

MINING ON CAR DATABASE EMPLOYING LEARNING AND CLUSTERING ALGORITHMS

Muhammad Rukunuddin Ghalib ^{#1}, Shivam Vohra ^{#2}, Sunish Vohra ^{#3}, Akash Juneja ^{#4}

[#] School of Computing Science and Engineering,
V.I.T University, Vellore-632014, Tamil Nadu, India

¹ruk.ghalib@vit.ac.in

³sunishvhr@gmail.com

²shivamvohra08@gmail.com

⁴aakashnjuneja@gmail.com

Abstract—In data mining, classification is a form of data analysis that can be used to extract models describing important data classes. Two of the known learning algorithms used are Naïve Bayesian (NB) and SMO (Self-Minimal-Optimisation). Thus the following two learning algorithms are used on a Car review database and thus a model is hence created which predicts the characteristic of a review comment after getting trained. It was found that model successfully predicted correctly about the review comments after getting trained. Also two clustering algorithms: K-Means and Self Organising Maps (SOM) are used and worked upon a Car Database (which contains the properties of many different CARS), and thus the following two results are then compared. It was found that K-Means algorithm formed better clusters on the same data set.

Keyword-Data Mining, Naïve Bayesian, SMO, K-Mean, SOM, Car review database

I. INTRODUCTION

Data mining [8][15][18][19][20] are the techniques that enable people with power to research and to manipulate the existing volume of data which is quite large. Data mining algorithms are of various types of which clustering algorithms are also one of the type. Basically, Clustering can be considered the most important unsupervised learning problem; so, it deals with finding a structure in a collection of unlabeled data. There are many types of clustering algorithms but we are using the CAR dataset to differentiate between K-Means Algorithm and SOM because of them having very similar characteristics and implementations.

K-Means Clustering [3][15] is the most vividly used clustering algorithm and is one among the oldest too. It is a prototype based technique performing clustering by providing the prototype in the form of median which is the mean of a group of points and is applicable to objects in a continuous n-dimensional space. K-Means Clustering is a simple but well known algorithms used for grouping objects, clustering. All objects should be represented as a set of numerical features. Each feature should represent some object leaving no object idle. K-Means algorithm is performed in a vector space containing several objects. The algorithm then randomly chooses k points in that vector space, these points serve as the initial centres of the clusters. Now each object is assigned to the centre they are closest to. Usually the distance measure is chosen by the user and determined by the learning task. Now, for the new clusters formed, a new median or centre is found by averaging all the objects in the new cluster. The process of assigning objects and recomputing centres is repeated until it converges. The algorithm can be proven to converge after a finite number of iterations.

Self-organizing feature maps (SOM)[1][2] learn to classify input vectors depending upon how the grouping is performed in the input space. They differ from competitive layers in that neighboring neurons in the self-organizing map learn to recognize neighboring sections of the input space. Thus, self-organizing maps are made to learn both the distribution and topology of the input vectors they are trained on. A self-organizing map consists of component namely nodes or neurons. Each node has a weight vector of the same dimension as the input data vectors and a position in the map space, associated with itself. The usual arrangement of nodes is a two-dimensional regular spacing in a hexagonal or rectangular grid. The self-organizing maps a mapping from a higher dimensional input space to a lower dimensional map space. The procedure for placing a vector from data space onto the map is to find the node with the closest (smallest distance metric) weight vector to the data space vector.

Also Learning algorithms[5][7][16][17] are used as a part of study of this paper in which the two learning algorithms are worked upon a Car review database in order to build a model which can be trained and thus efficiently predict the credibility of any user comment. The two learning algorithms used are: Naive Bayesian Classification and Sequential Minimal Optimization.

