RESEARCH ARTICLE OPEN ACCESS

Analysis Of Machine Learning Techniques By Using Blogger Data

Gowsalya.R, S.Veni and M.Hemalatha
Department of Computer Science, Karpagam University Coimbatore,TamilNadu.India.

ABSTRACT
Blogs are the recent fast progressing media which depends on information system and technological advancement. The mass media is not much developed for the developing countries are in government terms and their schemes are developed based on governmental concepts, so blogs are provided for knowledge and ideas sharing. This article has highlighted and performed simulations from obtained information, 100 instances of Bloggers by using Weka 3. 6 Tool, and by applying many machine learning algorithms and analyzed with the values of accuracy, precision, recall and F-measure for getting future tendency anticipation of users to blogging and using in strategical areas.

Keywords **-** Blog, Cyber Space, Data Mining, Random Forest.

I. INTRODUCTION

Data Mining has a great potential for investigating the concealed patterns from the large datasets of the web blogger. These patterns might be used for fetching the information from a new and/or future data. Nonetheless, the accessible raw blogger information is generally distributed by collecting the tremendous amounts of information. Initially, with the computers and means for huge storage, all sorts of data are started to collect and stored [2].These information requirement to be gathered in a composed structure. This gathered information could be coordinated to form a database. Data mining is an interdisciplinary subfield of Computer science, is the computational methodology of discovering patterns in huge data including methods at the convergence of artificial intelligence, machine learning, methods and database frameworks. The general objective of the data mining procedure is to extract information from a dataset and convert it into a understandable structure for further utilization. The structural planning exhibits the general procedure of Knowledge Discovery in Databases (KDD [21]). This initial phase has led to the creation of structured databases and Database Management Systems (DBMS). The DBMS have been very important and vital assets for managing a huge amount of data for effective and efficient retrieval of information from a large collection Information retrieval is not enough for decision-making. Automatic summarization of data is done and the extracted information are stored, as fresh data in discovery of patterns. Blog as a recent online networking in the internet is one of the Internet and web services, which normally give free space to clients to give them a chance to participate as a part of system and virtual community. It provides unlimited dynamic, intuitive relations, opinions and news about particular issues equipped for updating the others opinions about given issues or topics [2, 3]. A blog is an on line web journal that holds content, photos, representation, connections to different web journals, sites and related media. The term comes from the combination of the two words, "web" and "Log". The individual who makes and upgrades a site is known as a blogger. Anyone can be turned into a blogger;- all a person need is, a machine with an Internet connection, a free blogger account from a blog provider site, and the desire to share your ideas to the world. A blog (short for web log) is an individual online journal that is oftentimes upgraded and proposed for general public viewing. Regularly, a web journal holds a blending of contents, design and connections to important sites and different web journals of similar interest. The talents for the readers to leave suggestions on unique posts is generally a critical characteristic of web blogs. A blog may contains widgets and plug-in to upgrade its practicality. A widget is a small web requisition, holding element (evolving) substance, which could be added to any site.

II. RELATED WORKS

In this regard, it can be noted to the paper in the past such as Zafarani et al. [1] are used the Blogger system and selected proper data and then they begun to process it. Creating lists and extracting the keywords. And by measuring the importance and frequency of the collected data, and automatically pointed out the social and political issues.

Nachev and Ganchev [7] are proposed a new approach based on Art2 artificial neural network (ANN) as a kind of data analysis which contributes to the survey in data blog and uses it to provide

customized data. They filter the data to identify the users and then obtain result vectors and clustering results by using ANN. By analyzing data publish in blogging space, Kwonm et al [8] find out a different theory in contrast with social networks theory which relies on data publishing without any relationship. By clustering, they began to find increasing data explosion and correlation between tendency potentials and data explosion [8].

