iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** II **Month of publication:** February 2018

DOI:

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue II, February 2018- Available at www.ijraset.com

Comparison of Feature selection Methods and Algorithms

Ms. R. Durga Devi [1], Mrs. R. Merlin Packiam[2]
[1]Mphil Scholar,[2]Assist.Professor,Department of Computer Science, Cauvery College for Women, Bharathidhasan University, Trichy, Tamil Nadu(India)

Abstract: Feature selection is an main topic in data mining, particularly for high dimensional datasets. Feature selection (also known as subset selection) is a method normally used in machine learning, wherein subsets of the features presented from the data are selected for application of a learning algorithm. The most excellent subset contains the least number of dimensions that most supply to accuracy; we discard the remaining, unimportant dimensions. This is an important period of preprocessing and is one of two ways of avoiding the curse of dimensionality (the other is feature extraction). There are two approaches in Feature selection known as Forward selection and backward selection. Feature selection has been an dynamic research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to decide a subset of input variables by eliminating features with small or no predictive information. Feature selection methods can be decayed into three broad classes. One is Filter methods and another one is Wrapper method and the third one is Embedded method.. This paper provides the clear insight to different feature selection methods and algorithms reported in the literature and also compare all methods with each other. The experimental result shows that the feature selection algorithms provide better result for breast cancer data set.
Keywords: Filter model, Wrapper model, Clustering, Classification, Feature selection, Accuracy

I. INTRODUCTION

Difficulty of selecting some subset of a learning algorithms input variables leading which it should focus concentration, while ignoring the rest. Feature selection is the process of selecting the best feature between all the features because all the features are not helpful in constructing the clusters: various features may be discussed or irrelevant thus not contributing to the learning process. The top subset contains the least number of dimensions that most supply to accuracy; we remove the remaining, insignificant dimensions. This is an significant stage of preprocessing and is one of two ways of avoiding the curse of dimensionality (the additional is feature extraction). The main aim of feature selection is to decide a negligible feature subset from a problem domain while retaining a duly high accuracy in representing the unique features. In many real world problems Feature selection is a must due to the profusion of noisy, extraneous or misleading features. For example, by removing these factors, learning from data techniques can advantage. To be totally sure of the attribute selection, we would preferably have to test all the enumerations of attribute subsets, which is infeasible in most cases as it will result in 2n subsets of n attributes. Feature selection has been an dynamic research area in pattern recognition, statistics, and data mining communities.

The processing of accumulated data itself has become a big challenge for researchers in order to identify relevant and irrelevant features to improve the predictive accuracy and for this the number of data reduction techniques has been proposed so far. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance [1]. Feature selection is one of the important and frequently used techniques in data reduction or preprocessing for data mining. There are a number of advantages of feature selection includes it reduces the number of features, removes irrelevant, redundant, or noisy data, reduce the computational cost, speeding up a data mining algorithm and improve the classification accuracy [2].

Feature selection is a procedure that selects a subset of original features. The optimality of a feature subset is calculated by an evaluation criterion. The feature selection procedure consists of 4 essential steps, namely, subset evaluation, subset generation, stopping criterion, and result validation [3]. Subset generation is a discover process [4] that produces candidate feature subsets for evaluation based on an certain search strategy. Every candidate subset is evaluated and compared with the previous best one according to a convinced evaluation criterion. If the new subset twists out to be improved, it replaces the earlier best subset. The procedure of subset creation and evaluation is repeated awaiting a given stopping criterion is pleased. Finally, the selected best subset to be validated by domain experts or any other test and the selected best may be given as an input to any data mining task.

The feature selection methods generally confidential into three categories: the filter model [5, 6, 7], the wrapper model [8, 9, 10], and the embedded model [11, 12, 13]. The filter model relies on general character of the data to calculate and select feature subsets without connecting any mining algorithm. The wrapper model requires one prearranged mining algorithm and uses its presentation

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue II, February 2018- Available at www.ijraset.com

as the evaluation criterion. It seeks for features enhanced suited to the mining algorithm aiming to get improved mining presentation, but it also tends to be more computationally expensive than the filter model [14]. The embedded model attempts to take advantage of the two models by exploiting their special evaluation principle in different search stages. This paper is organized as follows. Section 1 gives the overview and introduction of feature selection, section 2 discusses the review of literatures and section 3 discusses the different feature selection methods used in this paper section 4 discusses the different feature selection algorithms used in this paper. The experimental results are shown and discussed in section 5 and finally the paper is concluded in section 6.

