is called "merge ". But this is not simple accumulation. There are complex operations in the process. Reducer makes a group of intermediate values set that associated with the same key smaller. In MapReduce framework, the programmer does not need to care about the details of data communication, so <key, value> is the communication interface for the programmer in Map Reduce model. <Key, value> can be seen as a "letter", key is the letter’s posting address, value is the letter’s content. With the same address letters will be delivered to the same place. Programmers only need to set up correctly <key, value>, Map Reduce framework can automatically and accurately cluster the values with the same key together. Map tasks and Reduce task is a whole, cannot be separated. They should be used together in the program. Map Reduce algorithm process is described as follows:

a. Map phase

Step 1: Hadoop and Map Reduce framework produce a map task. Each <Key, Value> corresponds to a map task.

Step 2: Execute Map task, process the input <key, value> to form a new <key, value>. This process is called "divide into groups". That is, make the correlated values correspond to the same key words. Output key value pairs that do not required the same type of the input key value pairs. A given input value pair can be mapped into 0 or more out-put pairs.

Step 3: Mappers output is sorted to be allocated to each Reducer.

b. Reduce phase

Step 4: Shuffle. Input of Reducer is the output of sorted Mapper. In this stage, Map Reduce will assign related block for each Reducer.

Step 5: Sort. In this stage, the input of reducer is grouped according to the key (because the output of different map-per may have the same key). The two stages of Shuffle and Sort are synchronized.

IV. RESULTS AND DISCUSSION

The metrics that are commonly used for evaluating the effectiveness of machine learning methods are precision, recall, F-measure and accuracy.[16] In order to find out this performance metrics we have to do the understanding of if the classification of a document was a true positive (TP), false positive (FP), true negative (TN), or false negative (FN) as showing in below Confusion matrix Table 1[4].

Table 1: Confusion matrix

		Predicted	
		Positive	Negative
Actual	Positive	A	B
	Negative	C	D

The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Precision (P) is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$\text{Precision} = \frac{tp}{tp + fp}$$

The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$\text{Recall} = \frac{tp}{tp + fn}$$

An F-measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

A. Scenarios

Following performance evaluation graphs are showing the performance of dataset for all methods. This configuration is applied for SVM, Parallel SVM, SVM-Map Reduce and Parallel SVM-Map Reduce algorithms. In this paper, we collected the movie reviews from Internet Blogs. In the preprocessing phase of the hypertext markup language (HTML) document, HTML-tag-removal process is required to extract the text information. Training data are necessary for SVM to train a classification model, and manual classification is performed to classify the training reviews into positive or negative reviews. In this performance of the system is analyzed by increasing the dataset from 100 to 400 reviews. We have evaluated our proposed approach with dataset, which is available at http://www.cs.cornell.edu/people/pabo/movie-review-data/polarity_html.zip.

Table 2: Table for Accuracy performance

Records	100	200	300	400
SVM	0.65	0.66	0.70	0.65
Parallel SVM	0.78	0.75	0.73	0.86
SVM-Map Reduce	0.73	0.67	0.69	0.65
Parallel SVM-Map Reduce	0.90	0.82	0.91	0.92

Table II show accuracy performance of SVM, Parallel SVM, SVM-Map Reduce and Parallel SVM-Map Reduce algorithms is evaluated by varying datasets from 100 to 400 records. The Figure 4 show accuracy graph for Algorithms scenario in which Parallel SVM-Map Reduce have better accuracy as compare to other algorithms.

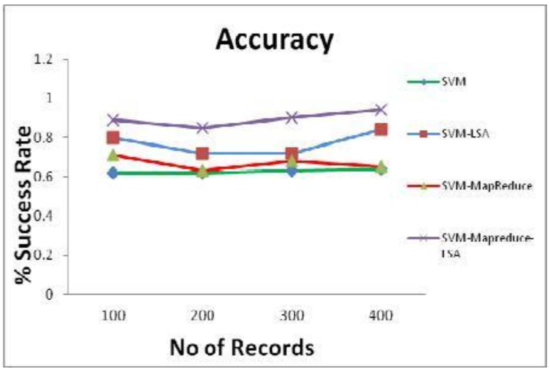

Figure 4: Accuracy curve for movie review dataset

Table 3: Table for Time in seconds

Records	100	200	300	400
SVM	11	20	31	42
Parallel SVM	9	7	9	9
SVM-Map Reduce	10	6	9	9
Parallel SVM-Map Reduce	5	6	9	9

Table III show time performance of SVM, SVM- Parallel SVM, SVM-Map Reduce and Parallel SVM-Map Reduce algorithms is evaluated by varying datasets from 100 to 400 records. The Figure 5 show time curve, in which Parallel SVM-Map Reduce take minimum time as compare to other algorithms.