Juffinger and Lex [4] have provided a system for blog analysis in all languages and by suggesting cross language data survey and imagery tendencies. They believe that the imagery tendencies would be based on recognition by providing pre-defined clustering and classification of blogs. Iraklis Varlamis et al. [9] have considered feature vector after classifying results. By the means of analysis techniques, classification and related vector graphics, integration detection and blog categories distribution along with different time intervals for reaching blogger approach reasons seem to be possible

Demartini et al. [5] have used analysis techniques of time intervals in groups and integration method in blogs data to improve ideas on politicians which, at the same time, the estimation of available political trend in blog societies is possible. Wyld [2] has been considered blogging as a social phenomenon; it acts as an unique opportunity to improve interactions and management in digital area.

III. CLASSIFICATION TECHNIQUES

Classification is a data mining (machine learning) technique used for predicting group membership for data instances. Classification techniques include decision trees and neural networks.

3.1 Random Tree

A random tree is a tree constructed randomly from a set of possible trees having that is K random features at every node. "At random" means that among the set of trees each tree has fairly an equal chance of being sampled. Or it can be said that the trees have "uniform" distribution. Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models. There has been extensive research in the recent years over Random trees in the field of machine Learning.

3.2 Random Forest

A random forest consisting of a collection of tree structured classifiers (h(x, k), k = 1. . .) where the _k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [33].

The algorithm

- Choose 'T' number of trees to grow,
- Choose 's' number of variables used to split each node. s<<S, where S is the number of input variables, 's' holds constant while growing the forest.
- Grow 'T' trees. While growing each tree construct a bootstrap sample of size 'n' sampled from 'Sn' with the replacement and grow a tree from this bootstrap sample.
- When growing a tree at each node select m variables at random and use them to find the best split.
- Grow the tree to a maximal extent and there is absolutely no pruning.

3.3 PART (Partial Decision Trees) Classifier

PART is a rule based algorithm that produces a set of if-then rules that can be used to classify the given data. PART is a modification of C4.5 and RIPPER algorithms and draws strategies from both the algorithms. PART takes the divide-and-conquer strategy of RIPPER [15] and combines it with the random tree approach of C4.5. PART generates a set of rules according to the divide-and conquers strategy, removes all instances from the training collection that are covered by this rule and proceeds recursively until no instance remains.

3.4 Naive Bayesian Classifier

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable y and a dependent feature vector x_1 through x_n, Bayes' theorem states the following relationship:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Advantages

- Fast to train (single scan). Easyt to classify
- Not sensitive to irrelevant features
- Handles real and discrete data
- Handles streaming data well

1.5 NB Tree

The NB Tree algorithm is a hybrid between decision-tree classifiers and Naive Bayes classifiers. It represents the learned knowledge in the form of a tree which is constructed recursively. However, the leaf nodes are Naive Bayes categorizers rather than nodes predicting a single class [6]. For continuous attributes, a threshold is chosen so as to limit the entropy measure. The utility of anode is evaluated by discretization the data and computing the fivefold cross-validation accuracy estimation using Naive Bayes at the node. The utility of the split is the weighted sum of utility of the nodes and this depends on the number of instances that go through that node. The NB Tree algorithm tries to approximate whether the generalization accuracy of Naive Bayes at each leaf is higher than a single Naive Bayes classifier at the present node[12, 13].

IV. WEKA TOOL

Weka (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

The Weka Knowledge Explorer is user friendly graphical user interface that harnesses the power of the Weka software [10]. The major Weka packages are Filters, Classifiers, Clusters, Associations, and Attribute Selection is represented in the Explorer along with a visualization tool, which allows datasets and the predictions of Classifiers and Clusters to be visualized in two dimensions.

The workbench contains a collection of visualization tools and algorithms for data analysis and predictive modelling together with graphical user interfaces for easy access to this functionality. It was primarily designed as a tool for analyzing data from agricultural domains. Now it is used in various other application areas especially for educational purposes and research.

4.1 WEKA DATA FORMAT

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software. This document describes the version of ARFF used with Weka versions 3.2 to 3.3; this is an extension of the ARFF format as described in the data mining book written by Ian H. Witten and Eibe Frank (the new additions are string attributes, date, attributes, and sparse instances[18, 19].