II. BACKGROUND STUDY

Feature selection is one of the active fields of research for decades in machine learning, data mining, genomic analysis [15], text mining [16], image retrieval [17], intrusion detection [18], etc. The paper [19] adopted an unbiased protocol to perform a fair comparison of frequently used multivariate and univariate gene selection techniques, in combination with a range of classifiers. In their conclusion they found that univariate and multivariate feature selection algorithms greatly improved the performance of cancer genes. Subset generation is a heuristic search in which each state specifies a candidate subset for evaluation in the search space. Two basic issues determine the nature of the subset generation process. First, successor generation decides the search starting point, which influences the search direction. To decide the search starting points at each state, forward, backward, compound, weighting, and random methods may be considered [20]. Second, search organization is responsible for the feature selection process with a specific strategy, such as sequential search, exponential search [22, 23] or random search [24]. A recently generated subset must be evaluated by confident evaluation criteria. So, many evaluation criteria have been proposed in the literature to determine the goodness of the candidate subset of the features. Base on their dependency on mining algorithms, evaluation criteria can be categorized into groups: independent and dependent criteria [21]. Independent criteria exploit the essential characteristics of the training data without involving any mining algorithms to evaluate the goodness of a feature set or feature. And dependent criteria involve predetermined mining algorithms for feature selection to select features based on the performance of the mining algorithm applied to the selected subset of features. Finally, to stop the selection process, stop criteria must be determined. Feature selection process stops at validation procedure. It is not the part of feature selection process, but feature selection method must be validate by carrying out different tests and comparisons with previously established results or comparison with the results of competing methods using artificial datasets, real world datasets, or both.

The relationship between the inductive learning method and feature selection algorithm infers a model. There are three general approaches for feature selection. First, the Filter Approach exploits the general characteristics of training data with independent of the mining algorithm. Second, the Wrapper Approach explores the relationship between relevance and optimal feature subset selection. It searches for an optimal feature subset adapted to the specific mining algorithm. And third, the Embedded Approach is done with a specific learning algorithm that performs feature selection in the process of training.

A. Filter Methods

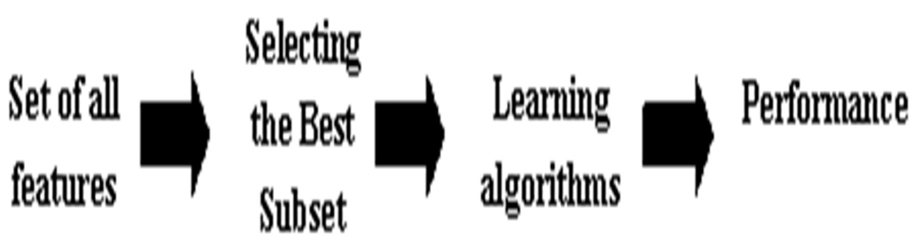

Fig 1. Filter method Performance

Filter methods are normally used as a preprocessing step. The selection of features is self-governing of any machine learning algorithms. In its place, features are selected on the base of their scores in different statistical tests for their correlation with the result variable. The correlation is a slanted term here. For basic guidance, you can refer to the following table for defining correlation co-efficients.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue II, February 2018- Available at www.ijraset.com

Table 1. correlation co-efficients

Feature\Response	Continuous	Categorical
Continuous	Pearson's Correlations	LDA
Categorical	Anova	Chi-Square

1) *Pearson's Correlation*: It is used as a calculate for quantifying linear dependence between two constant variables X and Y. Its value varies from -1 to +1. Pearson's correlation is given as:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

2) *LDA:* Linear Discriminant Analysis is used to find a linear combination of features that characterizes or separates two or more classes (or levels) of a categorical variable.
3) *ANOVA:* ANOVA stands for Analysis of variance. It is parallel to LDA excluding for the fact that it is operated using one or more definite independent features and one continuous needy feature. It provides a arithmetical test of whether the means of several groups are equal or not.
4) *Chi-Square*: It is a statistical test applied to the groups of categorical features to estimate the probability of correlation or association between them using their frequency distribution.

