VOL. 10, NO. 14, AUGUST 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A STATISTICAL COMPARISON OF LOGISTIC REGRESSION AND DIFFERENT BAYES CLASSIFICATION METHODS FOR MACHINE LEARNING

L. Mary Gladence[1], M. Karthi[2] and V. Maria Anu[1]
[1]Faculty of Computer Science, Department of Information Technology, JEPPIAR Nagar, Shollinganallur, Chennai, Tamil Nadu, India
[2]Karpagam College of Engineering, Coimbatore, Tamil Nadu, India
E-Mail: lgladence@gmail.com

ABSTRACT

Recent Machine Learning algorithms are widely available for various purposes. But which classifier is suitable for particular data is not yet defined. To consider this into account, well known classifiers Logistics Regression and Bayesian Classifier is taken to validate the work. To validate this, consider some factor such as Asymptotic error (i.e Normally Naive Bayes reaches its asymptotic error very quickly with regards to the number of training samples), how performance takes place when we increase the data set size etc.. Here we discuss how various bayes classifiers like Bayes Network, Naive Bayes, Naïve Bayes Multinomial Text, and Naïve Bayes Updateable are working and how they differ with each other based on given data and these results are effectively compared with Logistics Regression. Moreover, proposed work is compared using Naïve Bayes and Logistic Regression by using some standard dataset results as input. Finally it shows how the Bayes classifier methods and Logistic Regression differs each other in terms of performance factor.

Keyword: classification, logistic regression, Bayes classifier, projection.

1. INTRODUCTION

Generally the classification is done by manual, rule based, statistical or probabilistic based and some other techniques. Naive Bayes, Logistic Regression, SVM, Decision Trees are some of the methods based on statistical or probabilistic. Our proposed work tries to combine the background knowledge of some Bayes classifiers and Logistics Regression. Before going to see the detail we need to understand what is Generative models and discriminative models. Consider the classifier involves function of F(x): X→ Y or P (Y|X).The Discriminative classifiers are based on conditional models. So it considers the functional form for P (Y|X) and Estimate parameters of P (Y|X) directly from training data. But Generative classifiers are based on joint models. So it considers the functional form for P (X|Y), P(X) and Estimate parameters of P (X|Y), P(X) directly from training data. Using Bayes rule, we can calculate probability P (Y|X= x_j). Naive Bayes, Logistic Regression learning algorithms are based on understanding of probability by Interesting relationship between Generative and Discriminative classifiers.

2. RELATED WORK

This section discusses the research related to the classifiers. Generally logistic regression used for state-of-the-art categorization. The same approach is used in [1], [2] other high dimensional data analysis problems like predicting adverse drug events [3]. But the drawback of this work is recent applications more broaden in range. Some extension to polychromous logistic regression, and other related research have applied more complex models [4]. In richer prior distributions more informative in the statistical based result, but they are not based on any idea about language. In recent work [5], they propose some features are likely to be more useful than others, provide more significant high effectiveness. Here in [8] developed a classification tree based on standard long-term HRV for risk assessment in patients suffering from CHF. Totally 11 Attributes are used to find out particular person is having Heart Disease or not. Using Naïve Bayes Classifier Automatic Classifier is developed. The proposed classifier separates lower risk patients from higher risk ones, using standard long-term heart rate variability (HRV) measures. Moreover, proposed work is compared with CART (Classification and Regression Tree) the proposed method achieved the highest performance in terms of accuracy rate and sensitivity

3. PROBLEM STATEMENT

In KDD processing one of most difficult part is choosing the correct suitable data mining technique for specific data to perform decision making. Even though there are many techniques in the Data mining process, choosing the best classifier becomes challenging work. This makes troubles to the researchers who often use more data for their research work. While coming to classifications, choosing the methods as well as the relationship between them become more challenging. Here, Bayesian Classification Methods, Logistic regression is compared to find out which classifier suits for which data to perform better classification. This work suggests user to choose a classification method based on their data set. Also, it suggests the user to choose most