A Naive Bayes[4][6][15] classifier, based on applying Bayes' theorem (from Bayesian statistics), is a simple probabilistic classifier with strong (naive) independent assumptions. It can be also more descriptively terms as "independent feature classifier". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is not related to the presence (or absence) of the other entire feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Sequential Minimal Optimization (SMO)[9] is a simple algorithm that can quickly solve the SVM QP problem[18][19] easily without using optimization steps and without extra matrix storage. SMO decomposes the QP problem into a number of QP sub-problems, using Osuna's theorem for ensuring convergence. SMO chooses to solve the smallest optimization problem at every step which is not done in any of the previous algorithms. For the standard SVM QP problem, the smallest possible optimization problem involves two Lagrange multipliers, because the Lagrange multipliers must obey a linear equality constraint. At every step, SMO chooses two Lagrange multipliers to jointly optimize, finds the optimal values for these multipliers, and updates the SVM to reflect the new optimal values

Thus the paper performs its study in two phases:

Phase 1: The two learning algorithms are used to build a model which gets trained on a car review dataset and thus predicts credibility of any user comment. The model is created using a software: LightSide.

Phase 2: The two clustering algorithms are worked upon a different car dataset and thus the results obtained by two are compared. This is done using open source tools called Cluster 3.0 and Java Tree View.

II. METHODOLOGY

The paper performs its study in two phases:

In the first phase a dataset of car reviews is worked upon by two learning algorithms on an open source tool named LightSide. The method is as follows:

The learning algorithms are used to create a model which predicts the credibility of the reviews written by the users in terms of a GOOD or BAD comment.

- 1) First the input car data base is inputted for feature attraction and the input is only in CSV (comma separated variables) format.
- 2) Then extraction of features is performed upon the input file which generally means that it calculates the weight of each bigram and unigram in the file which helps it to calculate the total weight of the review when it is trained.
- 3) Then after the extraction has been done, then a model is built which actually finds the credibility of the reviews. In order to make it possible, the model is trained with a suitable learning algorithm.
- 4) After training is performed, our model is now ready for prediction which is done over a test data.
- 5) Then the prediction of the model is verified as the credibility of the reviews are already pre-known. Thus like this, efficiency of the model is observed.

In the second phase of the work the two different clustering algorithm are executed over the same car dataset due to which we are able to figure out the difference between the two clustering algorithms and also we are able to create clusters of data which are possessing same properties. This is done with a help of an Open Source tool named Cluster 3.0 which uses the two algorithms (K-Mean and SOM) to create the clusters of data and then these clusters of data is represented graphically using another Open Source tool named JavaTree which shows the clusters formed by the two cluster algorithm.

III. IMPLEMENTATION

The implementation is preceded in two phases:

- 1) In the first phase a model is created using Naïve Bayes Classifier and SMO (Sequential Minimal Optimisation)

Which works upon a car review database and thus the model is trained using the dataset in such a way that the model henceforth predicts the nature of a review comment whether it is GOOD or BAD. Also the results attained by the two algorithms is contrasted with one another. LightSide Machine Learning Software is used for creating the model.

Naïve Bayes Implementation

Naïve Bayes Algorithm [13] is as follows:

INPUT: training set T, hold-out set H, initial number of components k0, and convergence thresholds δ_{EM} and δ_{Add} .

```

Initialize M with one component.
k ← k0
repeat
Add k new mixture components to M, initialized using k
random examples from T.
Remove the k initialization examples from T.
repeat
E-step: Fractionally assign examples in T to mixture components, using M.
M-step: Compute maximum likelihood parameters for M,
using the filled-in data.
If logP(H|M) is best so far, save M in Mbest.
Every 5 cycles, prune low-weight components of M.
until logP(H|M) fails to improve by ratio  $\delta_{EM}$ .
M ← Mbest
Prune low weight components of M.
k ← 2k
until logP(H|M) fails to improve by ratio  $\delta_{Add}$ .
Execute E-step and M-step twice more on Mbest, using examples from both H
and T.
Return Mbest.

