Analysis of Bayes, Neural Network and Tree Classifier of Classification Technique in Data Mining using WEKA

Yugal kumar[1] and G. Sahoo[2]

[1]Assistant Professor in CSE/IT Dept., Hindu College of Engineering, Industrial Area, Sonepat, Haryana, India.
yugalkumar.14@gmail.com
[2]Professor in Dept. of Information Technology, Birla Institute of Technology, Mesra, Ranchi, Jhrakhand, India.
gsahoo@bitmesra.ac.in

ABSTRACT

In today's world, gigantic amount of data is available in science, industry, business and many other areas. This data can provide valuable information which can be used by management for making important decisions. But problem is that how can find valuable information. The answer is data mining. Data Mining is popular topic among researchers. There is lot of work that cannot be explored till now. But, this paper focuses on the fundamental concept of the Data mining i.e. Classification Techniques. In this paper BayesNet, NavieBayes, NavieBayes Uptable, Multilayer perceptron, Voted perceptron and J48 classifiers are used for the classification of data set. The performance of these classifiers analyzed with the help of Mean Absolute Error, Root Mean-Squared Error and Time Taken to build the model and the result can be shown statistical as well as graphically. For this purpose the WEKA data mining tool is used.

KEY TERM'S

BayesNet, J48, Mean Absolute Error, NavieBayes, Root Mean-Squared Error

1. INTRODUCTION

In recent years, there is the incremental growth in the electronic data management methods. Each companies whether it is large, medium or small, having its own database system that are used for collecting and managing the information, these information are used in the decision process. Database of any firm consist the thousands of the instance and hundreds of attributes. So, it is quite difficult to process the data and retrieving meaning full information from the data set in short span of time. The same problem is faced by researchers and scientists how to process the large data set for further research. To overcome this problem the term data mining come into existence. Data mining refers to the process of retrieving information from large sets of data. A number of algorithms and tools have been developed and implemented to retrieve information and discover knowledge patterns that may be useful for decision support [2]. The term Data Mining, also known as Knowledge Discovery in Databases (KDD) refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [1]. Several data mining techniques are pattern recognition, clustering, association and classification [4]. Classification has been identified as an important problem in the emerging field of data mining [3] as they try to find meaningful ways to interpret data sets. Some ethical

David C. Wyld, et al. (Eds): CCSEA, SEA, CLOUD, DKMP, CS & IT 05, pp. 359–369, 2012.
 DOI : 10.5121/csit.2012.2236

issue also related with Data mining for example process a data set that are belongs to racial, sexual, religious may occur some discernment.

2. CLASSIFICATION

Classification of data is very typical task in data mining. There are large number of classifiers that are used to classify the data such as bayes, function, rule based and Tree etc. The goal of classification is to correctly predict the value of a designated discrete class variable, given a vector of predictors or attributes [5].

2.1. BayseNet

BayesNet based on the bayes theorm. So, in BayesNet conditional probability on each node is calculated and formed a Bayesian Network. Bayesian Network is a directed acyclic graph. In BayesNet, it is assume that all attributes are nominal and there are no missing values any such value replaced globally. Different types of algorithms are used to estimate conditional probability such as Hill Climbing, Tabu Search, Simulated Annealing, Genetic Algorithm and K2. The output of the BayesNet can be visualized in terms of graph. Figure 1 shows the visualized graph of the BayesNet for a bank data set [9]. Visualize graph is formed by using the children attribute of the bank data set. In this graph, each node represents the probability distribution table within it.

