ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

Comparative Analysis of Bayes and Lazy Classification Algorithms

Ms S. Vijayarani[1], Ms M. Muthulakshmi[2]

Assistant Professor, Department of Computer Science, School of Computer Science and Engineering,
Bharathiar University, Coimbatore, Tamilnadu, India[1]

M.Phil Research Scholar, Department of Computer Science, School of Computer Science and Engineering,
Bharathiar University, Coimbatore, Tamilnadu, India[2]

Abstract: Data mining is the non-trivial extraction of implicit, earlier unknown and potentially useful information about data. There are several data mining techniques have been developed and used in data mining projects which includes classification, clustering, association rules, prediction, and sequential patterns. Data mining applications are used in various areas such as sales, marketing, banking, finance, health care, insurance and medicine. There are various research domains in data mining namely web mining, text mining, image mining, sequence mining, privacy preserving data mining, etc. Text mining is a technique which extracts information from both structured and unstructured data and also finding patterns which is novel and not known earlier. It is also known as knowledge discovery from text (KDT), deals with the machine supported analysis of text. Text mining is used in various areas such as information retrieval, document similarity, natural language processing and so on. Searching for similar documents is an important problem in text mining. The first and essential step of document similarity is to classify the documents based on their category. In this research work, we have analysed the performance of Bayesian and Lazy classifiers for classifying the files which are stored in the computer hard disk. There are two algorithms in Bayesian classifier namely BayesNet, and Naïve Bayes. In lazy classifier has three algorithms namely IBL, IBK and Kstar. The performances of Bayesian and lazy classifiers are analysed by applying various performance factors. From the experimental results, it is observed that the lazy classifier is more efficient than Bayesian classifier.

Keywords: Data mining, Text mining, Classification, Bayesian, BayesNet, Lazy, IBK, Naïve Bayes, IBL, Kstar.

I. INTRODUCTION

Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text mining is also known as text data mining, quick text analysis or knowledge-discovery in text (KDT) refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is an interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. Text mining is similar to data mining, except that data mining tools are intended to handle structured data from databases, but text mining can effort with unstructured or semi-structured data sets such as emails, html files and full-text documents etc. As a result, text mining is a much better solution for companies. [10]

Text mining usually involves the process of structuring the input text (usually parsing, along with the accumulation of some derived linguistic features and the removal of others, and consequent insertion into a database), deriving models within the structured data, and to finish evaluation and interpretation of the output. High quality in text mining typically refers to some combination of relevance of relevance, innovation, and interestingness. Text mining involves the application of techniques from areas such as data mining, information retrieval and natural language processing. Various stages of a text-mining process can be combined together into a single workflow. [11]

Some of the important applications of text-mining include Enterprise Business Intelligence, Data Mining Competitive Intelligence, E-Discovery, Records Management, National Security, Intelligence Scientific discovery especially Life Sciences, Search or Information Access and Social media monitoring. Some of the technologies that have been developed and can be used in the text mining process are information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, and question answering. [10]

The rest of this paper is organized as follows. Section 2 describes the review of literature. Section 3 discusses the Bayesian and lazy classifiers and the various algorithms used for classification. Experimental results are analysed in Section 4 and Conclusions are given in Section 5.

II. LITERATURE REVIEW

Mahendra Tiwari et al., [8] proposed the use of data mining technique to help retailers to identify customer profile for a retail store and behaviours, improve better customer fulfillment and retention. The aim is to evaluate the accuracy of different data mining algorithms on various data sets. The performance investigation depends on many factors about test mode, different nature of data sets and size of data set.

Dr. S.Vijayarani et al., [14] analyses the performance of different classification function techniques in data mining for predicting the heart disease from the heart disease dataset. The classification function algorithms is used and tested in this work. The performance factors used for analysing the efficiency of algorithms are clustering accuracy and error rate. The result illustrates shows LOGISTICS classification function efficiency is better than multilayer perception and sequential minimal optimization.