One thing that should be kept in mind is that filter methods do not confiscate multicollinearity. So, you must deal with multicollinearity of features as well before training models for your data.

B. *Wrapper Methods*

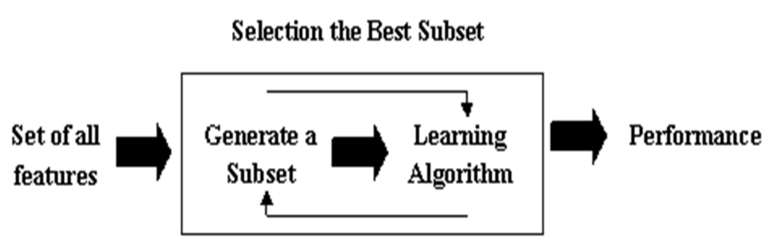

Fig 2. Wrapper method Performance

In wrapper methods, we attempt to use a subset of features and train a model using them. Based on the deductions that we illustrate from the previous model, we make a decision to add or remove features from your subset. The problem is basically reduced to a search problem. These methods are usually computationally very exclusive.

Some general examples of wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, etc.

1) *Forward Selection:* Forward selection is an iterative method in which we create with having no feature in the model. In every iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.
2) *Backward Elimination:* In backward elimination, we start with all the features and eliminate the least important feature at every iteration which improves the presentation of the model. We replicate this until no improvement is observed on elimination of features.
3) *Recursive Feature elimination:* It is a greedy optimization algorithm which aspires to find the best performing feature subset. It frequently creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are tired. It then ranks the features based on the order of their elimination. One of the top ways for applying feature selection with wrapper methods is to use Boruta wrap up that discovers the meaning of a feature by creating shadow features

4) *It works in the following steps*

a) Firstly, it adds randomness to the given data set by creating scuffled copies of all features (which are called shadow features).

b) Next, it trains a random forest classifier on the completed data set and applies a feature import calculate (the default is Mean Decrease Accuracy) to assess the importance of each feature where higher means more important.

c) At each iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z-score than the maximum Z-score of its shadow features) and continually removes features which are deemed highly insignificant.

d) Finally, the algorithm ends too when all features get confirmed or rejected or it reaches a particular maximum of random forest runs.

C. *Embedded Methods*

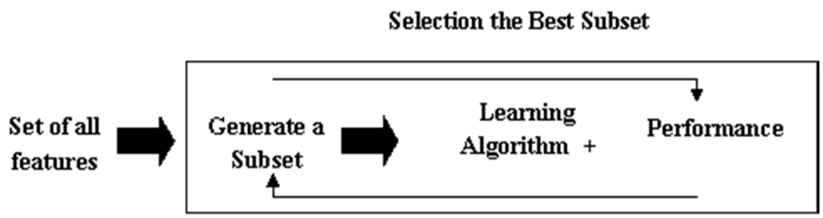

Fig 3. Embedded method Performance

Embedded methods separate the qualities' of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods. Several of the most popular examples of these methods are RIDGE and LASSO regression which have inherent penalization functions to decrease over fitting.

D. *Feature Selection Algorithms*

This part of this paper briefly introduces the feature selection algorithms that has been discovered and reported in the literatures. The feature selection algorithms are classified into three categories such as filter model, wrapper model and embedded model according to the computational models. The filter model relies on the common characteristics of data and evaluates features without connecting any learning algorithm. The wrapper model requires having a prearranged learning algorithm and uses its performance as evaluation criterion to choice features. The embedded model slot in variable selection as a part of the training process, and feature relevance is obtained logically from the objective of the learning model.