```

Now, the implementation is done in following steps:

- 1) First the dataset is inputted into the Light Side software [11] in CSV (Comma Separated Variable) format.
- 2) Now, first the features are extracted for Unigrams, Bigrams, Pos Bigrams and Trigrams. Extracting Features refer to that the model calculates weight of each word in the comment which enables it to be trained to predict its characteristic.
- 3) Then a model is created using Naïve Bayes algorithm which trains the model on the input dataset and also gives the confusion matrix which explains the number of good and bad comments in the input file predicted as Good and Bad by the model.

Table 1: Confusion Matrix of Naïve Bayes

Act/Pred	Bad	Good
Bad	103	22
Good	52	72

Table 1, explains that by training the model with Naïve Bayes Algorithm, there were actually 103 BAD comments which were predicted as BAD by the model and 22 were commented as GOOD while in the case of Good comments 52 were predicted as BAD and 72 were predicted as Good by the model.

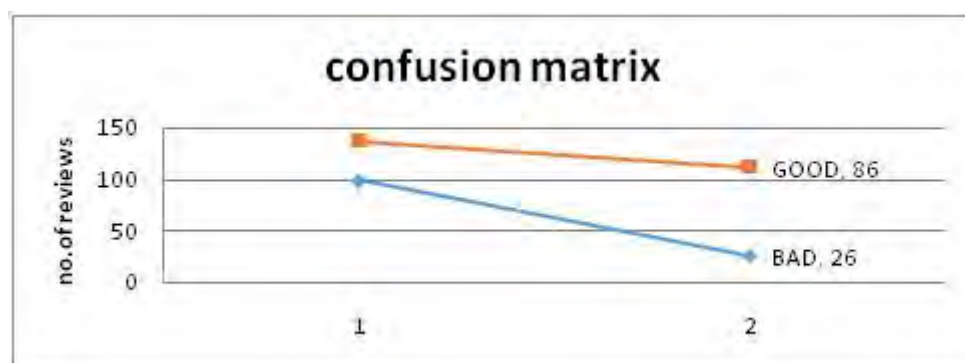


Fig. 1: Confusion Matrix Representation of Naïve Bayes

- 4) Now ore model is ready to predict the comments of a test file which is manually created to test the working of the model.

5) The Test file is loaded into the tool. There are 22 comments in the test file out of which 17 are predicted correct while 5 are predicted wrong. It is believed that as the dataset will be increased in future the accuracy can be increased or the number of wrong predictions can be decreased.

SMO (Sequential Minimal Optimisation) Implementation

SMO algorithm [14] is as follows:

1) Input: C, kernel, kernel parameters, epsilon

2) Initialize b 0

3) Repeat until KKT satisfied (to within epsilon):

Find an example e_1 that violates KKT (prefer unbound examples here, choose randomly among those)

Choose a second example e_2 . Prefer one to maximize step size (in practice, faster to just maximize $|E_1 - E_2|$).

If that fails to result in change, randomly choose unbound example. If that fails, randomly choose example. If that fails, re-choose e_1 .

Update α_1 and α_2 in one step

Compute new threshold b

Now for the implementation By SMO algorithm is done in the same manner as it is done by Naïve Bayes Classifier using the LightSide Machine Learning software [11] except that after feature attraction for building the model, SMO algorithm is chosen. Table 2, Represents the confusion matrix obtained by SMO algorithm and Fig 2. represents its confusion matrix.