Anshul Goyal et al., [3] proposed a performance evaluation of naïve bayes and J48 classification algorithms. The experimental results shown in the study are about classification accuracy and cost analysis. J48 gives more classification accuracy for class gender in bank dataset having two values Male and Female. The result in the study on these datasets also shows that the efficiency and accuracy of j48 and Naive bayes is good.

Kaushik H. Raviya et al., [6] presents the comparison on three classification techniques which are K-nearest neighbour, Bayesian network and Decision tree respectively. The aim of this research is to enumerate the best technique from the above three techniques. There is a direct relationship between execution time in building the tree model and the volume of data records and also there is an indirect relationship between execution time in building the model and attribute size of the data sets.

BS Harish et al., [16] presented various text representation schemes and compared different classifiers used to classify text documents to the predefined classes. The existing methods are compared and contrasted based on various parameters namely criteria used for classification, algorithms and classification time complexities. There is no single representation scheme and classifier can be recommended as a general model for any application. Different algorithms perform differently depending on data collections. None of them appears globally superior over the other. However, to the certain extent SVM with term weighted VSM representation scheme performs well in many text classification tasks.

Aurangzeb Khan et al., [15] proposed the important techniques and methodologies that are employed in text documents classification, while at the same time making awareness of some of the interesting challenges that remain to be solved, focused mainly on text representation and machine learning techniques. Several algorithms or combination of algorithms as hybrid approaches were proposed for the automatics classification of documents.

III. METHODOLOGY

Text classification is one of the important research issues in the field of text mining, where the documents are classified with supervised knowledge. The main objective of this research work is to find the best classification algorithm among Bayesian and lazy classifiers. The system architecture of the research work is as follows:

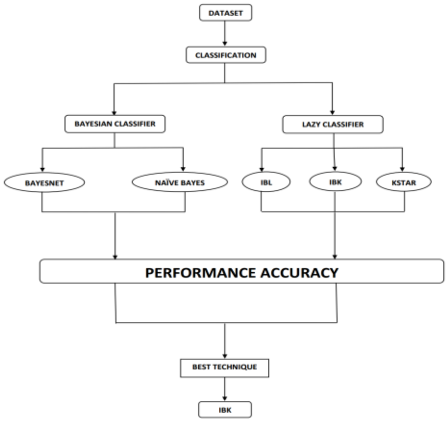


Fig 1: System Architecture of Classification Algorithms

A. Dataset

In order to compare the data mining classification techniques, computer files can be collected from the system hard disk and a synthetic data set is created. This dataset has 80000 instances and four attributes namely file name, file size, file extension and file path. Weka data mining tool is used for analysing the performance of the classification algorithms.

B. Classification

Classification is an important data mining technique with broad applications. It is used to classify each item in a set of data into one of predefined set of classes or groups. Classification algorithm plays an important role in document classification. In this research, we have analysed two classifiers namely Bayesian and lazy. In Bayesian classifier, we have analysed two classification algorithms namely BayesNet and naïve bayes, in lazy classifier we have analysed three classification algorithms such as IBL, IBK and Kstar.

C. Bayesian Classifier

Bayesian networks are a powerful probabilistic representation, and their use for classification has received considerable attention. Bayesian algorithms predict the class depending on the probability of belonging to that class. A Bayesian network is a graphical model for probability relationships among a set of variables features. This Bayesian Network consists of two components. First component is mainly a directed acyclic graph (DAG) in which the nodes in the graph are called the random variables and the edges between the nodes or random variables represents the probabilistic dependencies among the corresponding random variables. [1] Second component is a set of parameters that describe the conditional probability of each variable given its parents. The conditional dependencies in the graph are estimated by statistical and computational methods. Thus the Bayesian Network combines the properties of computer science and statistics.