ARPN Journal of Engineering and Applied Sciences

suitable data mining techniques by understanding the classifier based on a given problem. Here we chose a comparison of some classifier like Bayes Network, Naive Bayes, Naïve Bayes Multinomial Text, Naïve Bayes Updateable, Logistics Regression using standard datasets like Automobile, Contact-lenses, Thyroid, Housing etc.. And these data sets were extracted from the UCI repository. It is not possible to conclude that one particular method is best, since it depends on what type data we are going to handle and what type of result we expect like discrete, continuous and etc., So our novel work try to combine the Logistic Regression with Naive Bayes by comparing Naive Bayes and Logistic Regression. Also it discussed detail about the various Bayes method already proposed. If someone have to process there data, this work shows clear ideas about choosing various classification method especially Naive Bayes and Logistic Regression as their initial stage of implementation. The"Fig.1" shows the sample parallel project of data's in the dataset. Also, we compared the results based on precision, recall, MAE, RMSE, Kappa performance metric values.

4. CLASSIFICATION TECHNIQUES

A. Bayes network

A Bayesian network is a simple graphical notation or structured form for specifying probabilistic relationships among a group of variables X = {X_1, X_n,}, which contains a network structure of conditional independence about X and a set of local probability distributions. It represents dependency or independency of data via a directed graph. In a Bayes network the nodes are denoted by random variables and edges represent direct dependence. Bayesian Probability is the degree of belief in that particular event, but the classical Probability is the physical probability of a particular event. All the nodes in the graph are 1 to 1 relationship with the variables X and arc denotes a conditional independence which shows the strength of dependencies. A conditional distribution for nodes in the graph given its parents as P (X_j | Parents (X_j)). To construct bayes network, consider an ordering of variables $X_1, \dots, X_n$ .For each value of j (1 to n), X_j is added to the network and select parents (P) from $X_1, \dots, X_{j-1}$. Therefore,

$$P(X_j \mid Parents(X_j)) = P(X_j \mid X_1, \dots X_{j-1}) \quad (1)$$

This choice of parent's guarantees is form by chain rule

$$P(X_1, \dots, X_n) = \sum_{j=1} P(X_j \mid Parents(X_j)) \quad (2)$$

In general this is represented by,

$$p(X_1, \dots X_n) = P(\, p(X_j \mid parents(X_j))\,) \quad (3)$$

The main components of a Bayesian network are the graph structure which represents the conditional independence assumptions and the numerical probabilities for each variable. The Bayesian network graphs are no directed cycles (acyclic). But there is some limitation of this method. Bayesian Networks require some initial knowledge of more probabilities like quality and significant computational cost. Also, it provides a natural representation for the conditional independence, at the same time it is easy for domain experts to construct.

B. Multinomial Naive Bayes

Multinomial Naive Bayes is a new version of Naive Bayes which is designed for the text documents. Naive Bayes models the document based on presences and absences of particular words, whereas multinomial naive Bayes construct models explicitly by word counts. Denoting the classes as c_i and any relevant features F vector, the probability of a given class is given by Bayes Theorem,

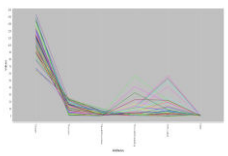


Figure-1. Parallel coordination plot of datapoint during classification.

$$p(C_i|\vec{B}) = \frac{P(\vec{B}|C_i)P(C_i)}{\sum_j p(\vec{B}|C_j)P(C_j)} \quad (4)$$

Now all we need to do is model the class likelihoods, $P(\vec{B}|C_i)$. The Naive Bayes assumption is that the features are independent given a class, i.e.

$$P(\vec{B}|C_i) = \prod_j p(B_j|C_i) \quad (5)$$

Popular choices of the $p(B_j|C_i)$ include the Bernoulli distribution (taking into account whether a binary feature occurs or not) and the Binomial distribution. Even it is similar to Bayes formula, the idea behind it, is text classification. So our work compares how results are deviated from other classifier.

C. Naive Bayes

VOL. 10, NO. 14, AUGUST 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

Naive Bayes Classifier is one of the practical learning Bayesian classifier method for data with several attributes (i.e. Vector data) and used to represents that attributes as independence of that class. It assigns a posterior probability to a class based on its prior probability and its likelihood for the given training data. It computes the maximum a posterior (MAP) hypothesis or the maximum likelihood (ML) hypothesis. Naive Bayes classifier assumes conditional independence between attributes and assigns the MAP class to new instances. Naïve Bayes is based on the independence assumption where training is very easy and fast. It requires considering each attribute in each class separately. The test is straightforward and looking up tables or calculating conditional probabilities with normal distributions. Naïve Bayes performance is competitive to most of state-of-the-art classifiers even in presence of violating the independence assumption.

