Volume 5, Issue 4, April 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Hybrid Classification Algorithm Based on Subtractive Clustering and Incremental Learning

Shweta Taneja, Charu Gupta, Swetambari Sinha, Soniya Singh

[#]CSE Department, Bhagwan Parshuram Institute of Technology,

Guru Gobind Singh Indraprastha University,

New Delhi, India

Abstract: As real world data is increasing day by day, so there is a great need of improved classification method to handle it. Therefore, a hybrid classification algorithm based on subtractive clustering and incremental learning is proposed in this work. The algorithm is divided into two phases. In the first phase of the algorithm, subtractive clustering is used to select the cluster centers and disposing of redundant data samples. In the second phase new samples are introduced. The angle between the new sample and the old samples is used to find out the support vectors and accordingly the hyper plane is optimized .We have implemented our algorithm on standard Iris data sets. Experimental results have shown that our proposed algorithm has equal accuracy as that of the conventional support vector machine algorithm and is also faster in speed. The memory requirement of our proposed algorithm is also better than the conventional SVM algorithm.

Keywords: Support Vector Machine(SVM), Subtractive Clustering, Incremental Learning.

I. INTRODUCTION

With the development of new computing technologies and different software and hardware, a huge number of data sets having high dimensions is also continuously getting stored in databases. So, the need of the time is to classify these large datasets into proper class labels. Different data classification methods are needed to analyze and understand these large data sets. Among all these classifiers, Vapnik's Support Vector Machine (SVM) classifier [1],[2] ,which obtains good accuracy on large data sets is the most popular.

Support Vector Machine (SVM) is a kind of machine learning method which is based on statistical learning and structural risk minimization principle [3].SVM is a supervised learning algorithm that analyzes data and recognizes patterns, and is highly used for text categorization, image classification and hand –written character recognition. We have chosen SVM over other classification methods like Naive Bayes,K-Nearest Neighbor etc. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

Though, SVM obtains good accuracy on large data sets but has the drawback that the running time complexity is $O(n^3)$ and the space complexity is $O(n^2)$[9]. To overcome these drawbacks, this paper introduces a new algorithm for classification which is based on subtractive clustering and incremental learning. Subtractive clustering scales down the original data sets and incremental learning, where we have applied the angle approach helps in reducing the number of support vectors. With the combination of these two methods, we have proposed an efficient algorithm. We have experimentally tested our algorithm using standard UCI dataset-Iris.

SVM can be used for both linear as well as non-linear classification.Suppose we are having the training set of binary classification task as:

$X=(x_1, y_1),(x_2,y_2)\ldots..(x_n ,y_n)$ ---eq. 1

i.e. $X =\{x_i ,y_i \}^n_{i=1}$ where $x_i \in R^d$ and $y_i \in (+1,-1)$. ---eq. 2

The set X can be separated by maximum margin hyperplane:

$w.x-b = 0$ ---eq. 3

where . denotes the dot product. The vector w is a Surface normal /normal vector: it is perpendicular to the hyperplane. The parameter $b/\|w\|$ determines the offset of the hyperplane from the origin along the normal vector w.

The aim is to choose w and b so as to maximize the margin, or distance between the parallel hyperplanes that are as far apart as possible while still separating the data. These hyperplanes[10] can be described by the equations[10] and shown by the Fig.1 as follow:

$w.x-b= 1$ ---eq. 4

and

$w.x-b= -1$ ---eq. 5

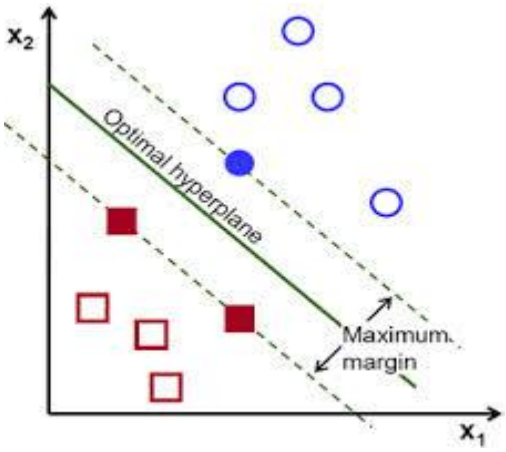

Fig.1 optimal hyperplane

For linear separable case,training SVM yields to solve a quadratic programming problem as follows:

$$\min W(\alpha) = -\sum_{i=1}^{l} \alpha_i + \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i . x_j) \quad \text{--- eq.6}$$

$$\text{s.t} \sum_{i=1}^{l} y_i \alpha_i = 0 \quad \text{---eq. 7}$$

$0 \leq \alpha_i < C \quad i=1,2,\ldots..l$

The classification function of a new datapoint x based on the plane is given by:

Predict(x) =sign (w.x-b) ---eq. 8

This algorithm is faster than conventional SVM algorithm and it also requires less memory space. The rest of the paper is designed as follows: In section-II, we have given the related wok to SVM algorithm. Section-III describes the new proposed algorithm. Section-IV shows the experimental results and comparisons. And finally, section-V states the conclusion and future work.

II. RELATED WORK

SVM, also called as supervised learning models analyzes data and recognizes patterns between them .It can further be used for classification and regression analysis.SVM is widely used for classifying high dimensional data as it is not based on the complete data sets but a subset of it called the support vectors[11].

Clustering is a way of grouping similar data together is widely used for finding the most informative data sets.The accuracy of any classification algorithm highly depends upon the quality of datasets. But, sometimes the dataset gets contaminated with noise by originating from different sources. So, Subtractive Clustering finds the most informative datasets and helps in reducing the large amount of redundant datasets [5].

In [4], SVM with subtractive clustering is given. Subtractive clustering is fast,one-pass algorithm for estimating the number of clusters and the cluster centers in a set of data. This approach increases the accuracy of classification .But SVM alone with subtractive clustering suffers from the problem of running out of time. The time for subtractive clustering reaches peak and the total time for running subclust+svm is over than that of other SVMs. So,we have proposed a hybrid approach where we have combined incremental learning with subtractive clustering to overcome the problem of time and space.

Incremental learning is a good approach for incrementally testing the data sets. When the data is presented to the learning algorithm sequentially in batches; one can compress the data of the previous batches to their support vectors. Then, for each new batch of data, a SVM is trained on the new data and the support vectors from the previous learning step [7].

III. PROPOSED ALGORITHM

We have proposed a novel algorithm. This algorithm is expected to reduce the drawbacks of conventional SVM algorithm. The proposed algorithm is divided into two phase. In the first phase subtractive clustering is used to reduce the massive number of datasets.

Subtractive Clustering is better than other clustering techniques because unlike other clustering techniques the centre here is also a data sample and thus helps in scaling down the massive training sets. And in the second phase of the algorithm incremental learning is used. In incremental learning, an angle approach is applied to find out the support vectors and thus constructing the support vector machine.The proposed algorithm can be shown by the flow diagram in

Fig.2 :
Input data
Subtractive clustering
Cluster centers
Reduced data sets
Search for maximum Potential
Search for minimum angle b/w new data and old data
Support vectors
Optimal hyperplane
SVM Classification

Fig 2: Flow chart of proposed algorithm

The algorithm goes as follows:

1. Let X^n be a set of N data points $x_1, x_2 \ldots x_n$. x_i can be real or vector .
2. Now, compute the potential value for each x_i as

$$P(x_i, x^n) = \sum \exp\left[\frac{- \| x_i - x_j \|^2}{(r_a/2)} \right] \quad i=1,2,3\ldots N \quad \ldots(7)$$

here $0 < r_a <= 1$ is the radius of the cluster

3. Select the data points x_i with the highest potential P_i as the first cluster center denoted by x_1^* and its potential P_1^* and reduce the potential value of remaining data points as

$P_i = P_i - P_i^* \exp[-\|x_i - x_j\|^2 / (r_b/2)^2]$ {here we have taken $r_b = 1.5\, r_a$} ….(8)