BayesNet:

BayesNet learns Bayesian networks made in nominal attributes (numeric ones are prediscretized) and no missing values (any such values are replaced globally). Bayes Nets or Bayesian networks are graphical representation for probabilistic relationships among a set of random variables. Given a finite set $\mathbf{X=\{X_{1\ldots}X_n\}}$ of discrete random variables where each variable X_i may take values from a finite set represented by Val (X_i). [5]

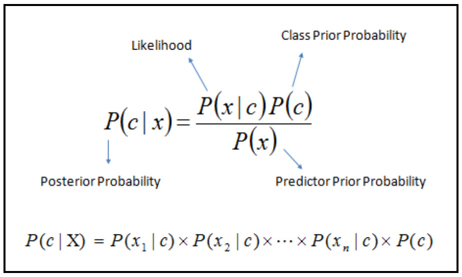
A Bayesian network is an annotated directed acyclic graph (DAG) G that encodes a joint probability distribution over X. The nodes of the graph correspond to the random variables $X_1\ldots X_n$. The links of the graph represent to the direct influence from one variable to the other. If there is a directed relationship from variable X_i to variable X_j, variable X_i will be a parent of variable X_j. Each node is annotated with a conditional probability distribution (CPD) that represents P (X_i | Pa (X_i)) where Pa (X_i) denotes the parents of X_i in G. [5]. The pair (G, CPD) encodes the joint distribution $P(X_{1\ldots}X_n)$. A unique joint probability distribution over X from G is factorized as:

$$\boxed{\mathbf{P(X_1\ldots X_n) = \Pi_i\ (P\ (X_i \mid Pa\ (X_i)))}}$$

Naïve Bayes:

Naive Bayes implements the probabilistic Naïve Bayes classifier. Naïve Bayes Simple uses the normal distribution to model numeric attributes. Naïve Bayes can use kernel density estimators, which develop performance if the normality assumption if grossly correct; it can also handle numeric attributes using supervised discretization. Naïve Bayes Updateable is an incremental version that processes one request at a time. It can use a kernel estimator but not discretization. [4]

The Naive Bayes algorithm is based on conditional probabilities. NB uses Bayes' Theorem that is a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes' Theorem determines the probability of an event occurring given the probability of another event that has already occurred.



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P (c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P (x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

D. Lazy Classifier

Lazy learners store the training instances and do no real work until classification time. Lazy learning is a learning method in which generalization beyond the training data is delayed until a query is made to the system where the system tries to generalize the training data before receiving queries.

The main advantage gained in employing a lazy learning method is that the target function will be approximated locally such as in the k-nearest neighbour algorithm. Because the objective function is approximated locally for each query to the system, lazy learning systems can concurrently solve multiple problems and deal successfully with changes in the problem arena. [5][8]

The disadvantages with lazy learning include the large space requirement to store the complete training dataset. Mostly noisy training data increases the case support unnecessarily, because no concept is made during the training phase and another disadvantage is that lazy learning methods are usually slower to evaluate, though this is joined with a faster training phase.

IBL (Instance Based Learning):

IBL is a basic instance-based learner which finds the training instance closest in Euclidean distance to the given test instance and predicts the same class as this training distance. If several instances qualify as the closest, the first one found is used. IBL algorithms do not construct extensional concept descriptions. Alternatively, concept descriptions are determined by how the IBL algorithm's selected similarity and classification function use the current set of saved distances. [8] These functions are two of the three components in the following framework that describes all IBL algorithms:

- Similarity Function: This calculates the similarity between training instances i and the instances in the concept depiction. Similarities are numeric-valued.
- Classification Function: This obtains the similarity function's results and the classification performance records of the instances in the concept description. It returns a classification for i.
- Concept Description Updater: This retains records on classification performance and decides which instances to include in the concept description. Inputs include i, the classification results, the similarity results, and a current concept description. It returns the modified concept description.[7]

Algorithm

```
The IBl algorithm (CD=Concept
Description)
CD←φ
For each x Є Training Set do
1.  For each y Є CD do
    Sim[y] ← Similarity(x, y)
2.  y_max←some y Є CD with maximal
    Sim[y]
3.  if class(x) = class(y_max)
    then classification ← correct
    else classification ← incorrect
```

IBK (K - Nearest Neighbour):

IBK is a k-nearest-neighbour classifier that uses the same distance metric. The number of nearest neighbours can be specified explicitly in the object editor or determined automatically using leave-one-out cross-validation focus to an upper limit given by the specified value. IBK is a k-nearest-neighbour classifier. A kind of different search algorithms can be used to speed up the task of finding the nearest neighbours. A linear search is the default but further options include KD-trees, ball trees, and so-called "cover trees".