Table no. 2-Confusion Matrix of SMO

Act/Pred	Bad	Good
Bad	99	26
Good	38	86

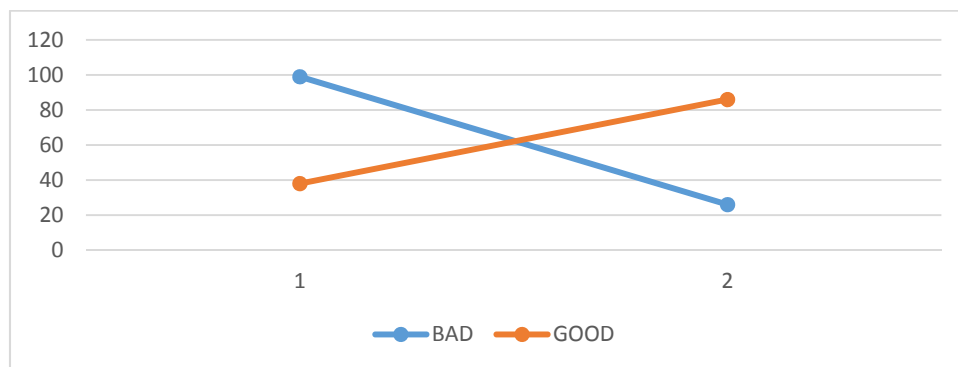


Fig.2: Confusion Matrix Representation of SMO

Now the implementation of phase 2 is carried on. In this the CAR Database is worked upon by two Clustering algorithms: K-Means Algorithm and SOM (Self Organising Maps), and the results obtained by two are contrasted among one another and analysed.

Now the implementation of K-Means Algorithm is done:

- 1) First the Cluster 3.0 software [10] is opened and thus the input car dataset is input in txt format. The results of the K-Mean algorithm obtained by Cluster 3.0[10] is then loaded in another open source tool Java Tree View[12] which displays the clusters properly by the means of dendrogram and the Scatter plot. Fig 3 shows the dendrogram which shows the clusters made in different colours.