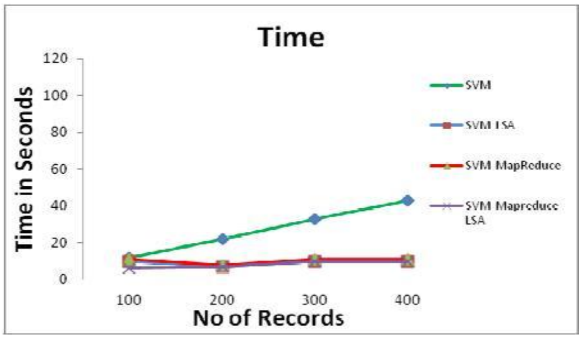

Figure 5: Time curve for movie review dataset

Table 4: Table for Precision

Records	100	200	300	400
SVM	0.63	0.64	0.64	0.64
Parallel SVM	0.81	0.74	0.73	0.90
SVM-Map Reduce	0.72	0.64	0.66	0.67
Parallel SVM-Map Reduce	0.89	0.87	0.88	0.96

Table IV show Precision performance of SVM, Parallel SVM, SVM-Map Reduce and Parallel SVM-Map Reduce algorithms is evaluated by varying datasets from 100 to 400 records. The Figure 6 show precision graph for Algorithms scenario in which Parallel SVM-Map Reduce have better performance as compare to other algorithms.

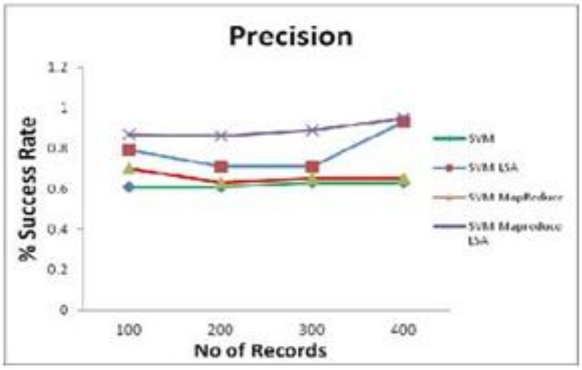

Figure 6: Precision curve for movie review dataset

Table 5: Table for Recall

Records	100	200	300	400
SVM	0.67	0.66	0.67	0.66
Parallel SVM	0.85	0.75	0.73	0.72
SVM – Map Reduce	0.72	0.74	0.73	0.73
Parallel SVM-Map Reduce	0.90	0.84	0.93	0.91

Table V show Recall performance of SVM, Parallel SVM, SVM-Map Reduce and Parallel SVM-Map Reduce algorithms is evaluated by varying datasets from 100 to 400 records. The Figure 7 show recall graph for Algorithms scenario in which Parallel SVM-MapReduce have better performance as compare to other algorithms.

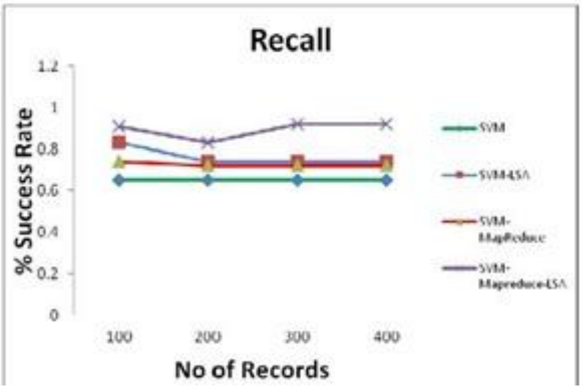

Figure 7: Recall curve for movie review dataset

Table 6: Table for F-Measure

Records	100	200	300	400
SVM	0.62	0.62	0.63	0.63
Parallel SVM	0.80	0.72	0.72	0.82
SVM-Map Reduce	0.71	0.67	0.68	0.68
Parallel SVM-Map Reduce	0.88	0.84	0.90	0.93

