ANALYSIS OF AN IMPROVED SPATIAL-BOOSTING BASED ON FITNESS FUNCTION

Aakash Panchal[1], Arvind Singh[2]

[1] *Research Scholar, Network Security ,GTU PG SCHOOL, India*
[2] *Reader, Geosciences, PRL,India*

ABSTRACT

"We are data rich, Information poor". Data need to be preprocessed before taking it to the data mining step. Some of the error and missing value are removed by preprocessing techniques. In this paper we are going to proposed fitness function to the existing adaboost algorithm . In this research paper we are going to have data in offline format and applying the adaboost algorithm to enhance the data classification. The S-boosting, just the extends of Adaboost is used for spatial analysis. The research proposed the S-boosting with fitness function to reduce the predication error in the result set. The proposed model and algorithm is shown in this paper. The paper also show result obtain in every iteration and also have graphical analysis for the same.

Keyword : - *data mining ,adaboost algorithm, spatial analysis, fitness function,*

1. Introduction

Due to globalization, there are enormous amount of data and information being collected and stored in the databases in every organization across the world. We can easily find repositories with Terabyte of data. We can find collection of data but difficult to find important pieces of information. From the large data set, Data Mining is also as a process of extracting useful information and finding patterns from huge/large database [3]. Mining also known as knowledge discovery process, knowledge mining from the data, pattern analysis or knowledge extraction [1].

We will discuss more Adaboost in this report. Adaboost stands for "Adaptive boosting" proposed by Yoav Freund and Robert Schapire. They won Gold price for their work in 2013. This algorithm in conjunction with many different types of learning algorithms which is combines into weighted sum that will represent the final output of the boosted classifier. It is adaptive in sense that the subsequent weak learners are tweaked in favor of boosted classifier [2].
Adaboost removes the noisy data and over fitting problem. The single learner may be weak but the performance of each one is just slightly better than the random guessing (<0.5 error) and integration of this entire weak leaner will give new strong classifier.

1.1 Goal of this paper

The goal of this thesis is to reduce the error rate and increase the accuracy in the adaboost algorithm. Well adaboost is known for its adaptive behavior and can be easily integrate with the other algorithm, so it is necessary to increase its accuracy so that result/output for the data mining is correct. To be more precise definition, I will do some mathematics to reduce the value of alpha, which represents the error rate in the algorithm. Instead of using log in the alpha, we are using user defined value and exponential value to reduce the error rate.
.

1.2 Purpose of paper

The purpose of my paper is work on the accuracy parameter in the adaboost algorithms which is widely used in the various applications across industry, education, image processing, etc. Due to its simple and adaptive behavior, it is also used in data mining purpose too.

2. Literature Review

Below are the paper review from various sources.

Table -1: Literature Review

SR. NO.	TITLE	Observation
1	**SpatialBoost: Adding Spatial Reasoning to Adaboost**	The conclusion is that the simple modification of Adaboost to manage the spatial coordinates and its information. Here they proposed the "Spatial classifier" which work on labels of data points.
2	**An Ensemble Approach for Cancerious Dataset Analysis using Feature Selection**	The paper conclude that no single classifier can produce the best result for every dataset. Also no single ensemble techniques can produce consistent results.
3	**An improved Adaboost algorithm based on uncertain functions**	The algorithm which is based on uncertain function, gives higher weight to enhance better performance over the weak classifier. This in turn make the result more accurate.
4	**MODELING AND DATA PROCESSING OF INFORMATION SYSTEMS**	The outcome of this paper shows 96% of processing time of the data are located within the first two interval as shown in graph. As a result, this allows optimization of the system in order to improve the speed of data processing.[4]
5	**Extracting Map Information from Trajectory and Social Media Data**	First they have spatial big data from various data sources. It composed of 2 parts : 1. Locations 2. Text Description. Two features are used: 1. Independent feature (shape, density) 2. Relational feature (topology, evolution). Techniques based on mathematical models, information processing algorithms are plotted to the screen. Geographic entities are extracted as: point element, line and area element. [5]

3. Proposed System

The Adaboost already exists in field of data mining and machine learning. It also reduce the over fitting problem and reduce noise. Due to its many advantages it is widely used in industry and in the field of data mining. In 2006, the S-boosting was concept introduced. Here they add SpatialBoost: Adaboost with Spatial Reasoning. Here I find the lack of fitness function which will decrease the predication error. Also it lack refinement of data i.e. data preprocessing.