V. DATASET COLLECTION

The data were collected from **UCI Machine Learning Repository** at the Kohkiloye and Boyer Ahmad Province in Iran. **Web blogger** data set is developed by collecting information to form database is done by questionnaire of **100 instances** with **7 different attributes**. This questionnaire is provided as oral, written and also programming of a website whichincludes an internet questionnaire and the users can answer the questions as they wish. They entered their used websites, blogs and social networks during the day. After collecting questionnaires, the wed addresses are gathered to get expected results [11,12]. And finally, their trustfulness is checked by analyzing their used web pages. As the results were the same, for getting better and noiseless response, they should be put in a database.

Table 5.1 Attribute details in dataset.

Attributes Name	Type	Attribute description
DEGREE	High, Medium, Low	Degree denotes the level of bloggers in web blogger.
CAPRICE	Left, Right, Middle	Sudden changes in the bloggers.
TOPIC	Impression, Political, Tourism, News	Many topics are discussed in web blogger. Based on users query.
LMT(local media turnover)	True/False	Performance status – local media turnover
LPSS(local, political and social space)	True, False	Details about the user's questionnaire according to their specified field.
PB(Pro bloggers)	True, False	Professional bloggers are effective in digital media and interested in digital writing continuous time intervals.

VI. EXPERIMENTAL RESULTS

6.1 Data Preparation

The variables are already categorized and represented by text values. The manner in which the collision occurred is categorized under two heads.

6.1.1 Topic Based

Topic is the specific combination of various titles discussed in the web blogger and in this dataset political, impression, tourism, and news are the types of topics discussed.

6.1.2 Degree Based

Degree is the important attribute in blogger dataset. It consists of three main categories such as high, medium, and low. Based on the qualification the degree attribute is performed.

Table 6.2 Algorithms and Values

Algorithm	Correctly classified instance	Testing percentage 70%	Testing percentage 80%	Testing percentage 90%
Random Tree	92%	80%	77.50%	80%
Random Forest	91%	78%	68.75%	76%
NB Tree	90%	76.66%	65%	74%
PART	87%	73.33%	53.75%	60%
Naïve Bayes	75%	70%	65%	70%

Figure 6.1.2 Graphical representations for Accuracy and Testing

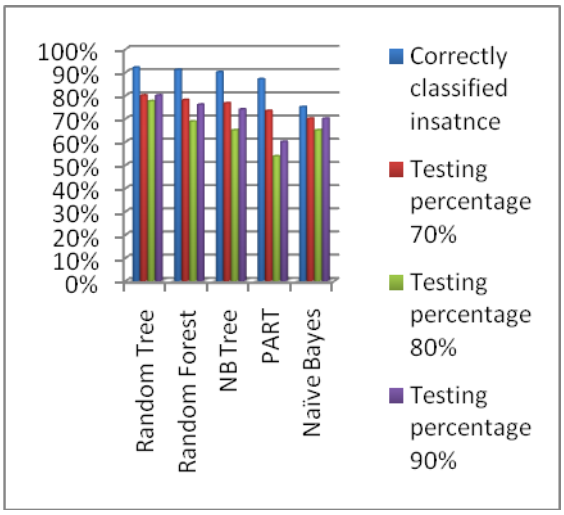


Table 6.3 Accuracy, Precision, Recall, FMEASURE, FOR EXISTING ALGORITHM.