Consider if data have many attributes, then attributes that describe data instances are conditionally independent for the given classification hypothesis. The Bayesian classifier that uses the Naive Bayes assumption and computes the MAP hypothesis is called Naïve Bayes classifier. To, Classify any new data instance x=(a_1,…a_t) as:

$$H_{NB} = {}^{Max}_{\ h}\, p(h)p(X|h) \tag{6}$$

To do this on training examples, the parameters need to estimate from the training examples for each target value of hypothesis (h)

$$\check{p}\,(h) = est\ p(h) \tag{7}$$

Where est p(h) denotes the estimated value of p(h)
Considering every attribute value k_t of each data instance

$$\check{p}(k_t|h) = est\ p(k_t|h) \tag{8}$$

Generative model for Naive Bayes is given below

$$\check{p}(y = q) \coloneqq \frac{r\{y=q\}}{\sum_j r\{y=j\}} \tag{9}$$

By comparing other, the Naive Bayes somewhat closely related to other methods compared especially, Logistic regression. Both are simple by considering implementing and performance on problems. Normally Naive Bayes reaches its asymptotic error very quickly with regards to the number of training examples.

D. Naïve Bayes updatable

Updatable classifier is one of the incremental learning which is used to train the learning model for each value within the dataset especially with large datasets. To minimize the objective function, it uses stochastic gradient descent, which is one of gradient descent optimization methods and it is written as a sum of differentiable function. This method is applicable to large datasets, because it evaluates instance of the training dataset one by one. The Naïve Bayes Updatable is the updatable version of Naive Bayes. While building Classifier a default precision used by this classifier is 0.1 for numeric attributes hence it is expressed as an incremental update.

E. Logistic regression

Logistic regression is used because of relationship between the discrete variable. It is known that regression give better result on numerical data values, but it allows the prediction of discrete variables by the mixed values of continuous and discrete predictors. The discriminate function analysis and multiple regressions have same functionality but with there is no distributional assumptions on the resultant predictors .The predictors are linearly related instead of normally distributed also it has equal variance in all group. Example consider the Thyroid dataset in table 1 ,the probability of disease changes very small with a 10 scale difference among people with low heart rate, at the same time a 10 scale make change in the probability of Thyroid disease in people with high heart rate. Here our novel work try to combine the Logistic Regression with Naive Bayes by comparing the Naive Bayes and Logistic Regression is given below,

$$\check{p}(y = q)|x, \alpha, \beta) = 1/(1 + e^{(-\sum_{i=1}^{n} \alpha_i x_i - \beta)}) \tag{10}$$

Where $\check{p}(y = q)|x, \alpha, \beta)$ is used to show relationship between the Naive Bayes and Logistic regression in derivation. Here α, β are discrete outcome values of Logistic Regression. Based on equation both classifiers are linear. If someone met the distributional assumptions from their data then discriminate function analysis gives better result. If processed data outcome is continuous then multiple regressions become better based on given assumptions. For example consider the Automobile dataset, the prediction of group of weights of tool which increase the likelihood conditional on training data by higher weights of tools. To make a view and comparison purpose the projection of data in the dataset discussed here is represented in three dimension views between the logistic regression and other Bayes methods which are shown in “Figure-2”. It denoted that particular feature is fit with best class on account of classifier. So it concludes that Naive Bayes is a special case of logistic regression that uses Bayes rule and conditional probabilities to set these weights.

4. EXPERIMENTS AND RESULTS

VOL. 10, NO. 14, AUGUST 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

To validate the Logistic regression with another classifier used some data set such as Automobile, Contact-lenses, Thyroid, and Housing etc. And these data sets were extracted from the UCI repository. Table-1 represents the characteristics of the dataset based on a number of patterns (N), number of attributes (K), number of classes (C) are taken.