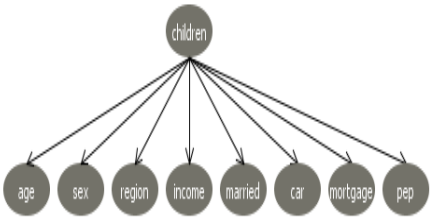


Fig. 1 Visualize Graph of the BayesNet for a bank data set

2.2. NaiveBayes

NaiveBayes is widely used for the classification due to its simplicity, elegance, and robustness. NavieBayes can be characterized as Navie and Bayes. Navie stands for independence i.e. true to multiply probabilities when the events are independent and Bayes is used for the bayes rule. This technique assumes that attributes of a class are independent in real life. The performance of the NavieBayes is better when the data set is actual. Kernel density estimators can be used to measure the probability in NavieBayes that improve the performance of the model. A large number of modifications have been introduced, by the statistical, data mining, machine learning, and pattern recognition communities, in an attempt to make it more flexible, but one has to recognize that such modifications are necessarily complications, which detract from its basic simplicity.

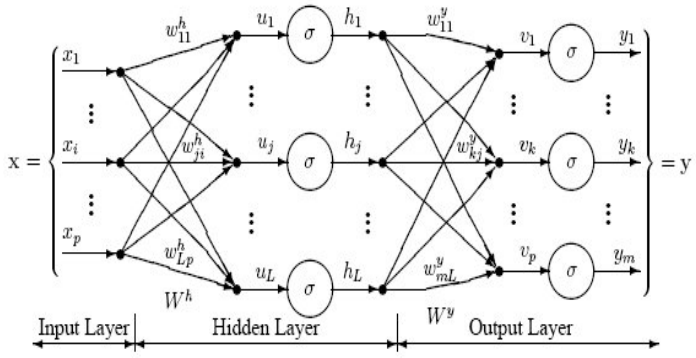
2.3. Navie Bayes Updatable

This is the updateable version of NaiveBayes. This classifier will use a default precision of 0.1 for numeric attributes when buildClassifier is called with zero training instances and also known as incremental update.

2.4. Multi Layer Percpeptron

Multi Layer Perceptron can be defined as Neural Network and Artificial intelligence without qualification. A Multi Layer perceptron (MLP) is a feedforward neural network with one or more

layers between input and output layer. The following diagram illustrates a perceptron network with three layers:



Each neuron in each layer is connected to every neuron in the adjacent layers. The training or testing vectors are presented to the input layer, and processed by the hidden and output layers. A Detailed analysis of multi-layer perceptrons has been presented by Hassoun [11] and by Żak[10].

2.5. Voted Perceptron

Voted Perceptron (VP) proposed by Collins can be viewed as a simplified version of CRF[1] and suggests that the voted perceptron is preferable in cases of noisy or un separable data[3]. Voted perceptron approaches to small sample analysis and taking advantage of the boundary data of largest margin. Voted perceptron method is based on the perceptron algorithm of Rosenblatt and Frank [2].

2.6. J48

J48 are the improved versions of C4.5 algorithms or can be called as optimized implementation of the C4.5. The output of J48 is the Decision tree. A Decision tree is similar to the tree structure having root node, intermediate nodes and leaf node. Each node in the tree consist a decision and that decision leads to our result. Decision tree divide the input space of a data set into mutually exclusive areas, each area having a label, a value or an action to describe its data points. Splitting criterion is used to calculate which attribute is the best to split that portion tree of the training data that reaches a particular node. Fig. 2 shows the decision tree using J48 for a bank data set whether a bank provide loan to a person or not. Decision tree is formed by using the children attribute of the bank data set.

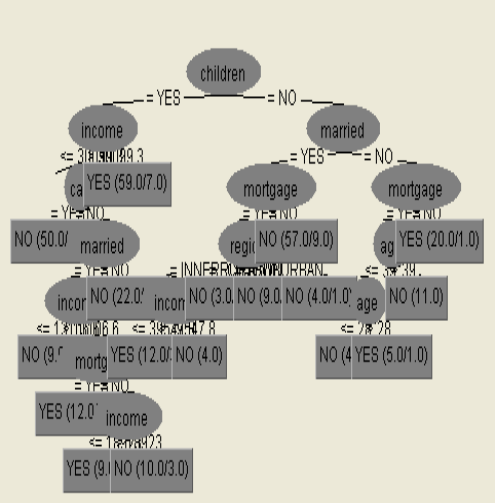


Fig. 2 Decision Tree using J48 for Bank Data Set

3. TOOL

The WEKA toolkit is used to analyze the dataset with the data mining algorithms [7]. WEKA is an assembly of tools of data classification, regression, clustering, association rules and visualization. The toolkit is developed in Java and is open source software issued under the GNU General public License [8]. The WEKA tool incorporates the four applications within it.