Here I will get data from Ecotaxa portal (www.ecotaxa.obs.vlfr) in the form of image or .tsv or summary of both. Then the data will go preprocessing i.e. removing some noise or missing value.

Then applying the feature selection i.e a process for selecting attributes of data which are most relevant to the mining. Well, this will differ from dimensional reduction as this method reduce by creating new combination of attributes, whereas, feature selection include and exclude attributes present in the data without changing them. We will be using harmony search algorithms for this.

Sbooting: This algorithm gets input from the feature selection as well as applying the fitness function to this algorithm. It is the simple extension of adaboost to handle the spatial information.

Fig -1: Proposed System

The model used here allows for classification, regressing, data visualization/finding pattern and finding common characteristics in groups of object. Decision tree is used for classification of unknown objects in the database. Decision tree allows [4]
1. Automated decision making,
2. Universality of description
3. Intelligibility for the user.

Managing the high quality information in information system as per their classification is very important factor for its efficiency. Without relevant information, process of training and research will not be sufficiently effective for the desired result.

The algorithms derived by changing the data distribution to classify the correction over the sample of the dataset. The plus point of the algorithm is "accelerated good fitness function" which calculates weights of the weak classifier. The integration process emphasize on the weak classifier. The value of c1 and c2 are manually specified. The result set shows the reduce in prediction error on the different data set.

3.1 Proposed Algorithm

Input: Training set $\{\mathbf{x_i}, y_i\}_{i=1}^{N}$
Number of iterations T
Output: A strong classifier $H(\mathbf{x})$

1. Initialize weights $\{w_i\}_{i=1}^{N}$ to $\frac{1}{N}$
2. Initialize estimated margins $\{\hat{y}_i\}_{i=1}^{N}$ to zero
3. For $t = 1...T$
   (a) Make $\{w_i\}_{i=1}^{N}$ a distribution
   (b) Set $\mathbf{x_i'} = \{\hat{y}_j | \mathbf{x_j} \in Nbr(\mathbf{x_i})\}$
   (c) Train weak *data* classifier h_t on the data $\{\mathbf{x_i}, y_i\}_{i=1}^{N}$ and the weights $\{\mathbf{w_i}\}_{i=1}^{N}$
   (d) Train weak *spatial* classifier h_t' on the data $\{\mathbf{x_i'}, y_i\}_{i=1}^{N}$ and the weights $\{\mathbf{w_i}\}_{i=1}^{N}$
   (e) Set $\epsilon = \sum_{i=1}^{N} w_i |h_t(\mathbf{x_i}) - y_i|$
   (f) Set $\epsilon' = \sum_{i=1}^{N} w_i |h_t'(\mathbf{x_i'}) - y_i|$
   (g) Set $\lambda_t = \begin{cases} 1 \text{ if } \epsilon < \epsilon' \\ 0 \text{ otherwise} \end{cases}$
   (h) Set $err = \lambda_t \epsilon + (1 - \lambda_t)\epsilon'$
   (i) $\alpha_t = (c_1 - c_3) * exp\left(\frac{(f_t(x) - minf)}{(maxf - minf)}\right) + c_3$ (1)

   Where:

   $c_3 = \frac{(c_2 - c_1 * \exp(1))}{(1 - \exp(1))}$ (2)

   (j) Update examples weights

$$w_i = w_i e^{(\alpha_t(\lambda_t |h_t(\mathbf{x_i}) - y_i| + (1 - \lambda_t)|h_t'(\mathbf{x_i}) - y_i|)}$$

   (k) Update margins $\hat{y}_i$ to be

$$\hat{y}_i = \hat{y}_i + \alpha_t(\lambda_t h_t(\mathbf{x_i}) + (1 - \lambda_t)h_t'(\mathbf{x_i'}))$$

4. The strong classifier is given by $sign(H(\mathbf{x}))$ where $H(x) = \sum_{t=1}^{T} \alpha_t(\lambda_t h_t(\mathbf{x}) + (1 - \lambda_t)h_t'(\mathbf{x}))$

Here we have added the fitness function to the S-boost algorithm to increase its accuracy and reduce prediction error. Based on the value of the c1 and c2 the accuracy increases.