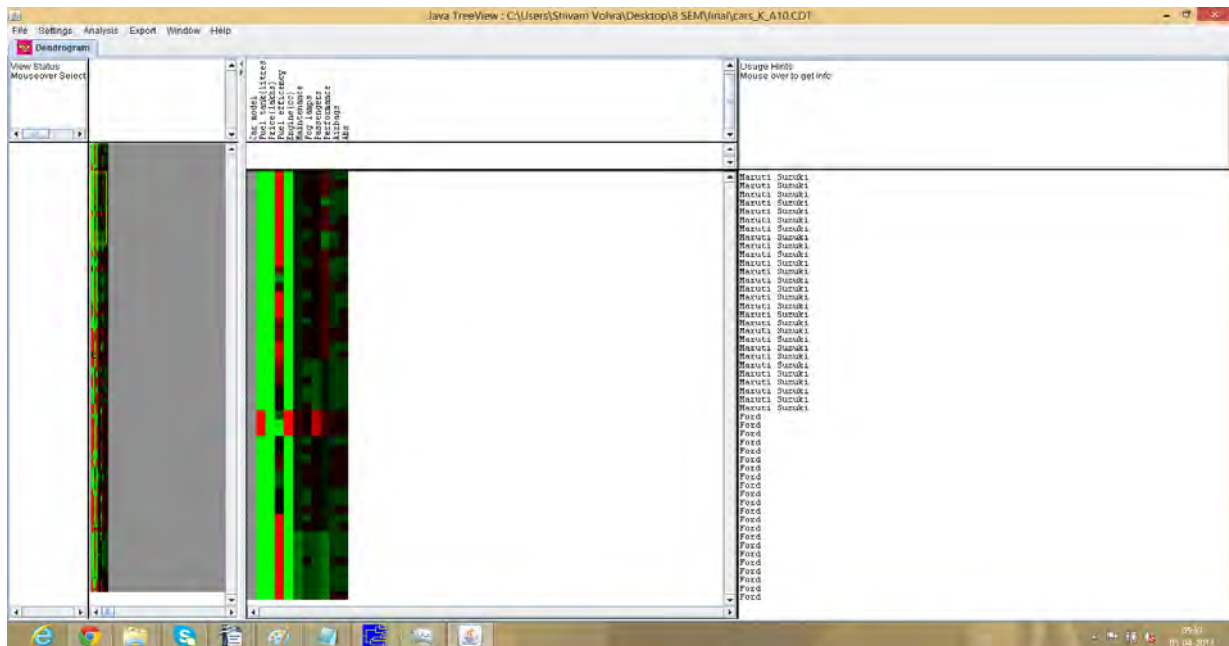


Fig. 3-Dendrogram of K-Means Clustering

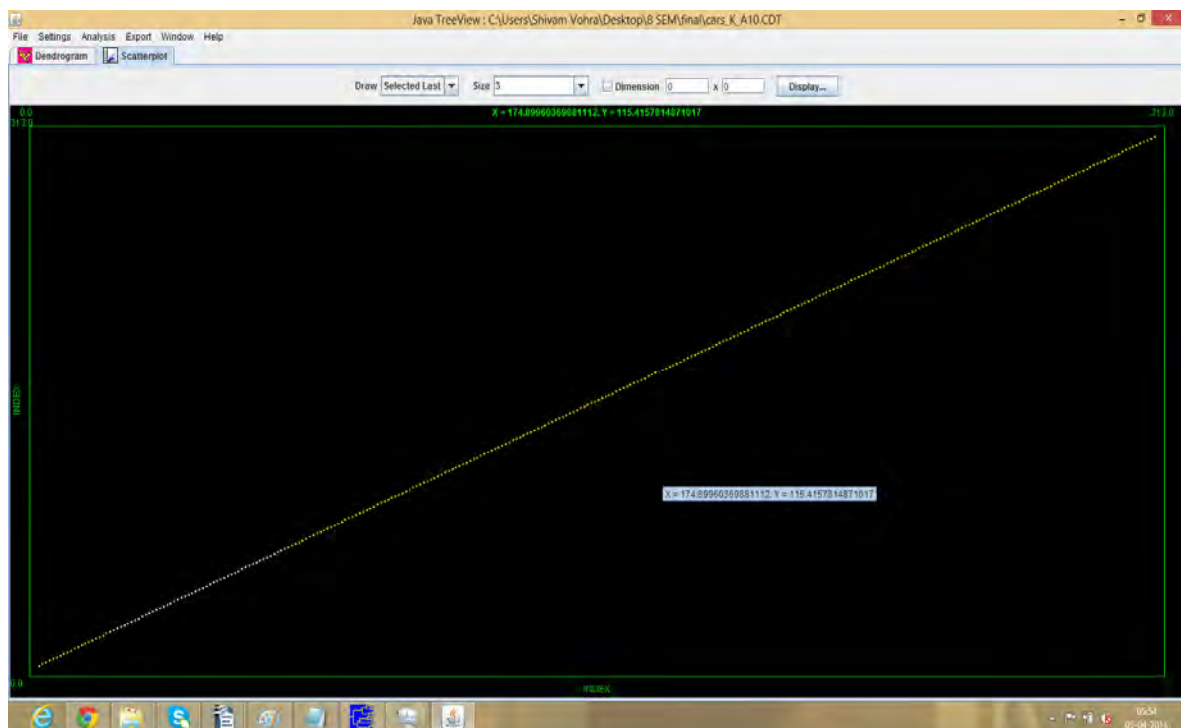


Fig 4-Scatter Plot of K-Means Clustering

2) Then the SOM algorithm is applied in the same tool and the dendrogram and the scatterplot is obtained from the JavaTreeView [12] for it too. Fig 5 shows the dendrogram which represents the clusters in different colours.