- Weka Explorer
- Weka Experiment
- Weka Knowledge Flow
- Simple CLI

For the Classification of Data set, weka explorer is used to generate the result or statistics. Weka Explorer incorporates the following features within it:-

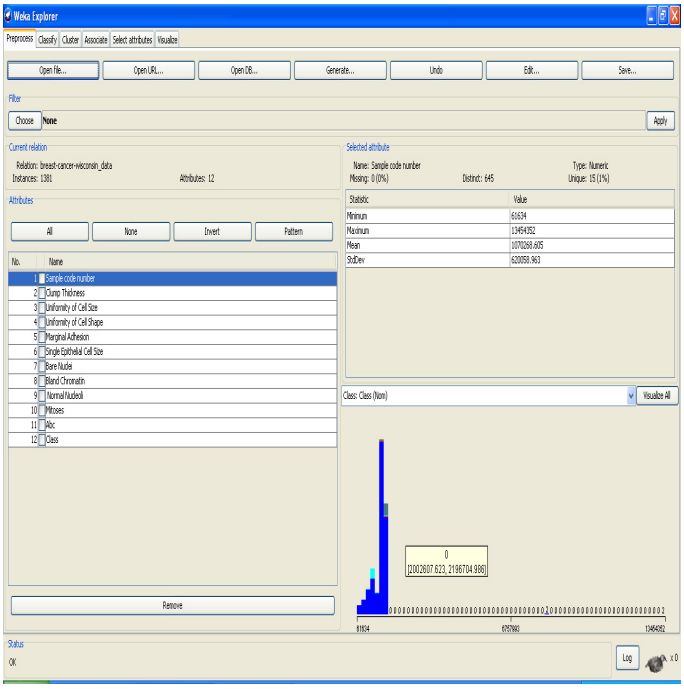


Fig. 3 Pre process of data using weka

- Preprocess: It is used to process the input data. For this purpose the filters are used that can transform the data from one form to another form. Basically two types of filters are used i.e. supervised and unsupervised.

- Classify. Classify tab are used for the classification purpose. A large number of classifiers are used in weka such as bayes, function, rule, tree and meta etc. Four type of test option are mentioned within it.

- Cluster: It is used for the clustering of the data.

- Associate: Establish the association rules for the data.

- Select attributes: It is used to select the most relevant attributes in the data.

- Visualize: View an interactive 2D plot of the data.

Data set used in Weka is in Attribute-Relation File Format (ARFF) file format that consist of special tags to indicate different things in the dataset such as attribute names, attribute types, attribute values and the data. This paper includes the two data sets such as sick.arff and breast-cancer-wisconsin. Sick.arff data set has been taken from the weka tool website while the brest cancer data set has been taken from the UCI repository i.e. real time multivariate data set [7, 9]. Brest cancer data set is in the form of text file. Firstly it converts into the .xls format; .xls format to .csv format and then .csv format convert into the .arff format. The .arff format of both data sets given as:-