3. Implementation

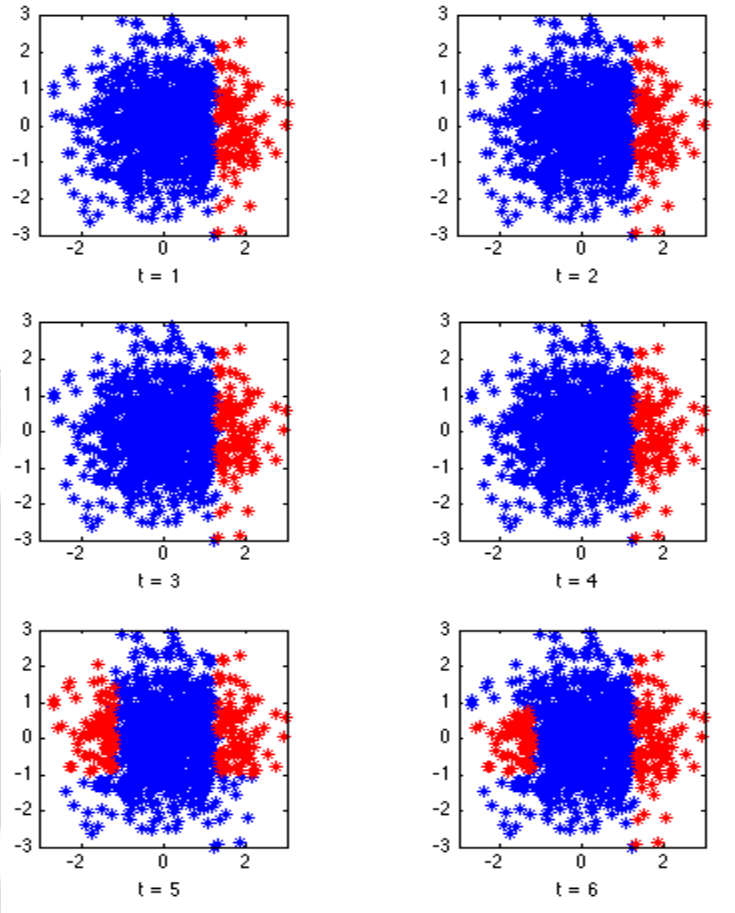
We can use any language to implement the proposed system like C, C++, python, java, matlab etc. I select the python because of the following reasons:

- Open Source
- Community Support
- Easy to use
- Robust
- Easily integrate with other algorithms

To implement my proposed work, I will be using these configurations.

Processor: Intel(R) Core(TM) i3 CPU M330 @ 2.13GHz
RAM : 3 GB
OS : Kali Lunix 2.0
Bit : 32-bit