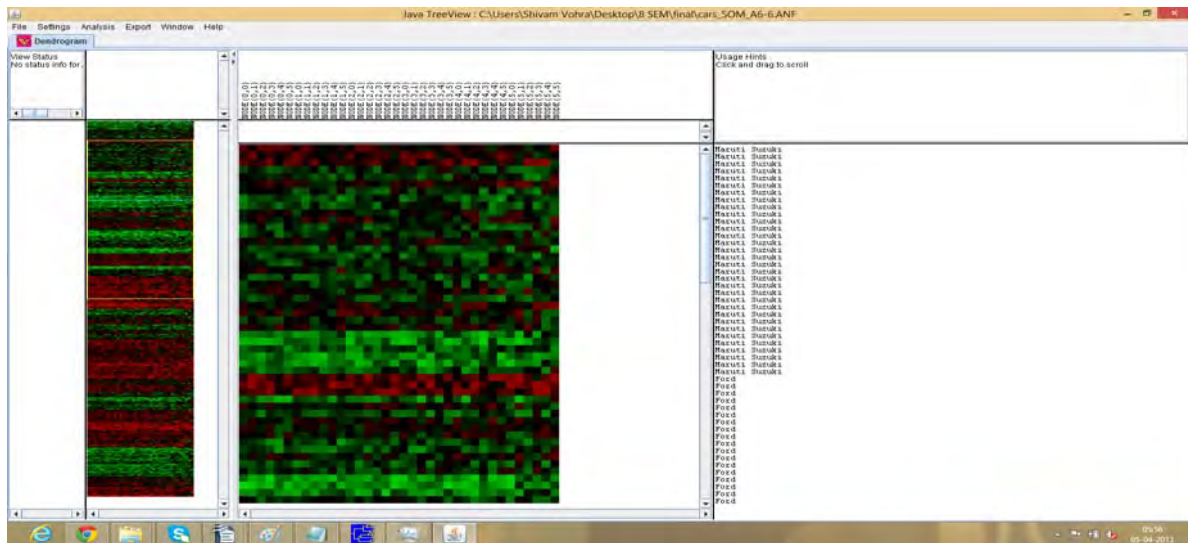


Fig 5-Dendrogram of SOM

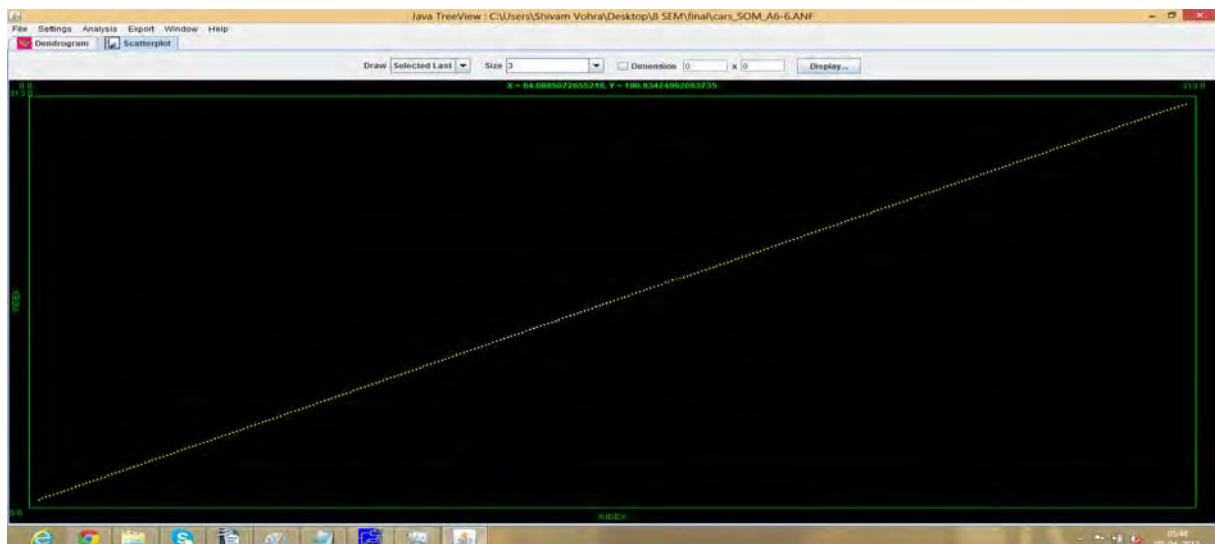


Fig 6 : Scatterplot of SOM

IV. RESULTS AND DISCUSSION

The results obtained by both the learning algorithms are observed. Table Numbers 1 and 2 clearly show that SMO algorithm generates comparatively better accuracy than the Naïve Bayes algorithm as the number of comments predicted correct by the model using SMO algorithm if compared to that of Naïve Bayes. Thus it is seen that for such a model, SMO algorithm is more accurate than Naïve Bayes.