Sick.arff Data Set:

```
@relation sick.nm
@attribute age    real
@attribute sex    {M,F}
@attribute on_thyroxine           {f,t}
@attribute query_on_thyroxine   {f,t}
@attribute on_antithyroid_medication     {f,t}
@attribute sick   {f,t}
@attribute pregnant       {f,t}
@attribute thyroid_surgery        {f,t}
@attribute I131_treatment         {f,t}
@attribute query_hypothyroid    {f,t}
@attribute query_hyperthyroid   {f,t}
@attribute lithium          {f,t}
@attribute goitre           {f,t}
@attribute tumor            {f,t}
@attribute hypopituitary          {f,t}
@attribute psych            {f,t}
@attribute TSHmeasured            {f,t}
@attribute TSH  real
@attribute T3measured   {f,t}
@attribute T3     real
@attribute TT4measured            {f,t}
@attribute TT4  real
@attribute T4Umeasured            {f,t}
@attribute T4U  real
@attribute FTImeasured            {f,t}
@attribute FTI    real
@attribute TBGmeasured            {f,t}
@attribute TBG real
@attribute referral_source        {WEST,STMW,SVHC,SVI,SVHD,other}
@attribute class  {sick,negative}
@data
```

Breast-cancer-wisconsin_data,arff Data Set:

```
@relation breast-cancer
@attribute age {'10-19','20-29','30-39','40-49','50-59','60-69','70-79','80-89','90-99'}
@attribute menopause {'lt40','ge40','premeno'}
@attribute tumor-size {'0-4','5-9','10-14','15-19','20-24','25-29','30-34','35-39','40-44','45-49','50-
54','55-59'}
@attribute inv-nodes {'0-2','3-5','6-8','9-11','12-14','15-17','18-20','21-23','24-26','27-29','30-
32','33-35','36-39'}
@attribute node-caps {'yes','no'}
@attribute deg-malig {'1','2','3'}
@attribute breast {'left','right'}
@attribute breast-quad {'left_up','left_low','right_up','right_low','central'}
@attribute 'irradiat' {'yes','no'}
@attribute 'Class' {'no-recurrence-events','recurrence-events'}
@data
```

4. RESULT & DISCUSION

In this paper, the following parameters are used to evaluate the performance of above mentioned classification techniques:

- Mean Absolute Error (MAE): It can define as statistical measure of how far an estimate from actual values i.e. the average of the absolute magnitude of the individual errors. It is usually similar in magnitude but slightly smaller than the root mean squared error.

- Root Mean-Squared Error (RMSE): The root mean square error (RMSE)) calculates the differences between values predicted by a model / an estimator and the values actually observed from the thing being modeled/ estimated. RMSE is used to measure the accuracy. It is ideal if it is small.

- Time: The amount of time required to build the model.

Table 1 Comparison of the different classifiers

Algorithm (Total Instance: 2800/286)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (Value)	Time Taken (seconds)	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
BayesNet	97.1429 %	2.8571 %	0.2	0.7662	0.0476	0.1651	41.3751 %	68.9458 %
	72.028 %	27.972 %	0.02	0.2919	0.3297	0.4566	78.7898 %	99.9047 %
NaiveBayes	97.2857 %	2.7143 %	0.13	0.7756	0.0456	0.1594	39.6417 %	66.5598 %
	71.6783 %	28.3217 %	0.02	0.2857	0.3272	0.4534	78.2086 %	99.1872 %
Naïve Bayes Updateable	92.5714 %	7.4286 %	0.03	0.5182	0.0886	0.2296	77.0785 %	95.8833 %
	71.6783 %	28.3217 %	0	0.2857	0.3272	0.4534	78.2086 %	99.1872 %
Multilayer Perceptron	97.8214 %	2.1786 %	110.94	0.7937	0.0265	0.1355	23.0698 %	56.6009 %
	64.6853 %	35.3147 %	8.91	0.1575	0.3552	0.5423	84.8811 %	118.654 %
Voted Perceptron	93.6429 %	6.3571 %	0.77	0.0335	0.0636	0.2521	55.2995 %	105.292 %
	71.3287 %	28.6713 %	0.03	0.212	0.2848	0.5322	68.0628 %	116.4466 %
J48	99.6786 %	0.3214 %	0.3	0.972	0.0066	0.0533	5.7569 %	22.2529 %
	75.5245 %	24.4755 %	0.02	0.2826	0.3676	0.4324	87.8635 %	94.6093 %