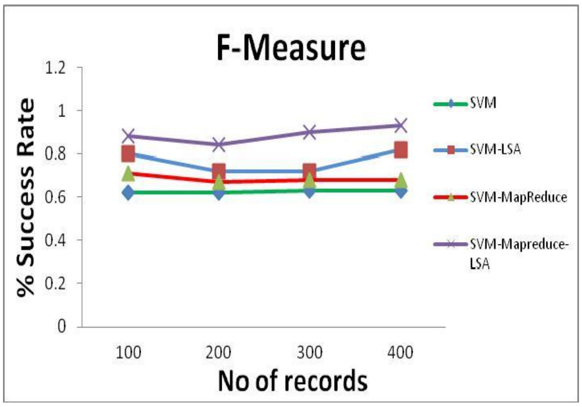

Figure 8: F-Measure curve for movie review dataset

We have observed all the expected results for precision, recall, F-measure, time and accuracy rates. We claim that from above results, our proposed or extended method of sentiment classification is more accurate and efficient as compared to SVM method and hence we will further like to do analysis and investigation over the same.

CONCLUSION

Sentiment classification is applied to the reviews, and summarization is based on sentiment-classification results. In this paper, we have discussed first most commonly used SVM, Parallel SVM, and then most recent is MapReduce based approach for sentiment classification. However we found that there is still place for improvement in terms of accuracy and efficiency of SVM method, and hence we have proposed to add the approach of Hadoop together with SVM to improve the accuracy and efficiency of sentiment classification approach. We will further like do carry more investigation over the same to achieve better accuracy level.

References

[1] Jun Zhao, Zhu Liang, and Yong Yang,” Parallelized Incremental Support Vector Machines Based on MapReduce and Bagging Technique” IEEE Inter-national Conference on Information Science and Technology Wuhan, Hubei, China; March 23-25, 2012.

[2] L. Zhuang, F. Jing, and X.-Y. Zhu, “Movie review mining and summarization,” in Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. 2006, pp. 43–50.

[3] Y. Lu, C. Zhai, and N. Sundaresan, “Rated aspect summarization of short comments,” in Proc. 18th Int. Conf. World Wide Web, New York: ACM, 2009, pp. 131–140.

[4] Jayashri Khairnar and Mayura Kinikar,”Machine Learning Algorithms for Opinion Mining and Sentiment Classification”, Interna-tional Journal of Scientific and Research Publications (IJSRP), Vol-ume 3, Issue 6, ISSN 2250-3153, June 2013.

[5] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proc.10thACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2004, pp.168–177.

[6] Godwin Caruana, Maozhen Li, and Man Qi,” A MapReduce based Parallel SVM for Large Scale Spam Filtering” IEEE, Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011.

[7] Sergio Herrero-Lopez, “Accelerating SVMs by integrating GPUs into MapReduce Clusters” IEEE, 2011.

[8] Mahalakshmi, K., and R. Prabhakar. "Performance Evaluation of Non Functional Requirements." Global Journal of Computer Science and Technology 13.8 (2013)

[9] B.Pang, L.Lee, and S.Vaithyanathan, “Thumbs up: Sentiment classification using machine learning techniques,”inProc.ACL-02Conf.Empirical Methods Natural Lang. Process., 2002, pp. 79–86.

[10] S. H. Choi, Y.-S. Jeong, and M. K. Jeong, “A hybrid recommendation method with reduced data for large-scale application,” IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 5, pp. 557– 566, Sep. 2010.

[11] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou, “Movie Rating and Review Summarization in Mobile Environment”, IEEE VOL. 42, NO. 3, MAY 2012.

[12] (2001). LIBSVM: A library for support vector machines [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[13] V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives,” in Proc. 8th Conf. Eur. Chap. Assoc. Comput. Linguist, Morristown, NJ: Assoc. Comput. Linguist. 1997, pp. 174–181.

[14] Mahalakshmi, K., Prabhakar, R., & Balakrishnan, V. (2014). “Optimizing Support Vector Machine for Classifying Non Functional Requirements”. Research Journal of Applied Sciences, Engineering and Technology, 7(17), 3643-3648. Hadoop. http://hadoop.apache.org/

[15] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In Proceedings of the 6th USENIX Symposium on Operating Systems Design and Implementation, pag-es 137-149, 2004.

[16] Mahalakshmi, K., Prabhakar, R., (2015). “Hybrid Optimization of SVM for Improved Non-Functional Requirements Classification”. International Journal of Applied Engineering Research, Vol. 10 No.20 pp-20157-20174.

[17] Mahalakshmi, K., R. Prabhakar, and V. Balakrishnan. (2015) "Kernel Optimization For Improved Nonfunctional Requirements Classification." Journal of Theoretical and Applied Information Technology 60.1: 64-72.