The results obtained by clustering algorithms are observed as follows:

- 1) Clusters formed by K-Mean Algorithm is shown in Fig. 7 in graphical manner.

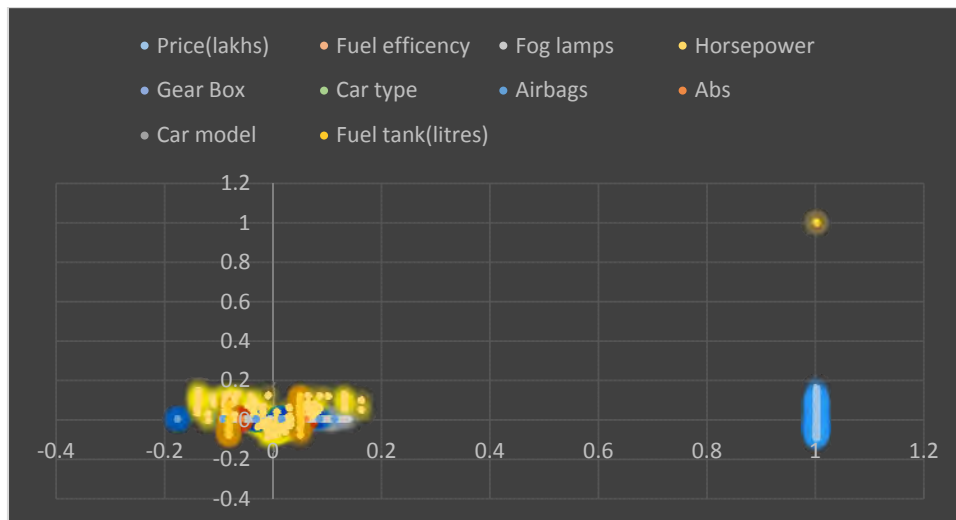


Fig. 7: Cluster distribution using K-Means Clustering

2) Clusters formed by SOM algorithm is shown in Fig 8 in graphical manner.

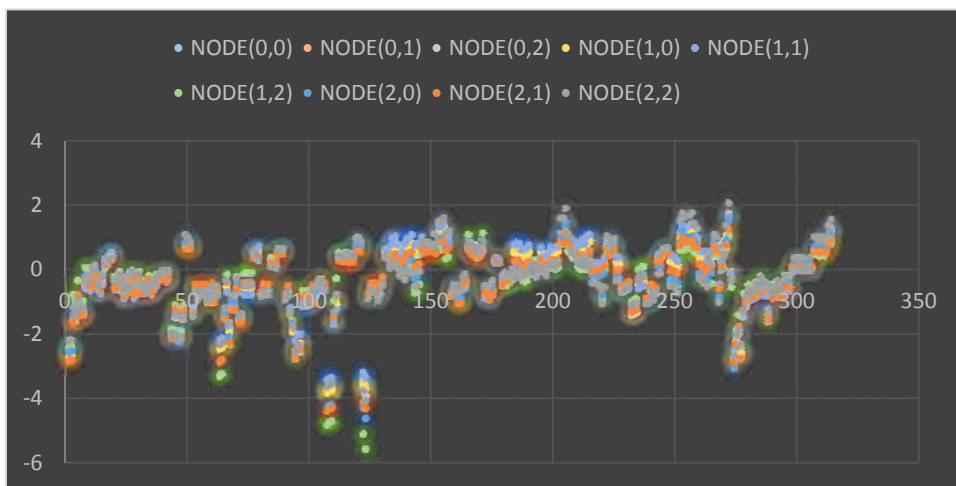


Fig. 8: Cluster distribution using SOM Clustering

Thus, analyzing Figures 3 and 5 it is seen that the clusters formed by K-Mean algorithm are more properly defined and accurate while those formed by SOM algorithm are not that properly defined and hence are thus stretched.