The distance function used is a parameter of the search method. The remaining thing is the same as for IBL—that is, the Euclidean distance; other options include Chebyshev, Manhattan, and Minkowski distances. [10] Predictions from more than one neighbour can be weighted according to their distance from the test instance and two different formulas are implemented for converting the distance into a weight. [5][13]

The number of training instances kept by the classifier can be restricted by setting the window size option. As new training instances are added, the oldest ones are detached to maintain the number of training instances at this size.

Algorithm:

```
K -Nearest neighbour algorithm
Training
Build the set of training examples D.
Classification
Given a query instance $x_q$ to be classified,
Let $x_1$... $x_k$ denote the k instances from D that are nearest to $x_q$
Return
$F(x_q) = \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i))$
where (a, b) =1, if a = b, and -(a, b)=0 otherwise.
```

Kstar:

The K* algorithm can be defined as a method of cluster analysis which mainly aims at the partition of 'n' observation into 'k' clusters in which each observation belongs to the cluster with the nearest mean. We can describe K* algorithm as an instance based learner which uses entropy as a distance measure. The benefits are that it provides a consistent approach to handling of real valued attributes, symbolic attributes and missing values. [12]

K* is a simple, instance based classifier, similar to K-Nearest Neighbour (K-NN). New data instances, *x*, are assigned to the class that occurs most frequently amongst the k-nearest data points, y_j, where j = 1, 2…k. Entropic distance is then used to retrieve the most similar instances from the data set. By means of entropic distance as a metric has a number of benefits including handling of real valued attributes and missing values. The K* function can be calculated as:

$$\boxed{K^*(y_i, x) = -\ln P^*(y_i, x)}$$

Where *P** is the probability of all transformational paths from instance *x* to *y*. It can be useful to understand this as the probability that *x* will arrive at *y* via a random walk in IC feature space. It will performed optimization over the percent blending ratio parameter which is analogous to K-NN 'sphere of influence', prior to assessment with other Machine Learning methods.

IV. EXPERIMENTAL RESULTS

A. *Accuracy Measure and Error Rate*

The following tables show the accuracy measure of classification techniques. They are the True Positive rate, F Measure, Receiver Operating Characteristics (ROC) Area and Kappa Statistics. The TP Rate is the ratio of play cases predicted correctly cases to the total of positive cases. It is a probability corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. F Measure is a way of combining recall and precision scores into a single measure of performance. Recall is the ratio of relevant documents found in the search result to the total of all relevant documents. Precision is the proportion of relevant documents in the results returned. ROC Area is a traditional to plot this same information in a normalized form with 1-false negative rate plotted against the false positive rate.

They are the Mean Absolute Error (M.A.E), Root Mean Square Error (R.M.S.E), Relative Absolute Error (R.A.E) and Root Relative Squared Error (R.R.S.R) [10]. The mean absolute error (MAE) is defined as the quantity used to measure how close predictions or forecasts are to the eventual outcomes. The root mean square error (RMSE) is defined as frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. It is a good measure of accuracy, to compare the forecasting errors within a dataset as it is scale-dependent. Relative error is a measure of the uncertainty of measurement compared to the size of the measurement. The root relative squared error is defined as a relative to what it would have been if a simple predictor had been used. More specifically, this predictor is just the average of the actual values.

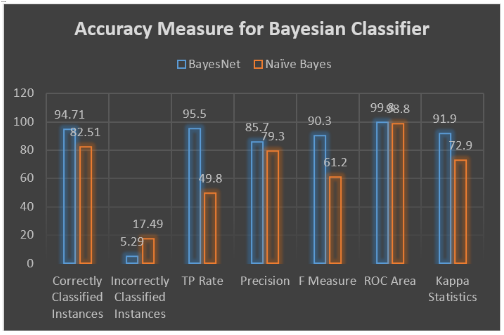


Fig 2: Accuracy Measure for Bayesian Classifier

From the analysis of Accuracy Measures of Bayesian Algorithm from the Table 1, BayesNet performs well when compared to all accuracy measures namely TP rate, F Measure, ROC Area and Kappa Statistic. As a result BayesNet outperforms well when compared to other Bayesian algorithm.