ALGORITHMS	Random Tree	Random Forest	NB Tree	PART	Naïve Bayes
Accuracy	92%	91%	90%	87%	75%
Precision	0.929	0.905	0.893	0.848	0.759
Recall	0.985	0.956	0.945	0.945	0.926
FMeasure	0.944	0.942	0.937	0.912	0.834

Figure 6.3.1Analysis of Accuracy

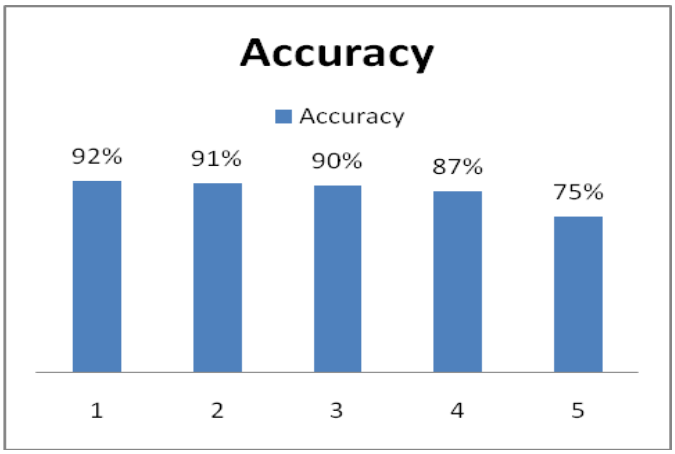


Figure 6.3.2 Analysis of Precision

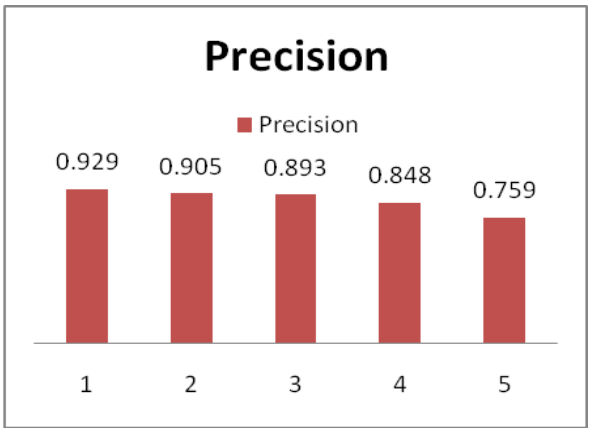


Figure 6.3.3 Analysis of Recall

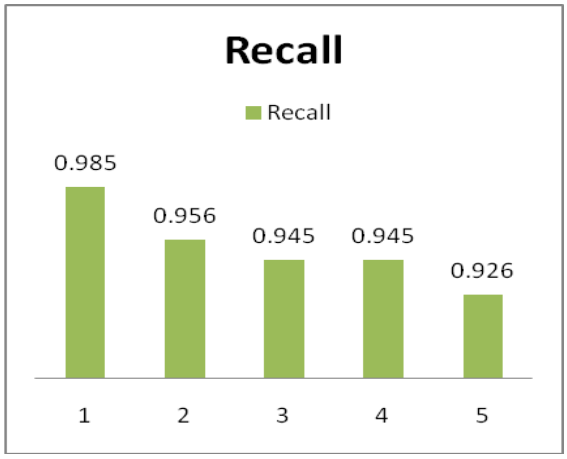
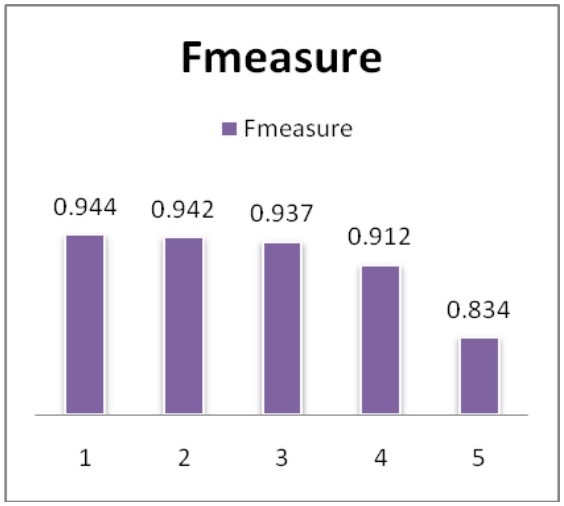