Fig 7 and 8 shows us that the result obtained to us by K-Mean algorithm ranges between (-1,1) while that of SOM has higher and lower values. Thus the result obtained by K-Mean algorithm is more reliable to the user.

V. CONCLUSION AND FUTURE WORK

Thus, we have seen that the model created by using the learning algorithm was successful in predicting the characteristic of each review comment and the accuracy can be improved by increasing the depth of the database which is a part of the future work. It is also seen that using SMO (Sequential Minimal Optimization) was produced more accurate results if compared to that of Naïve Bayes.

The implementation of clustering algorithm showed us that K-Means algorithm produced better and more proper clusters if compared to SOM clusters and so, using K-Means algorithm can be more beneficial in certain cases.

REFERENCES

- [1] Fernando Bação, Victor Lobo and Marco Painho(2005), "Self-organizing Maps as Substitutes for K-Means Clustering", IEEE research paper, pages 476-483
- [2] Juha Vesanto and Esa Alhoniemi(2000), "Clustering of the Self-Organizing Map", IEEE research paper, pages 586-600
- [3] Juntao Wang and Xiaolong(2011), "An improved K-Means clustering algorithm", IEEE research paper, page 45
- [4] Zhang, Harry. "The Optimality of Naive Bayes". FLAIRS 2004

- [5] Caruana, R.; Niculescu-Mizil, A. (2006). "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning.
- [6] S.L. Ting, W.H. Ip, Albert H.C. Tsang (2011) "Is Naïve Bayes a Good Classifier for Document Classification?" International Journal of Software Engineering and Its Applications Vol. 5, No. 3, page 37.
- [7] Rifkin, Ryan (2002), "Everything Old is New Again: a Fresh Look at Historical Approaches in Machine Learning"
- [8] E.W.T Ngaia, LiXiub, D.C.K Chaua(2008), "Application of data mining techniques in customer relationship management: A literature review and classification"
- [9] Pai-Hsuen Chen, Rong-En Fan, Chih-Jen Lin(2006), "A study on SMO-type decomposition methods for support vector machines" .
- [10] Cluster 3.0-a software used to implement K-Means Algorithm and SOM algorithm.
- [11] LightSide-a software used to implement the learning algorithm in order to generate the required comment predicting model.
- [12] Java Tree View-software used to display the results of the clustering algorithm.
- [13] Daniel Lowd and Pedro Domingos, "Naive Bayes Models for Probability Estimation" Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2350, USA
- [14] Algorithm by John Platt, Lecture by David PageCS 760: Machine Learning. "SVM by Sequential Minimal Optimisation (SMO)"
- [15] Jiawei Han and Micheline Kamber(2006), "Data Mining: Concepts and Techniques", 2nd ed.The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers,
- [16] K Murugesan, Md Rukunuddin Ghalib, J. Geetanjali, J. Indumathi(2009), "A Pioneering Cryptic Random Projection Based Approach for Privacy Preserving Data Mining", Proc. of IEEE IRI2009 Intl.conf. pp 437-439
- [17] I Sumaiya Thaseen, Fazal E Sattar, Muhammad Rukunuddin Ghalib(2011), "A Secure Adaptive Agent Model using Model Driven Architecture", Intl.J. of Computer Information Systems, Vol.2, Issue 2, pp 56-62.
- [18] Li Zhuang, Feng Jing(2006) , "Movie Review Mining and Summarization", proc. of ACM Intl Conf., pp 43-50, doi-10.1145/1183614.1183625
- [19] Olfa Nasraoui, Maha Soliman, and Antonio Badia(2008), "Mining and Tracking Evolving User Profiles and More – A Real Life Case Study", Proc. of IEEE Intl.J. of computer science, vol.20 Issue 2, pp 202-215
(doi.ieeecomputersociety.org/10.1109/TKDE.2007.190667)
- [20] Tipawan Silwattananusarn and Dr. Kulthida Tuamsuk(2012), "Data Mining and Its Applications for Knowledge Management", Intl.J. of IJDKP, vol.2, Issue 5.