Algorithm	MAE	RMSE	RAE	RRSR
IBL	0.49	6.97	6.89	37.14
IBK	0.50	5.95	7.06	31.67
Kstar	2.71	14.9	42.95	85.23

TABLE 3
ERROR RATE OF BAYESIAN CLASSIFIER

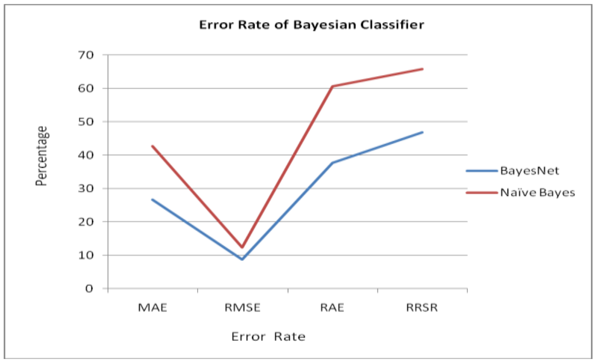


Fig 3: Error Rate for Bayesian Classifier

From the graph, it is observed that Naïve bayes attains highest error rate. Therefore the BayesNet classification algorithm performs well because it contains least error rate when compared to Naïve bayes algorithm.

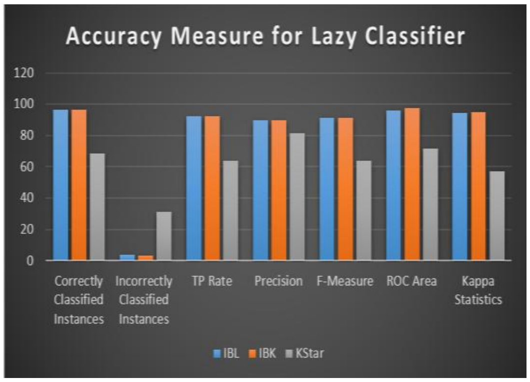


Fig 4: Accuracy Measure for Lazy Classifier

From the analysis of Accuracy Measures of Lazy Classifier from the table 2, IBK performs well when compared to all accuracy measures namely TP rate, F Measure, ROC Area and Kappa Statistic. As a result IBK outperforms well when compared to other Lazy algorithms.

TABLE 4
ERROR RATE OF LAZY CLASSIFIER

Algorithm	MAE	RMSE	RAE	RRSR
BayesNet	26.70	8.80	37.83	46.88
Naïve Bayes	42.80	12.38	60.71	65.92

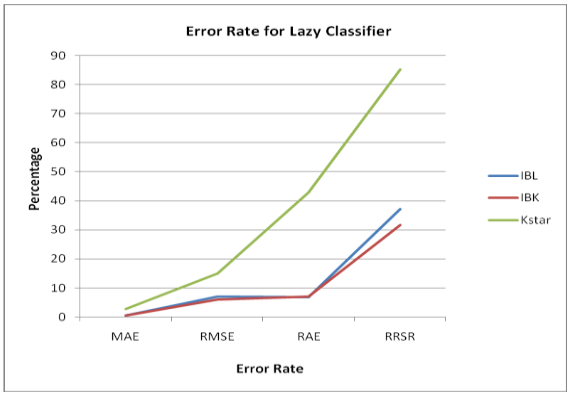


Fig 5: Error Rate of Lazy Classifier

From this graph, it is observed that IBL and Kstar algorithms attains highest error rate. Therefore, the IBK classification algorithm performs well because it contains least error rate when compared to IBL and Kstar algorithms.