Figure 6.3.4 Analysis of Fmeasure



VII. K-Fold Cross-validation

In *k*-fold cross-validation, the original sample is randomly partitioned into *k* equal size subsamples. Of the *k* subsamples, a single subsample is retained as the validation data for testing the model, and the remaining *k* − 1 subsamples are used as training data. The cross-validation process is then repeated *k* times (the *folds*), with each of the *k* subsamples used exactly once as the validation data. The *k* results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling (see below) is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used,[6] but in general *k* remains an unfixed parameter.

$$CV = \frac{1}{k}\sum_{i=1}^{k} PM_i$$

Where CV stands for the cross-validation, k is the number of folds used, and PM is the performance measure for each fold (Olson & Delen, 2008). In this study, a stratified 10-fold cross-validation approach was used to estimate the performance of classifiers. Empirical studies have shown that 10 is the optimal number of folds that optimizes the time it takes to complete the test while minimizing the bias and variance associated with the validation process (Kohavi, 1995). In 10-fold cross-validation, the entire dataset is divided into 10 mutually exclusive subsets (or folds) with approximately the same class distribution as the original dataset (stratified). Each fold is used once to test the performance of the classifier that is generated from the combined data of the remaining nine fold, leading to 10 independent performance estimates.

VIII. CONCLUSION

This paper applies several probabilistic topic models to discourse blogs. Thereby introducing a novel comment prediction task to assess these models in an objective evaluation with possible practical applications. The results show that predicting political discourse behaviour is challenging, in part because of considerable variation in user behaviour across different blog sites. The results show that using topic modelling, one can start to make reasonable predictions as well as qualitative discoveries about language in blogs. By continuing the work further one can identify better techniques for blog detection based on machine learning models. Based on the accuracy, better models among them are selected. This experimental result shows that the Random tree gives better performance than all the other machine learning models.

IX. FUTURE WORK

In this work, a new approach is proposed for presenting the analysis of bloggers that tend to recognize parameters by using data mining. Due to the performed information from input data of 100 users and bloggers are using weka 3.6 tools and c4.5 algorithm to provide Random tree and achieve to future anticipation of users approach, results are shown with %92 precision. If the users interested in writing memos, the basic factor of professional approach will be political thinking and academic education as the next step. If the political issues are considered important in blogging, the attitude toward local media function, political and social conditions would be basic factors in recognizing professional approach. If the users are interested in digital writing in the tourism sector, academic education and political thinking would be the basic factors in recognizing professional approach. And finally, if they choose subjects such as news and science, their professional approach would be political thinking.

The precision of random tree in our paper and its results have made one to provide strategic programs and software to planners based on decision tree in the future.

REFERENCE

[1] Zafarani,R, Jashki, M.A, Baghi,H.R , Ghorbani,A., 2008, A Novel Approach for Social Behavior Analysis of the Blogosphere, springer-Verlag Berlin Heidelberg, S. Bergler (Ed.): Canadian AI, 356–367. Techniques.

[2] Dr.M.Hemalatha,S.Krishnaveni.,et.,al, A Prespective analysis of Traffic Accident Using Data Mining Techniques.

[3] Dr.M.Hemalatha,S.Krishnaveni,G.V.Nadaimmai. ,et.,al, Evaluating the YieldOfHybrid Napier Grass with Data Mining

[4] Juffinger,A., Lex, E., 2009, Cross language Blog Mining and Trend Visualization ,WWW 2009, 2009, Madrid, Spain.1149-1150.

[5] Demartini, G., Siersdorfer,S., Chelaru, S., Nejdl,W., Analyzing Political Trends in the Blogosphere, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media,465-469.

[6] Moertini, V.S., 2003, Towards the Use of C4.5 Algorithm for classifying Banking Dataset , Oktober 2003, Integral, Vol. 8 No. 2. 105-117.

[7] Lavanya, D., Usha R., 2011, Performance Evaluation of Decision Tree Classifiers on Medical Datasets, International Journal of Computer Applications (0975 – 8887(, Volume 26– No.4, 1-4.