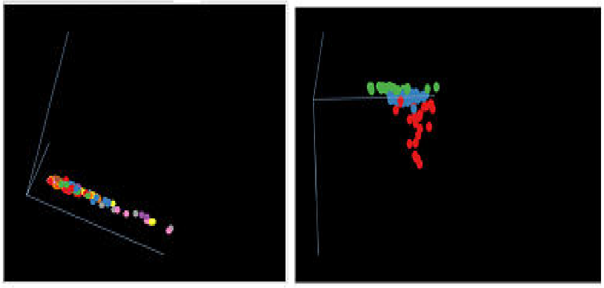
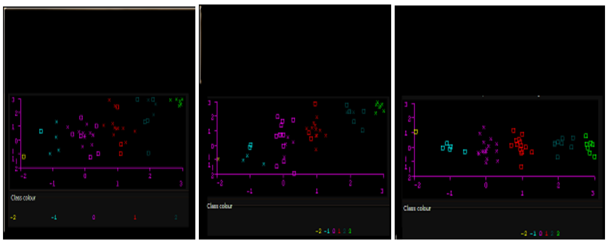
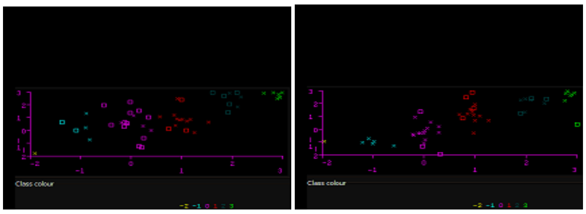


Figure-2. Different between projections of data in varies Bayes and Logistic Regression classifier.



(a) (b) (c)



(d) (e)

Figure-3. Margin curve representation of dataset after classification (a) Bayes network (b) Naive Bayes (c) Logistic regression (d) Naïve Bayes updateable (e) Naïve Bayes multinomial text.

A. Performance measures

In this work Precision, Recall, Mean Absolute Error, Average Mean Absolute Error, Kappa's used to validate the classifier methods. Precision and recall are based on True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). The Mean Absolute error is close to the Mean Squared Error. The only difference is, it uses absolute values instead of squaring. Here average of these absolute value is taken to consider the mean absolute and is measured using an equation

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} e(x_i) \quad (11)$$

Root Mean Square Error is measured using equation

$$\mathrm{RMSE} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{N_n}\sum_{i=1}^{N_n} e(x_i) \quad (12)$$

Table-2 shows about the comparison between different methods like Bayes Network, Naive Bayes, Naïve Bayes Multinomial Text, Naïve Bayes Updateable, Logistic Regression for different datasets based on Precision. Here bold face data represent maximum precision achieved for each data set and italic font represents the second highest precision achieved for each data set. For example, in Table-3 Logistic regression has achieved maximum Recall in all data sets and second highest Recall in Bayes Network method which is denoted by italic font style.

Table-1. Characteristics of datasets based on their classes.

Dataset	N	K	C
Automobile	52	26	6
Bond rate	5	42	15
Contact-lenses	18	6	3
Thyroid	161	6	3
Housing	3	39	25

After classification how the class labels were projected into particular margin for each classifier is mentioned in "Figure-3"

It is the representation of Bayes Network, Logistic Regression, Naïve Bayes Updateable, Naive Bayes updatable, Naïve Bayes Multinomial Text for Automobile data set.

Table-2. Results of the precision for various methods compared.

VOL. 10, NO. 14, AUGUST 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

Method/Dataset	**Automobile**	**Contact lenses**	**Bond rate**	**Thyroid**	**Housing**
LR	**0.808**	**0.917**	0.489	**1.000**	0.333
NB	0.582	**0.917**	0.610	0.582	**0.582**
NBMT	0.107	0.444	0.410	0.495	0.360
NBU	0.582	0.429	0.426	0.429	0.429
BN	*0.692*	*0.915*	**0.890**	*0.932*	*0.570*

Table-3. Results of the recall for various methods compared.