4. Then, find the cluster center on the basis of the value of P_K^* / P_1^* . P_K^* is the highest potential from the reduced potential and P_1^* as the first cluster center . If the value of $P_K^* / P_1^* < \epsilon_{up}$ then accept x_k^* as the next cluster center and if $P_K^* / P_1^* < \epsilon_{down}$ then reject it. a

Fig. 3: Optimal Hyperplane between the cluster centers

5. We compute the angle between $(x_N - x_A)$ and the separation plane as $<\alpha'$ and between $(x_N - x_B)$ and the separation plane as $<\beta'$. The more less than angle, the more better.

β α A N
B

Fig. 4: The angles which are between subtraction new sample from neighborhoods and the separation plane

6. $U_i = \dfrac{-2k(x0, xi) + k(xi, xj)}{\left[\sum j \in \delta v\, yj\alpha j \left(k(x0, xj) - k(xi, xj)\right)\right]}$ ...(9)

Here, we have used a variable Ui to minimize the new samples. The new samples x_i, which minimizes the above equation transformed into support vectors. Using these support vectors, hyperplane is drawn.
Our proposed SVM algorithm is better than the conventional algorithm. This proposed algorithm improves the execution time and the memory requirement to a great extend.

IV. EXPERIMENTAL RESULTS

We have used the standard UCI repository Iris dataset[8] to evaluate the performance of our proposed algorithm. This data set is comprised of 150 instances. Each instance is characterized by 4 attributes-sepal length, sepal width, and petal length and petal width and classified as either "Iris-setosa" or "Iris-versicolor" or "Iris-virginica" classes.121 datasets are being selected as training sets randomly and the remaining as testing set.

All the experiments are performed on a Dell, Intel Core i3 Duo processor, 2.66GHz PC with 1 GB RAM and MATLAB 7.10.0.

The experimental result shows better classification performance with increasing values of r_a. But, we see that under the same value of r_a, our proposed algorithm shows time and space requirement. Also the number of support vectors also gets drastically scaled down. The following Tables show the obtained results.
Table 1 shows the results obtained when the value of r_a is 0.5 . Our algorithm shows better results in comparison to conventional SVM algorithm with respect to CP, space time and also the no. of support vectors are also got drastically reduced. Table 2. shows the comparison of the algorithm with conventional SVM when the value of r_a is 0.3. Although, the no. of support vectors, time and space is reduced. But it happens at the cost of Classification Performance Table 3. shows the comparison of the algorithm with conventional SVM when the value of r_a is 0.7. Although, the no. of support vectors, time and space is reduced. But, Classification Performance came out to be equivalently same.

TABLE 1 : Comparison of the proposed algorithm with conventional SVM when r_a=0.5
CP: Classification performance
NS: No. of support Vectors

	SVM(Linear)	Proposed Algorithm
CP	0.9867	0.9864
NS	40	16
TIME(sec)	6	4
SPACE(MB)	460	284

TABLE 2: Comparison of the proposed algorithm with conventional SVM when r_a=0.3
CP: Classification performance
NS: No. of support Vectors

	SVM(Linear)	Proposed Algorithm
CP	0.9835	0.9765
NS	30	12
TIME(sec)	9	6
SPACE(MB)	345	242

TABLE 3 : Comparison of the proposed algorithm with conventional SVM when r_a=0.7
CP: Classification performance
NS: No. of support Vectors

	SVM(Linear)	Proposed Algorithm
CP	0.9867	0.9864
NS	30	12
TIME(sec)	6	4
SPACE(MB)	556	335