E. *Relief (RF)*

Relief F [26] is a supervised multivariate feature selection algorithm of the filter model which is the addition of Relief is a univariate model. Assuming that p instances are randomly sampled from data, the evaluation criterion for managing multiclass problems is of the form

$$SC_R\,(f_i) \;=\; \frac{1}{P}\;.\;\sum_{t=1}^{p}\left\{\begin{matrix} -\frac{1}{m_{xt}}\;\sum_{x_{j\in NM(x_t)}} d\left(f_{t,i}-f_{i,j}\right)\;+ \\ \sum_{y\neq yx_t}\frac{1}{mx_{t,y}}\;\frac{P(y)}{1-P(yx_t)}\;\sum_{x_i\,\in NM(x_t,y)} d\,\left(f_{t,i}-f_{i,j}\right)\end{matrix}\right\}$$

where yx_t is the class label of the instance x_t and P(y) is the probability of an example being from the class y. NH(x) or NM(x, y) indicates a set of next points to x with the same class of x, or a different class (the class y), correspondingly. mx_t and mx_t,y are the sizes of the sets NH(x_t) and NM(x_t, y), respectively. Typically, the size of both NH(x) and NM(x, y); ¥ y ≠ yx_t , is set to a pre-specified constant k.

F. *Information Gain (IG)*

Information Gain [6] is supervised univariate feature selection algorithm of the filter model which is a measure of dependence between the feature and the class label. It is one of the most powerful feature selection techniques and it is easy to compute and simple to interpret. Information Gain (IG) of a feature X and the class labels Y is calculated as

IG(X,Y) = H(X) – H(X|Y)

Entropy (H) is a measure of the uncertainty associated with a random variable. H(X) and H(X/Y) is the entropy of X and the entropy of X after observing Y, respectively.

H(X) = $-\sum_i P(x_i) log_2 (P(x_i))$

The maximum value of information gain is 1. A feature with a high information gain is applicable. Information gain is evaluated separately for each feature and the features with the top-k values are selected as the relevant features. This feature selection algorithm does not reduce redundant features.

H(X|Y) = $-\sum_j P(y_i) \sum_i P(x_i|y_j) log_2(P(x_i|y_j))$

G. Gain Ratio

The Gain Ratio is the non-symmetrical calculate that is introduced to recompense for the bias of the IG [31]. GR is given by

$$GR = \frac{IG}{H(X)}$$

As the above equation presents, when the variable Y has to be predicted, the Information Gain has to regularized by dividing by the entropy of X, and vice versa. Due to this normalization, the Gain Ratio values always fall in the range [0, 1]. A value of Gain Ratio = 1 indicates that the knowledge of X completely predicts Y, and Gain Ratio = 0 means that there is no relation between Y and X. The Gain Ratio works well variables with fewer values where as the Information Gain works well variables with larger values.

H. Gini Index (GI)

Gini index [16] is supervised multivariate feature selection algorithm of the filter model to measure for quantifying a feature's ability to distinguish between classes. Given C classes, Gini Index of a feature *f* can be calculated as Gini Index can take the maximum value of 0.5 for a binary classification. The more relevant features have smaller Gini index values. Gini Index of each feature is calculated independently and the top k features with the smallest Gini index are selected. Like Information gain, it also not eliminates redundant features.

$$Gii\ Index(f) = 1 - \sum_{i=1}^{C} [p(i|f)]^2$$

Random Forest inhabited by Leo [4] is a group of original classification or regression trees finished from the random selection of samples of the training data. Random features are selected using the introduction process. Prediction is made by aggregating the predictions of the collection. Random Forest generally confirms a significant performance improvement as compared to single tree classifier C4.5

I. Comparison of Feature Selection Algorithms

Comparison of feature selection techniques is shown in the following table. For each feature selection type, we highlight a set of characteristics which can guide the choice for a technique suited to the goals and resources of practitioners in the field

MODEL SEARCH		ADVANTAGES	DISADVANTAGES	EXAMPLES
FILTER	UNIVARIATE	Fast, Scalable, Independent of the classifier	Ignores feature dependencies, Ignores interaction with the classifier	X^2 , Euclidian distance, t-test, Information gain
	MULTIVARIATE	Models feature dependencies, Independent of the classifier, Better computational complexity than wrapper methods	Slower than univariate techniques, Less scalable than univariate techniques, Ignores interaction with the classifier	Correlation-based feature selection(CFS), Markov blanket filter (MBF), Fast correlation-based feature selection (FCBF)