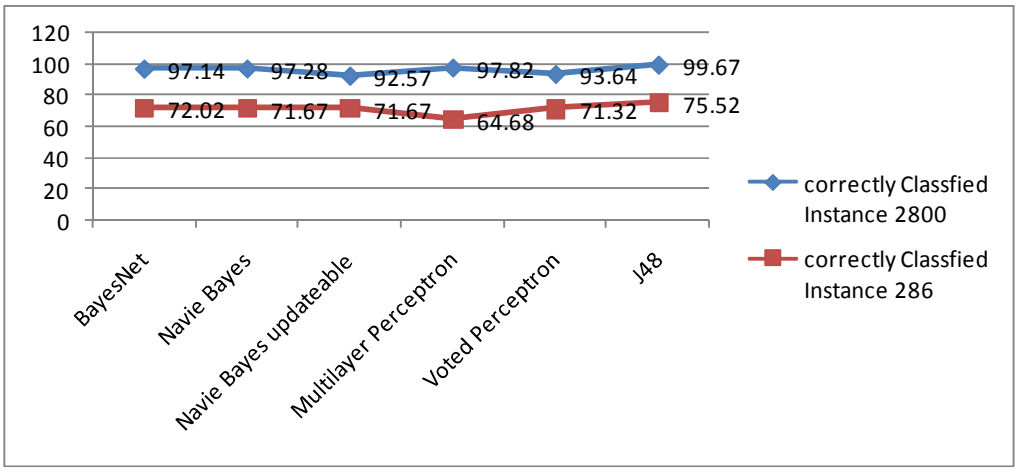


Fig. 4 Comparison of Correctly Classified Parameter of Datasets

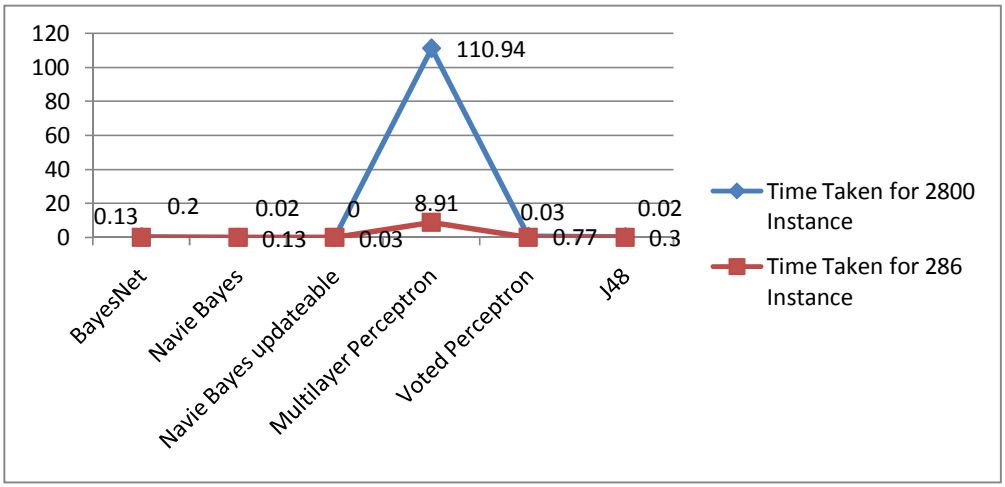


Fig. 5 Comparison of Time Taken Parameter of Datasets

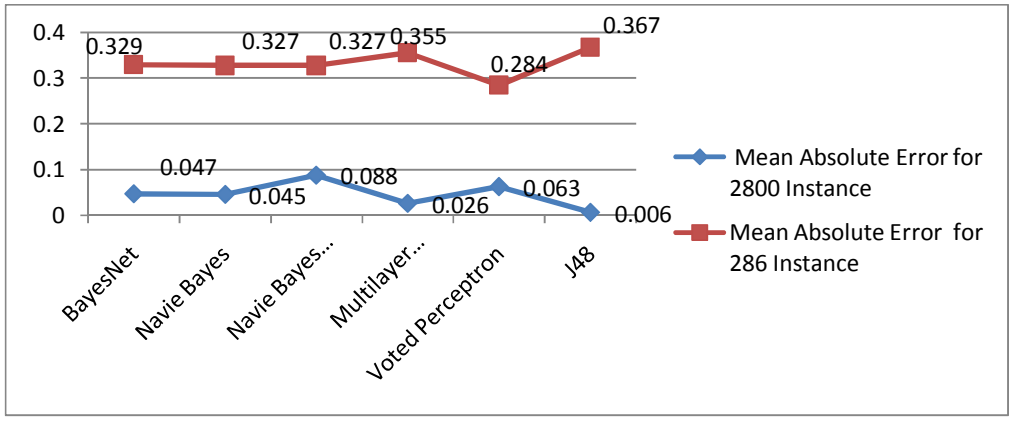


Fig. 6 Comparison of Mean Absolute Error Parameter

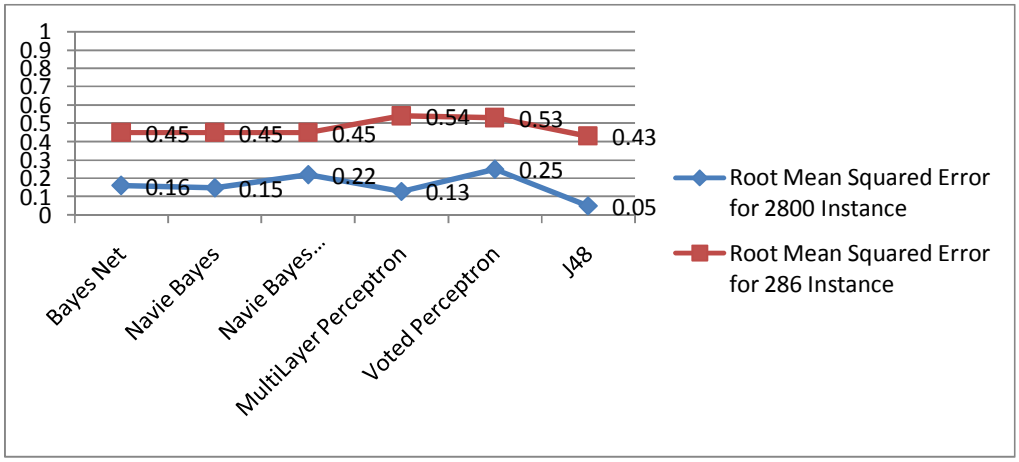


Fig. 7 Comparison of Root Mean Squared Error Parameter

Table 1 shows the comparison of the BayesNet, NavieBayes NavieBayes Uptable, Multilayer perceptron, Voted perceptron and J48. For the analysis of discussed classifiers the two data sets has been used in which breast cancer data set has 286 instance and 10 attributes while the sick data set has 2800 instance and 30 attributes. From the table 1, it is clear that the time taken by the NavieBayes uptable classifiers to build the model is smallest for both of data set i.e. 0.03s and 0.0s whereas the time taken by the multilayer perceptron is the largest. So, in terms of time taken the NavieBayes uptable classifier is the best among these. But the analysis of another two parameter i.e. MAE and RMSE, the model formed by J48 classifier is better. J48 classfier classified the instance more correctly as compare to BayesNet and Navie Bayes. It is also seen that the performance of Naviebayes uptable and navie bayes classifiers almost same when the dataset is small.