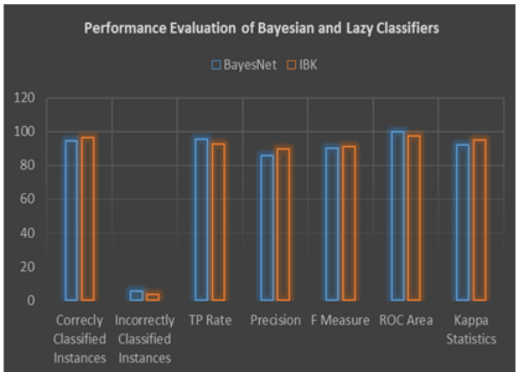


Fig 6: Accuracy measure of Bayesian and Lazy Classifiers

From the graph, it is observed that IBK algorithm performs better than BayesNet algorithms. Therefore the IBK classification algorithm performs well because it contains highest accuracy when compared to BayesNet.

TABLE I
ACCURACY MEASURE FOR BAYESIAN CLASSIFIER

Algorithm	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**TP Rate**	**Precision**	**F Measure**	**ROC Area**	**Kappa Statistics**
BayesNet	*E.* 94.71	*F.* 5.29	95.50	85.70	90.30	99.80	91.90
Naïve Bayes	*G.* 82.51	*H.* 17.49	49.80	79.30	61.20	98.80	72.90

TABLE 2
ACCURACY MEASURE FOR LAZY CLASSIFIER

Algorithm	Correctly Classified Instances (%) value	Incorrectly Classified Instances (%) value	TP Rate	Precision	F-Measure	ROC Area	Kappa Statistics
IBL	96.36	3.64	92.60	89.90	91.20	96.20	94.44
IBK	96.72	3.28	92.60	89.90	91.20	97.70	94.98
KStar	68.47	31.53	63.90	81.30	63.90	71.60	57.32

V. CONCLUSION

Data mining can be defined as the extraction of useful knowledge from large data repositories. In this paper, the classification algorithms namely Bayesian and Lazy classifier are used for classifying computer files which are stored in the computer. The Bayesian Algorithm includes two techniques namely Bayes Net, Naïve Bayes and the Lazy algorithms includes IBl (Instance Based Learning), IBK (K-Nearest Neighbour) and KStar techniques. By analysing the experimental results it is observed that the lazy classifier's IBK classification technique has yields better result than other techniques.

REFERENCES

[1] Abdullah H. Wahbeh, Mohammed Al-Kabi, "Comparative Assessment of the Performance of three WEKA text classifiers applied to Arabic Text".
[2] Anshul Goyal, Rajni Mehta, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms".
[3] S.Deepajothi, Dr.S.Selvarajan, "A Comparative Study of Classification Techniques On Adult Data Set"
[4] Ian H. Witten, Eibe Frank, "Data Mining Tools and Techniques practical Machine Learning".
[5] Kaushik H. Raviya, Biren Gajjar, "Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA".
[6] Lan Huang, David Milne, Eibe Frank, Ian H. Witten, "Learning a Concept-based Document Similarity Measure".
[7] Mahendra Tiwari, Manu Bhai Jha, Om Prakash Yadav, "Performance analysis of Data Mining algorithms in Weka".
[8] Mohd Fauzi bin Othman, Thomas Moh Shan Yau, "Comparison of Different Classification Techniques Using WEKA for Breast Cancer".
[9] Petra Kralj Novak, "Classification in WEKA".
[10] Mrs. Sayantani Ghosh, Mr. Sudipta Roy, Prof. Samir K. Bandyopadhyay, "A tutorial review on Text Mining Algorithms".
[11] Sunila Godara, Ritu Yadav,"Performance analysis of clustering algorithms for character recognition using Weka tool".
[12] Trilok Chand Sharma, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm".
[13] Dr. S.Vijayarani, S.Sudha,"Comparative Analysis of Classification Function Techniques for Heart Disease Prediction".
[14] Dr. S.Vijayarani, S. Sudha, "An Effective Classification Rule Technique for Heart Disease Prediction".
[15] B S Harish, D S Guru, S Manjunath, "Representation and Classification of Text Documents: A Brief Review".
[16] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan, "A Review of Machine Learning Algorithms for Text-Documents Classification".

BIOGRAPHY

Dr. S. Vijayarani has completed MCA, M.Phil and PhD in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy, security, bioinformatics and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. M. Muthulakshmi has completed M.Sc in Computer Science and Information Technology. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Mining, Text Mining and Semantic web mining.