Method/ Dataset	**Automobile**	**Contact lenses**	**Bond rate**	**Thyroid**	**Housing**
LR	**0.788**	**0.833**	**0.756**	**1.000**	**0.667**
NB	0.577	**0.833**	0.421	0.577	0.577
NBMT	0.327	0.667	0.502	0.704	*0.600*
NBU	0.577	0.333	0.356	0.333	0.333
BN	*0.678*	*0.823*	*0.584*	*0.926*	0.333

Table-4. Results of The MAE for various methods compared.

Method/ Dataset	**Automobile**	**Contact lenses**	**Bond rate**	**Thyroid**	**Housing**
LR	**0.0685**	**0.1111**	**0.017**	**0.0001**	0.2667
NB	0.1371	0.2584	***0.1953***	0.2008	**0.2008**
NBMT	0.2575	0.3651	0.3112	0.3125	0.2485
NBU	0.1371	0.2404	0.214	0.2404	*0.2404*
BN	***0.1208***	***0.2186***	0.2843	***0.0444***	**0.2008**

Table-5. Results of The RMSE for various methods compared.

Method/ Dataset	**Automobile**	**Contact lenses**	**Bond rate**	**Thyroid**	**Housing**
LR	**0.2555**	0.3333	**0.1231**	**0.0006**	0.5079
NB	0.3316	*0.319*	*0.307*	*0.0106*	**0**
NBMT	0.3582	0.4142	0.402	0.3919	0.3461
NBU	0.3316	0.4029	0.3942	0.4029	0.4029
BN	*0.2957*	**0.2979**	0.312	0.1547	*0.321*

Table-6. Results of the KAPPA for various methods compared.

Method / Data set	Auto mobile	Contact lenses	Bond rate	Thyroid	Housing
LR	**-0.087**	0.7143	0.3267	1.000	**-0.087**
NB	0.3534	0.7143	0.4517	0.3534	0.3534
NBMT	0.001	*0.023*	*0.001*	*0.0071*	0.006
NBU	*-0.0135*	**-0.0135**	**-0.0134**	**-0.0135**	*-0.0135*
BN	0.3534	0.7143	0.4573	0.8457	0.3534

Similarly, Table-3, Table-4 and Table-5 are derived for different datasets based on Mean absolute Error (MAE), RMSE and Kappa performance metric. From these we can identify LR gave best result while compared with other methods. In addition to this, how a value varies between various measures are also compared. It shows that Logistic Regression clearly project the data in margin comparing to others. However, Ordinal dataset gave better results in LR but in some other cases it gave second best values.

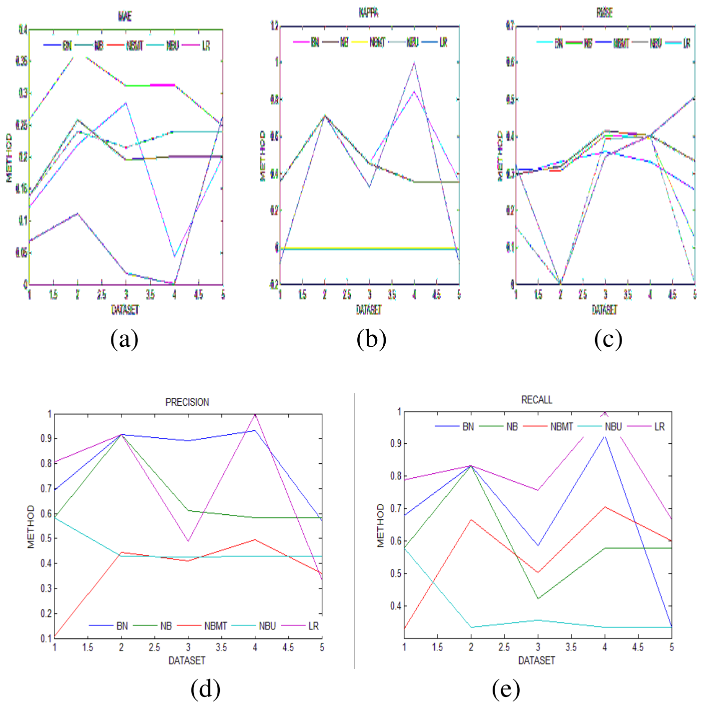


Figure-4. Graphical representation of result obtained by different performance measures (a) MAE (b) Kappa (c) RMSE (d) Precision (e) Recall for classifier used.