5. CONCLUSION

In this paper, six different classifiers are used for the classification of data. These techniques are applied on two dataset in which one of data set has one tenth of instance and one third attribute as compare to another data set. The fundamental concept to take two datasets is to analyze the performance of the discussed classifiers for small as well as large dataset. But, it cannot say easily which one is better. For example, mean absolute error of J48 is minimum for breast cancer data set (i.e. small data set) but not minimum for sick data set (i.e. large data set) for from the table 1, it says that the performance of the J48 classifier/technique is better as compare to another classifier/technique.

6. FUTURE WORK

In weka, there are the large numbers of classifiers such as fuzzy rules, REP tree, Random tree, Gaussian Function, Regression and so on. So the future work will be based on these classifiers i.e. apply these classifiers on the data set and analyze the performance of these classifiers. In this paper, six parameters are used for the analysis the performance of the classifiers. In future, numbers of parameter will be increased such that better result will be obtained.

REFERENCES

[1] J. Han and M. Kamber, (2000) "Data Mining: Concepts and Techniques," Morgan Kaufmann.

[2] Desouza, K.C. (2001) ,Artificial intelligence for healthcare management In Proceedings of the First International Conference on Management of Healthcare and Medical Technology Enschede, Netherlands Institute for Healthcare Technology Management.

[3] Rakesh Agrawal,Tomasz Imielinski and Arun Swami, (1993)" Data mining : A Performance perspective". IEEE Transactions on Knowledge and Data Engineering , 5(6):914-925.

[4] Ritu Chauhan, Harleen Kaur, M.Afshar Alam, (2010) "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases", International Journal of Computer Applications (0975 – 8887) Volume 10– No.6.

[5] Daniel Grossman and Pedro Domingos (2004). Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood. In Press of Proceedings of the 21st International Conference on Machine Learning, Banff, Canada.

[6] Ridgeway G, Madigan D, Richardson T (1998) Interpretable boosted naive Bayes classification. In: Agrawal R, StolorzP, Piatetsky-Shapiro G (eds) Proceedings of the fourth international conference on knowledge discovery and data mining.. AAAI Press, Menlo Park pp 101–104.

[7] Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/

[8] Ian H.Witten and Elbe Frank, (2005) "Datamining Practical Machine Learning Tools and Techniques," Second Edition, Morgan Kaufmann, San Fransisco.

[9] www.ics.uci.edu/~mlearn

[10] Zak S.H., (2003), " Systems and Control" NY: Oxford Uninversity Press.

[11] Hassoun M.H, (1999), " Fundamentals of Artificial Neural Networks", Cambridge, MA: MIT press.

[12] Yoav Freund, Robert E. Schapire, (1999) "Large Margin Classification Using the Perceptron Algorithm." In: Machine Learning, 37(3).

[13] Yunhua Hu, Hang Li, Yunbo Cao, Li Teng, Dmitriy Meyerzon, Qinghua Zheng, (2006), " Automatic extraction of titles from general documents using machine learning", in Information Processing and Management(publisheb by elesvier) 42, 1276–1293.

[14] Michael Collins and Nigel Duffy, (2002), "New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron" in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 263-270.

Authors Bibliography

G. Sahoo received his MSc in Mathematics from Utkal University in the year 1980 and PhD in the Area of Computational Mathematics from Indian Institute of Technology, Kharagpur in the year 1987. He has been associated with Birla Institute of Technology, Mesra, Ranchi, India since 1988, and currently, he is working as a Professor and Head in the Department of Information Technology. His research interest includes theoretical computer science, parallel and distributed computing, cloud computing, evolutionary computing, information security, image processing and pattern recognition.

Mr. Yugal Kumar received his B.Tech in Information Technology from Maharishi Dayanand University, Rohtak, (India) in 2006 & M.Tech in Computer Engineering from Maharishi Dayanand University, Rohtak, India in 2009. His research interests include fuzzy logic, computer network and Data Mining & Swarm Intelligence system. At present, he has been worked as working as an Assistant Professor in Department of Computer Science and Engineering, Hindu College of Engineering, Sonepat, Haryana, India.