WRAPPER	DETERMINISTIC	Simple, Interacts with the classifier, Models feature dependencies, Less computationally intensive than randomized methods	Risk of over fitting, More prone than randomized algorithms to getting stuck in a local optimum (greedy search), Classifier dependent selection	Sequential forward selection (SFS), Sequential backward elimination (SBE), Plus L Minus R, Beam search
	RANDOMIZED	Less prone to local optima, Interacts with the classifier, Models feature dependencies	Computationally intensive, Classifier dependent selection, Higher risk of over fitting than deterministic algorithms	Simulated annealing, Randomized hill climbing, Genetic algorithms, Estimation of distribution algorithms
EMBEDDED		Interacts with the classifier, Better computational complexity than wrapper methods, Models feature dependencies	Classifier dependent selection	Decision trees, Weighted naïve Bayes, Feature selection using the weight vector of SVM

III. RELATED WORK

S.NO	AUTHOR	TITLE	METHOD OR ALGORITHM USED	DATASET	ACHIEVEMENT	DRAWBACK
1	Huan Liu and Lei Yu	Toward Integrating Feature Selection Algorithms for Classification and Clustering	Filter Method	Real-world data sets	A unifying platform is proposed as an intermediate step	A preprocessing step in very large databases collected from Internet
2	M. Dash and H. Liu	Feature selection for classification	Relief Algorithm	Various dataset	Overview of many existing methods from the 1970's to the present	Test different combinations that previously exist
3	M.Dash, K.Choi and P.Scheuermann	Feature Selection for Clustering-A Filter Solution	Filter Method	Synthetic, Benchmark and real datasets	To evaluate feature subsets and choose the best subset for clustering by considering their effect on the underlying clusters	Lack of unanimous agreement in evaluating the clusters
4	P.Langley	Selection of relevant features in machine learning	Wrapper Method	Real World dataset	Feature selection can improve the behavior of induction algorithms in a variety of situations	Former's computational cost,which results from calling the induction

						algorithm for each feature set considered
5	H.Liu and R.Setiono	A probabilistic approach to feature selection- a filter solution	Filter or Wrapper method	Artificial and real world dataset	Proves the effectiveness and scalability	It works hard to find the optimal solution
6	E.Xing, M.Jordan and R.Karp	Feature selection for high-dimensional genomic microarray data	Filter and Wrapper method	Gene Expression dataset	Performed significantly better in the reduced feature space than in the full feature space	More computation time
7	Sanmay Das	Filters,Wrappers and a Boosting-based Hybrid for Feature Selection	Filter,Wrapper method and hybrid algorithm	Mushroom,Chess,Ads,DNA,Vote and Lymphography dataset	Improves the performance of the learning algorithm, filter method scale much better to large datasets	Size of the feature set to select needs to be pre-specified
8	R.Kohavi and G.H.John	Wrappers for feature subset selection	Wrapper,filter method,FOCUS and Relief algorithm	Artificial and real dataset	Identifying the relevant features in a dataset and giving only that subset to the learning algorithm	Large amount of CPU time required
9	Y.Kim, W.Street and F.Menczer	Feature Selection in Unsupervised Learning via Evolutionary Search	Evolutionary local selection algorithm	Synthetic and real dataset	Improved understandability,scalability and possibly,accuracy of the resulting models	More Computational time
10	Yvan saey, Inaki inza and pedro Larranaga	A review of feature selection techniques in bioinformatics	Filter,Wrapper and Embedded method	Various dataset	Better computational complexity	Large input dimensionality and the small sample sizes

IV. DISCUSSION

Feature selection is an important part of most of the data processing applications including data mining, machine learning and computational intelligence. It helps in removing the irrelevant features and redundant information which affects the accuracy of the model. This paper presents a survey about types of feature selection techniques and processes as discussed by various authors. Performance of different algorithms varies according to the data collection and requirements. However, all the discussed classifiers can only predict the class of unknown document; they do not provide degree of relevance of a particular document to a particular class. Also the data needs to be certain, precise and accurate. These difficulties can be overcome by using soft computing methodologies that aim to exploit the tolerance for imprecision, uncertainty, partial truth and approximation.