3.1 Output



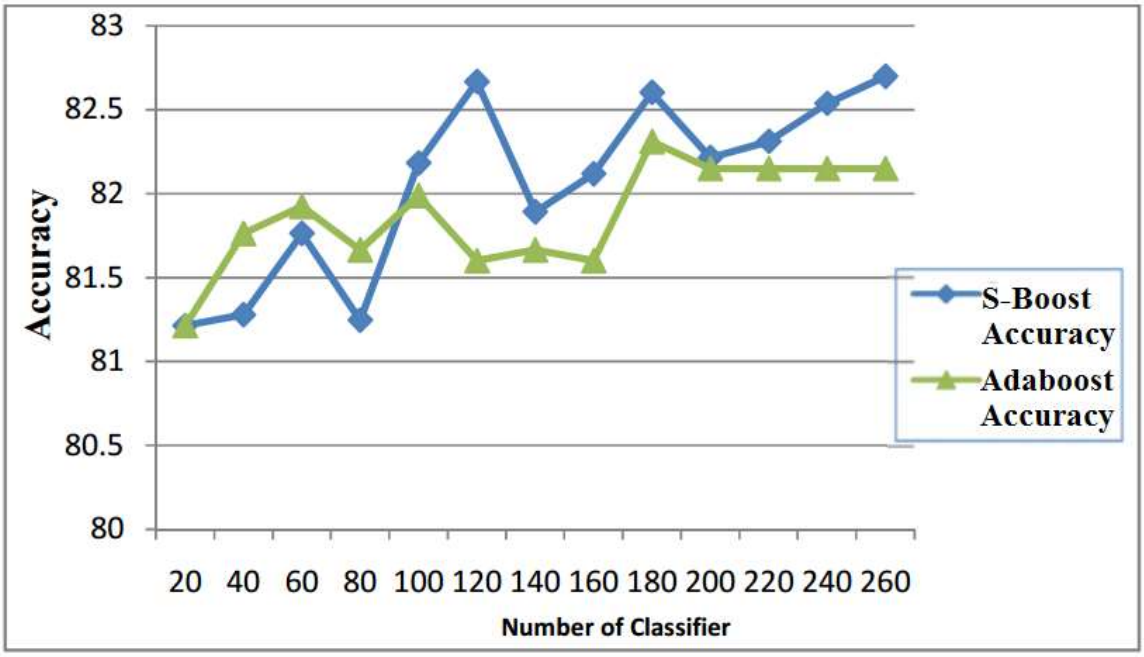
3.2 Observation

I have applied both Adaboost classifier and Proposed S-Boost Classifier on the dataset from UCI repository. For this purpose I have used various dataset 18 feature (attributes) and run the code up to 55 iterations. The table below shows the value after certain iteration and its value for both the classifier.

Iteration	AdaBoost Classifiers	Ada-GA Classifiers (Proposed)	Pop.-Size	Num.of Generation	AdaBoost Accuracy	Ada-GA Accuracy (Proposed)
5	11	4	100	10	81.2137	81.2137
10	26	7	100	10	81.7624	**81.7947**
15	41	13	100	15	**81.9238**	81.8270
20	56	17	100	15	81.6656	**81.7301**
25	71	26	150	20	81.9884	**82.2466**
30	86	22	150	25	81.6010	**82.2466**
35	101	30	150	20	81.6656	**82.0529**
40	116	48	200	25	81.6010	**82.1498**
45	131	54	250	30	82.3112	**82.6985**
50	146	58	350	35	82.1498	82.1498
55	161	73	350	35	82.1498	**82.2466**
Average	**86**	**32**			**81.8485**	**82.0771**

3.3 Analysis

From the below graph it is visible that our proposed algorithm is just slightly better then the existing one.



4. CONCLUSIONS

Today there is a lot of data available. We have to do data analysis/mining on the available data. Our Proposed model helps to do this achieve task. Also it will reduce the error present in the result set.

Also in the field of spatial analysis, this will enhance the spatial data and this modified algorithm can increase its efficiency up to optimal level.

Our Proposed algorithms has proved that it is slightly better than the existing one. The graph in chapter 5 shows that our algorithm is 0.85% more accurate than exiting algorithms.

4.1 Future Work

In the field of spatial analysis and adaboost, we will be working on the lambda to increases the efficiency of the algorithm. Maybe we can get better result while working on lambda parameter and increase the accuracy, hence the output result will get more meaningful.

5. ACKNOWLEDGEMENT

I would like to thanks GTU and PRL to allow me to do this research. Also I am thanks to my guide Dr, Arvind Singh who guides me to carry out my research. Last but not the least, I would heart thanks to Mr. Gaurang Kadiya and Mr. Bhadreshsinh Gohil who motivate me.

6. REFERENCES

[1]. Shai Avidan " SpatialBoost: Adding Spatial Reasoning to AdaBoost," Leonardis, H. Bischof, and A. Prinz (Eds.): ECCV 2006, Part IV, LNCS 3954, pp. 386–396, 2006. © Springer-Verlag Berlin Heidelberg 2006

[2] Shu Xinqing, Wang Pan " An improved Adaboost algorithm based on uncertain functions," International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration 2015

[3] Payal P. Dhakate, Dr. K. Rajeswari , Deepa Abin "An Ensemble Approach for Cancerious Dataset Analysis using Feature Selection" Proceedings of 2015 Global Conference on Communication Technologies(GCCT 2015)

[4] Jiasong Zhao, Lizhen Wang , Xuguang Bao, Yaqing Tan , " Mining Co-location Patterns with Spatial Distribution Characteristics," in 978-1-5090-0690-8/16/ © 2016 IEEE

[5] Galina Panayotova, Georgi Petrov Dimitrov, Pavel Petrov, Bychkov OS , " MODELING AND DATA PROCESSING OF INFORMATION SYSTEMS," ISBN: 978-1-4673-9187-0 ©2016 IEEE