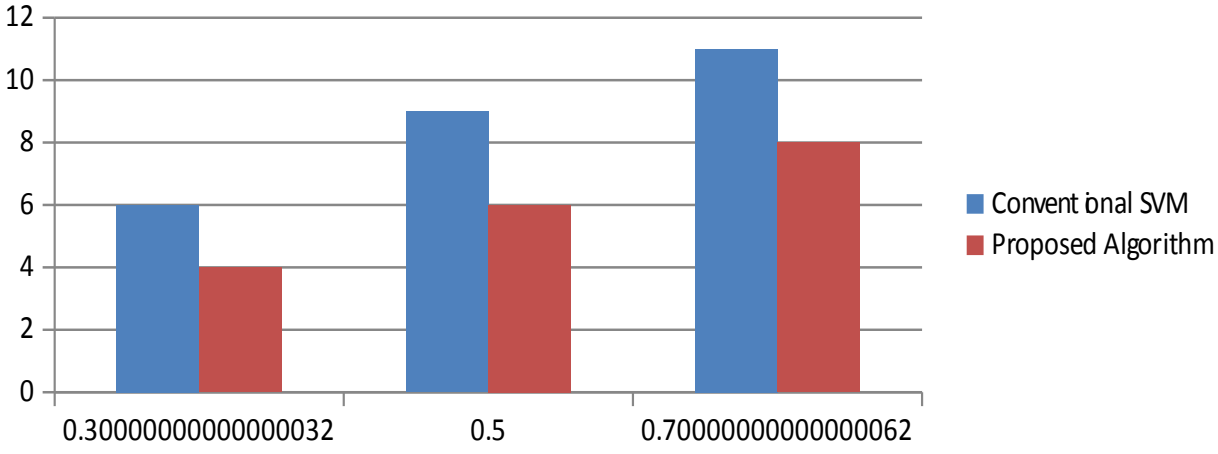

Fig. 5: Graphical Representation based on time with varying r_a

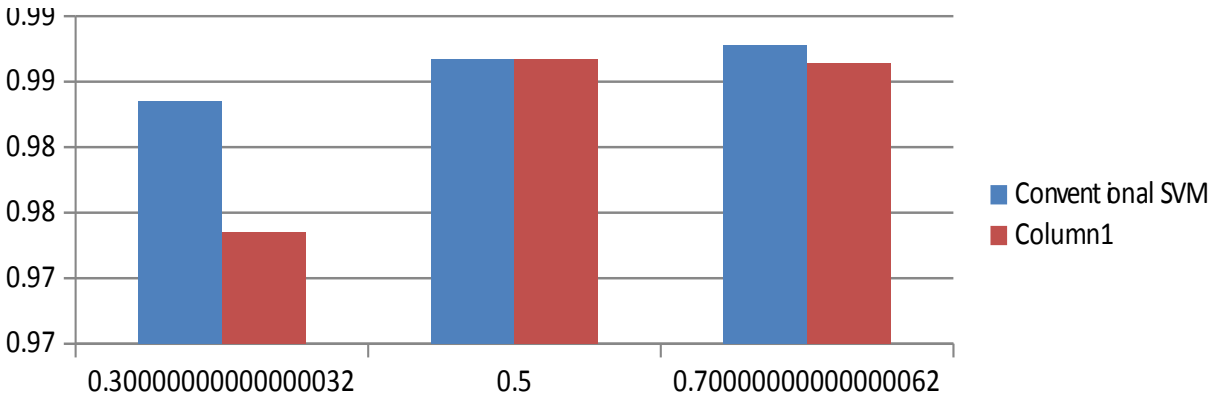

Fig. 6: Graphical Representation based on CP with varying r_a

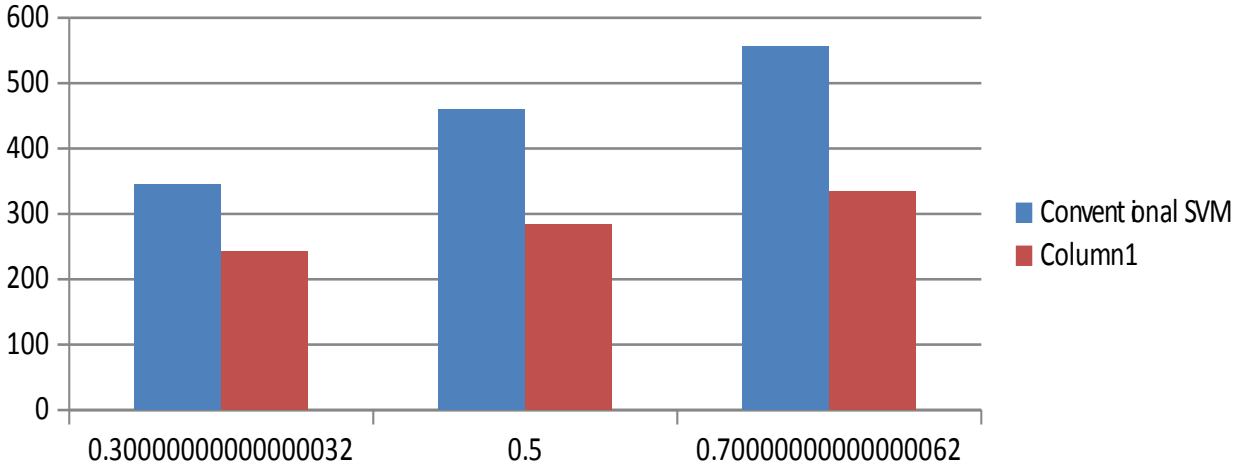

Fig. 6:Graphical Representation based on space(MB) with varying r_a

V. CONCLUSION AND FUTURE WORK:

In this paper, we have tried to solve the time and space complexity problem of conventional SVM . We have introduced an improved algorithm to overcome these problems. This algorithm is a hybrid method based on subtractive clustering and incremental learning is as accurate as other SVM implementations. But the time and the storage requirement is significantly smaller. Furthermore, the number of support vectors also get reduces to a minimum ensuring high precision and thus helps in finding maximum margin hyperplane.

We have experimentally tested our algorithm in MATLAB 7.3.1, using the standard UCI repository dataset-Iris. Experimental results have proved that our proposed algorithm performs better than conventional SVM algorithm. In future, we will implement our proposed algorithm on various other standard UCI datasets and also incorporate the different kernel methods in comparison.

REFERENCES

[1] V.N.Vapnik,Statistical Learning Theory John Wiley and sons,New York, 1998

[2] Fa Zhu, Ning Ye, Dongyin Pan, Wen Ding, "Incremental Support Vector Machine Learning: An Angle Approach",Fifth International Joint Conference on Computational Sciences and Optimization, pp. 288-292