V. CONCLUSION

Feature selection has been a statement as ever green research topic with practical implication in many areas such as statistics, pattern recognition, machine learning, and data mining, web mining, text mining, image processing, and gene microarrays analysis. These feature selection algorithms are very well helpful to construct simpler and more comprehensible models, improving data mining tasks presentation and accuracy, and helps to understand main data. In this element, it is shown the different activities of the algorithms to different data particularities and thus the risk in relying in a single algorithm. This point in the direction of using new hybrid algorithms or combinations thereof for a more dependable assessment of feature relevance. As future behavior, this work can be extended in many ways to take up richer evaluations such as considering features powerfully correlated with the class or with one another, noise in the data sets, other kinds of data (e.g., continuous data), missing values, and the use of mutual evaluation measures.

REFERENCES

[1] Han & Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, 2006.
[2] Huan Liu and Lei Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 4, Pages: 491 - 502, 2005
[3] M. Dash and H. Liu. Feature selection for classification. Intelligent Data Analysis: An International Journal, 1(3):131–156, 1997.
[4] P. Langley. Selection of relevant features in machine learning. In Proceedings of the AAAI Fall Symposium on Relevance, pages 140–144, 1994.
[5] M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature selection for clustering – a filter solution. In Proceedings of the Second International Conference on Data Mining, pages 115–122, 2002.
[6] M.A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 359–366, 2000.
[7] H. Liu and R. Setiono. A probabilistic approach to feature selection - a filter solution. In Proceedings of the Thirteenth International Conference on Machine Learning, pages 319– 327, 1996.
[8] R. Caruana and D. Freitag. Greedy attribute selection. In Proceedings of the Eleventh International Conference on Machine Learning, pages 28–36, 1994.
[9] J. G. Dy and C. E. Brodley. Feature subset selection and order identification for unsupervised learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 247–254, 2000.
[10] Y. Kim, W. Street, and F. Menczer. Feature selection for unsupervised learning via evolutionary search. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 365–369, 2000.
[11] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. In Proceedings of the Eighteenth International Conference on Machine Learning, pages 74–81, 2001.
[12] A. Y. Ng. On feature selection: learning with exponentially many irrelevant features as training examples. In Proceedings of the Fifteenth International Conference on Machine Learning, pages 404–412, 1998.
[13] E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. In Proceedings of the Eighteenth International Conference on Machine Learning, pages 601–608, 2001.
[14] R. Kohavi and G.H. John. Wrappers for feature subset selection. Artificial Intelligence, 97(1-2):273–324, 1997.
[15] Inaki Inza, Pedro Larranaga, Rosa Blanco, and Antonio J. Cerrolaza. Filter versus wrapper gene selection approaches in dna microarray domains. Arti_cial Intelligence in Medicine, 31:91-103, 2004.
[16] George Forman. An extensive empirical study of feature selection metrics for text classi_cation. Journal of Machine Learning Research, 3:1289-1305, 2003.
[17] R. Gonzalez and R. Woods. Digital Image Processing. Addison-Wesley, 2nd edition, 1993.
[18] W. Lee, S. J. Stolfo, and K. W. Mok. Adaptive intrusion detection: A data mining approach. AI Review, 14(6):533-567, 2000.
[19] A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets , Carmen Lai, Marcel JT Reinders, Laura J van't Veer and Lodewyk FA Wessels, BMC Bioinformatics 2006, 7:235 doi:10.1186/1471-2105-7-235
[20] Clustering Time Series Data Stream – A Literature Survey, V.Kavitha, Computer Science Department, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, Tamilnadu, India. M. Punithavalli, Sri Ramakrishna College of Arts & Science for Women, Coimbatore ,Tamil Nadu, India., Vol 8, Apr, 2010.