[8] Hartati,K, 2007, Implementation Of C4.5 Algorithm To Evaluate The Cancellation Possibility Of New Student Applicants At Stmik Amikom Yogyakarta, Proceedings of the International Conference on Electrical Engineering and Informatics Institute Technology Bandung, Indonesia June 17-19, 623-626

[9] Quinlan, J.R, 1986, Induction of Decision Trees, Machine Learning 1, Kluwer Academic Publishers, Boston. 81-106.

[10] Rosanna, E., Cassie, A. Bradley, E., Okdie, M, 2010, Personal Blogging Individual Differences and Motivations, IGI Global.292-301.

[11] Alan Bivens, Chandrika Palagiri, Rasheda Smith, Boleslaw Szymanski, "Network-Based Intrusion Detection Using Neural Networks", in Proceedings of the Intelligent Engineering Systems Through Artificial Neural Networks, St.Louis, ANNIE-2002, and Vol: 12, pp- 579-584, ASME Press, New York.

[12] Aly Ei-Semary, Janica Edmonds, Jesus Gonzalez-Pino, Mauricio Papa, "Applying Data Mining of Fuzzy Association Rules to Network Intrusion Detection", in the Proceedings of Workshop on Information Assurance United States Military Academy 2006, IEEE Communication Magazine, West Point, NY,DOI:10.1109/IAW.2006/652083.

[13] Amir Azimi, Alasti, Ahrabi, Ahmad Habibizad Navin, Hadi Bahrbegi, "A New System for Clustering & Classification of Intrusion Detection System Alerts Using SOM", International Journal of Computer Science & Security, Vol: 4, Issue: 6, pp-589-597, 2011.

[14] Anderson.J.P, "Computer Security Threat Monitoring & Surveilance", Technical Report, James P Anderson co., Fort Washington, Pennsylvania, 1980.

[15] Data Mining:Concepts and Techniques, 2nd Edition , Jiawei Han and Kamber,Morgan kaufman Publishers, Elsevier Inc,2006.

[16] Denning .D.E, "An Intrusion Detection Model", Transactions on Software Engineering, IEEE Communication Magazine, 1987,SE-13, PP-222-232,DOI:10.1109/TSE.1987.232894.

[17] Dewan Md, Farid, Mohammed Zahidur Rahman, "Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm", Journal of Computers, Vol 5, pp-23-31, Jan 2010, DOI:10.4.304/jcp 5.1.

[18] ZeroR available at http://en.Wikipedia.org/wiki/ZeroR

[19] Decision tree, available at http://en.Wikipedia.org/wiki/Decision_tree

[20] Random Forest, available at http://en.Wikipedia.org/wiki/Random_Forest

[21] KDD Cup 1999 Data, available at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[22] Jake Ryan, Meng - Jang Lin, Risto Miikkulainen, "Intrusion Detection With Neural Networks", Advances in Neural Information Processing System 10, Cambridge, MA:MIT Press,1998,DOI:10.1.1.31.3570.

[23] Jian Pei, Upadhayaya.S.J, Farooq.F, Govindaraju.V,"Data Mining for Intrusion Detection: Techniques, Applications & Systems, in the Proceedings of 20th International Conference on Data Engineering, pp-877-887, 2004.

[24]] Jin-Ling Zhao, Jiu-fen Zhao ,Jian-Jun Li, "Intrusion Detection Based on Clustering Genetic Algorithm", in Proceedings of International Conference on Machine Learning & Cybernetics (ICML),2005, IEEE Communication Magazine,ISBN:0-7803-9091-1,DOI: 10.1109/ICML.2005.1527621.

[25] Macros .M. Campos, Boriana L. Milenora, " Creation & Deployment of Data Mining based Intrusion Detection Systems in Oracle Db 10g", in the proceedings of 4th International Conference on Machine Learning & Applications, 2005.

[26] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", in Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications, pp. 53-58, Ottawa, Ontario, Canada, 2009.