"Figure-4" shows how different dataset results are compared with various methods based on precision, recall, MAE, RMSE, Kappa performance metric respectively. To differentiate each method graphical representation is plotted with different colors according to datasets.

5. CONCLUSIONS

This paper shows how Bayesian classification methods can be applied to regression problems. Proposed work combined the knowledge of classification using Bayesian method and logistic regression. A result shows that Logistic Regression outperforms Bayesian Classification Methods in terms of various performance measures such as Precision, Recall, Mean Absolute error, Kappa, RMSE. These measures are checked using the data set such as Automobile, Contact-lenses, Bond rate, Thyroid, Housing and this data set were extracted from UCI Repository. Here, Bayesian has better result with respect to Average Mean absolute Error but has the worst performance with respect to Average Mean Squared Error. Bayesian Classification methods play better role, if there is a situation to use less number of training samples.

ACKNOWLEDGEMENT

Thanks to Sathyabama University for their encouragement to finish this work successfully. We would like to express our gratitude to those who helped during this research.

REFERENCES

[1] Cambon AC, Baumgartner KB, Brock GN, Cooper NGF, Wu D., and Rai SN. 2014. Classification of Clinical Outcomes Using High-Throughput Informatics: Part 1- Nonparametric Method Reviews. Model Assisted Statistics and Applications, In Press.

[2] Alexander Genkin, David D. Lewis, David D. Lewis. 2007. Large-Scale Bayesian Logistic Regression for Text Categorization. Technometrics. 49(3).

[3] Hauben M., Madigan D., Gerrits C. and Meyboom R. 2005. The Role of Data Mining in Pharmacovigilence. Expert Opinion in Drug Safety. 4: 929-948.

[4] Li F. and Yang Y. 2004. Recovering Genetic Regulatory Networks from Micro-Array Data and Location Analysis Data. Genome Informatics. 15: 131-140.

[5] Dayanik A., Lewis D., Madigan D., Menkov V. and Genkin A. 2006. Constructing Informative Prior Distributions from Domain Knowledge in Text Classification. In: Proceedings of SIGIR 2006: The Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM. pp. 493-500.

[6] Eibe Frank, Leonard Trigg, Geoffrey Holmes, Ian H. Witten. 2000. Technical Note: Naive Bayes for Regression. Machine Learning, 41, 5-25, Kluwer Academic Publishers.

[7] Madigan D., Genkin A., Lewis D. D. and Fradkin D. 2005b. Bayesian Multinomial Logistic Regression for Author Identification. In Bayesian Analysis and Maximum Entropy Methods in Science and Engineering: 25th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, AIP Conference Proceedings. Vol. 803, Melville, NY: AIP, pp. 509-516.

[8] L.Mary Gladence, K.Ravi, M.Karthi. 2014. Heart Disease Prediction using Naive Bayes Classifier – Sequential Pattern Mining. International Journal of Applied Engineering Research (IJAER).

[9] L. Mary Gladence, K. Ravi, M. Karthi. 2014. An Enhanced Method for Detecting Congestive Heart Failure -Automatic Classifier. IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT).

[10] John Halloran. 2009. Classification: Naive Bayes vs Logistic Regression.

[11] Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani. 2009. Evaluation Measures for Ordinal Regression. Ninth International Conference on Intelligent Systems Design and Applications.

[12] A. Y. Ng and M. I. Jordan. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Neural Information Processing Systems.

[13] Revathy S., Parvathavarthini B. 2011. Integrating rough clustering with Fuzzy sets. Sustainable Energy and Intelligent Systems (SEISCON 2011), International Conference on. pp. 865-869.

[14] Andrew Ng. Cs229 lecture notes part III: Generalized linear models.

[15] Smith R. L. 1999. Bayesian and Frequentist Approaches to Parametric Predictive Inference (with discussion), in Bayesian Statistics 6, J. M.Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), Oxford, U.K.: Oxford University Press. pp. 589-612.

[16] Y. BevishJinila, K. Komathy, "Cluster oriented ID based multi-signature scheme for traffic congestion warning in Vehicular Ad hoc Networks ", Advances in Intelligent Systems and Computing, Vol.338, 2015, pp. 337 - 345.