[3] Syed N, Liu H, Sung K. "Incremental learning with support vector machines," Proceedings of the Workshop on Support Vec-tor Machines at the International Joint Conference on Artificial Intelligence (IJCAI-99).Stockholm, Sweden: Morgan Kaufmann, 1999, pp.876-892.

[4] Sheng-Wu Xiong, Xiao-Xiao Niu,Hong-Bing Liu"Support Vector Machine Based On Subtractive Clustering"Proceeding of the Fourth Conference On Machine Learning And Cybernetics , china,2002.

[5] C. D. Doan, S. Y. Liong and Dulakshi S. K. Karunasinghe "Derivation of effective and efficient data set with subtractive clustering method and genetic algorithm" Journal of Hydroinformatics | 07.4 | 2005

[6] Jair Cervantes1, Xiaoou Li2, Wen Yu "SVM Classification for Large Data Sets by Considering Models of Classes Distribution" Jair Cervantes1, Xiaoou Li2, Wen Yu,Sixth Mexican International Conference On Artificial Intelligence,Special Session, Mexico,2008

[7] Stefan R¨uping "Incremental Learning with Support Vector Machines"in IEEE International Conference on Data Mining,2011

[8] The UCI website [Online] http://archive.ics.uci.edu/ml/datasets/Iris

[9] S.Abe,Support Vector Machine for Pattern Classification. London:Springer Verlag,2005

[10] Nello Cristianimi, John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. China Machine Press, 2005

[11] Jair Cervantes, Xiaoou Li, Wen Yu" SVM Classification for Large Data Sets by Considering Models of classes distribution" Sixth Mexican International Conference on Artificial Intelligence, Special Session , 2008