[27] Norouzian.M.R, Merati.S, "Classifying Attacks in a Network Intrusion Detection System Based on Artificial Neural Networks", in the Proceedings of 13th International Conference on

Advanced Communication Technology(ICACT), 2011,ISBN:978-1-4244-8830-8,pp-868-873.

[28] Oswais.S, Snasel.V, Kromer.P, Abraham. A, "Survey: Using Genetic Algorithm Approach in Intrusion Detection Systems Techniques", in the Proceedings of 7th International Conference on Computer Information & Industrial Management Applications (CISIM), 2008,

[29] . Abdel-Aty, M., and Abdelwahab, H., Analysis andPrediction of Traffic Fatalities Resulting From AngleCollisions Including the Effect of Vehicles‟Configuration and Compatibility. Accident Analysisand Prevention, 2003.

[30] Bedard, M., Guyatt, G. H., Stones, M. J., & Hireds, J.P., The Independent Contribution of Driver, Crash,and Vehicle Characteristics to Driver Fatalities.Accident analysis and Prevention, Vol. 34, pp. 717-727, 2002.

[31] . Domingos, Pedro & Michael Pazzani (1997) "On theoptimality of the simple Bayesian classifier underzero-one loss". Machine Learning, 29:103–137.

[32] Evanco, W. M., The Potential Impact of RuralMayday Systems on Vehicular Crash Fatalities.Accident Analysis and Prevention, Vol. 31, 1999, pp.455-462.

[33] E. Frank and I. H. Witten. Generating accurate rulesets without global optimization. In Proc. of the Int‟lConf. on Machine Learning, pages144–151. MorganKaufmann Publishers Inc., 1998.

[34] Gartner Group High Performance ComputingResearch Note 1/31/95

[35] Gartner Group Advanced Technologies &Applications Research Note 2/1/95

[36] Data Mining and Data Warehousing available at:http://databases.about.com/od/datamining/g/Classification.htm

[37] Genetic algorithm available at:http://en.wikipedia.org/wiki/Genetic_algorithm

[38] Road Traffic Accident Statistics available at:http://www.td.gov.hk/en/road_safety/road_traffic_accident_statistics/2008/index.html

[39] Statistical Analysis Software, Data Mining, PredictiveAnalytics available at:http://www.statsoft.com/txtbook/stdatmin.html

[40] Data Mining: Bagging and Boosting available at:http://www.icaen.uiowa.edu/~comp/Public/Bagging.pdf

[41] Kweon, Y. J., & Kockelman, D. M., Overall InjuryRisk to Different Drivers: Combining Exposure,Frequency, and Severity Models. Accident Analysisand Prevention, Vol. 35, 2003, pp. 441-450.

[42] Miaou, S.P. and Harry, L. 1993, "Modeling vehicleaccidents and highway geometric designrelationships". Accidents Analysis and Prevention,(6), pp. 689–709.27. Desktop Reference for CrashReduction Factors Report No. FHWA-SA-07-015,Federal Highway Administration September,2007http://www.ite.org/safety/issuebriefs/Desktop%20Reference%20Complete.pdf

[43] Martin, P. G., Crandall, J. R., & Pilkey, W. D., InjuryTrends of Passenger Car Drivers In the USA.Accident Analysis and Prevention, Vol. 32, 2000, pp.541-557.

[44] National Highway Traffic Safety Administration,Traffic Safety Facts 2005, 2007, P. 54. http://wwwnrd.nhtsa.dot.gov/Pubs/TSF2006.PDF

[45] Ossenbruggen, P.J., Pendharkar, J. and Ivan, J. 2001,"Roadway safety in rural and small urbanized areas".Accidents Analysis and Prevention, 33 (4), pp. 485–498.

[46] . Quinlan, J. R. C4.5: Programs for Machine Learning.Morgan Kaufmann Publishers, 1993.

[47] Rish, Irina. (2001). "An empirical study of the naiveBayes classifier". IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.