Filter Based Approach for Genomic Feature Set Selection (FBA-GFS)

V.Bhuvaneswari[1] and K.Poongodi[2]

[1]Assistant Professor, Department of Computer Application, Bharathiar University, Coimbatore, India

bhuvanes_v@yahoo.com

[1]M.Phil Research Scholar,, Department of Computer Application, Bharathiar University, Coimbatore, India

poongodikmca@gmail.com

Abstract

Feature selection is an effective method used in text categorization for sorting a set of documents into certain number of predefined categories. It is an important method for improving the efficiency and accuracy of text categorization algorithms by removing irredundant terms from the corpus. Genome contains the total amount of genetic information in the chromosomes of an organism, including its genes and DNA sequences. In this paper a Clustering technique called Hierarchical Techniques is used to categories the Features from the Genome documents. A framework is proposed for Genomic Feature set Selection. A Filter based Feature Selection Method like x^2 *statistics, CHIR statistics are used to select the Feature set. The Selected Feature set is verified by using F-measure and it is biologically validated for Biological relevance using the BLAST tool.*

Keywords:

Feature selection, Clustering, Genome

1. INTRODUCTION

Data Mining is also known as Knowledge Discovery in Database (KDD). Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, potentially useful and ultimately understandable patterns in data. It is the process of extracting novel information and knowledge from large databases. This process consists of many interacting stages performing specific data manipulation and transformation operations with an information flow from one stage onto the next [4].

Document clustering is an automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other clusters.Hierarchical techniques produce a nested sequence of partitions, with

DOI : 10.5121/ijcsea.2012.2104

a single, all inclusive cluster at the top and singleton clusters of individual points at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level (or splitting a cluster from the next higher level). The result of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendogram. This tree graphically displays the merging process and the intermediate clusters. The dendogram at the right shows how four points can be merged into a single cluster. For document clustering, this dendogram provides a taxonomy, or hierarchical index.

The task of feature selection is generally divided into two aspects eliminating *irrelevant* features and *redundant* ones. Irrelevant features usually disturb the learner and degrade the accuracy, while redundant features add to computational cost without bringing in new information. Feature Selection techniques can be divided into three main approaches as filter, wrapper and embedded approaches. In embedded approaches, where feature selection is part of the classification algorithm, i.e. decision tree. In Filter approaches, the features are selected before the classification algorithm and in the Wrapper approaches the classification algorithm is used as a black box to find the best subset of attributes [14].

In filter based approach, the features are selected according to data intrinsic values, such as information dependency or consistency measures. In bioinformatics, datasets are often very large. Therefore, the filter approach is mostly used to select the features. The advantage of this approach is that we can use any classifier to evaluate the accuracy of the test set. In the wrapper approach, we have to use a different classifier for the test set and training set, results may be negatively affected. In the proposed work, the Filter based Feature Selection approach is studied and analyzed for selecting features from genomic databases for clustering.

The paper is organized as follows. Section 2 provides the literature study of the various Feature selection methods for Bio- logical database. Section 3 explores the methodology for Genomic Feature set Selection (GFS). In section 4 the implemented results are verified and validated. The final section draws the conclusion of the paper.

2. REVIEW OF LITERATURE

Feature selection methods have been successfully applied to text categorization but seldom applied to text clustering due to the unavailability of class label information [13].The goal of feature selection for unsupervised learning is to find the smallest feature subset that best uncovers clusters form data according to the preferred criterion [1].Feature selection techniques do not alter the original representation of the variables, but merely select a subset of them [14].

In[3] , Bassam Al-Salemi (2011) Used Feature Selection techniques such as Mutual Information (MI), Chi-Square Statistic (CHI), Information Gain (IG), GSS Coefficient (GSS) and Odds Ratio (OR) to reduce the dimensionality of feature space by eliminating the features that are considered irrelevant for the category [3].

Feature selection can help improve the efficiency of training. In information retrieval, especially in web search, usually the data size is very large and thus training of ranking models is computationally costly[15]. In [2] , Balaji Krishnapuram et al., (2004) have adopted a Bayesian approach for both an optimal nonlinear classifier and a subset of predictor variables (or features)

that are most relevant to the classification task. The Bayesian approach uses heavy-tailed priors to promote sparsity in the utilization of both basis functions and features.
Feng Tan et al.,(2006) have proposed a hybrid approach to combine useful outcomes from different feature selection methods through a genetic algorithm. Their experimental results demonstrate that the approach can achieve better classification accuracy with a smaller gene subset than each individual feature selection algorithm [6].

In[9],Gouchol Pok etal., (2010) have done a effective feature selection framework suitable for two-dimensional microarray data. The correlation-based on parametric approach allows compact representation of class-specific properties with a small number of genes. In[11],Lin Sun et al(2011) have discussed a new rough entropy to measure the roughness of knowledge and its properties were proposed in decision information system, and concluded that rough entropy decreased monotonously as the information granularities became finer was obtained.

In[15],Yiming Yang et al., (1992) have reported a controlled study with statistical significance test on five text categorization methods : Support Vector Machine(SVM), a K-Nearest Neighbor (KNN) Classifier,a Neural Network(NNet) approach, the Linear Least Square Fit(LLSF) mapping and a Naïve-Bayes(NB) Classifier.

In[8], George H.John et al., (1994) have described a method for feature subset selection using cross-validation that is applicable to any induction algorithm. Discussed the experiments conducted with ID3 and C4.5 on artificial and real datasets. Pabitra Mitra et al., (2002) have described an unsupervised feature selection algorithm suitable for data sets, large in both dimension and size. The method is based on measuring similarity between features whereby redundancy therein is removed [12]. Spectral feature selection identifies relevant features by measuring their capability of preserving sample similarity [21].

3. METHODOLOGY

The Genomic sequence data are stored in public databases like NCBI, Uniport in various formats. The information related to sequence is represented as attributes. Finding the important attributes for comparing the genomic sequence data based on annotation, becomes the challenging task. Feature selection methods can be used to analyze and study the best features used for representing sequence information for association and clustering of documents using supervised and techniques. The proposed framework given in Figure 1 is used for Genomic Feature set Selection (GFS).

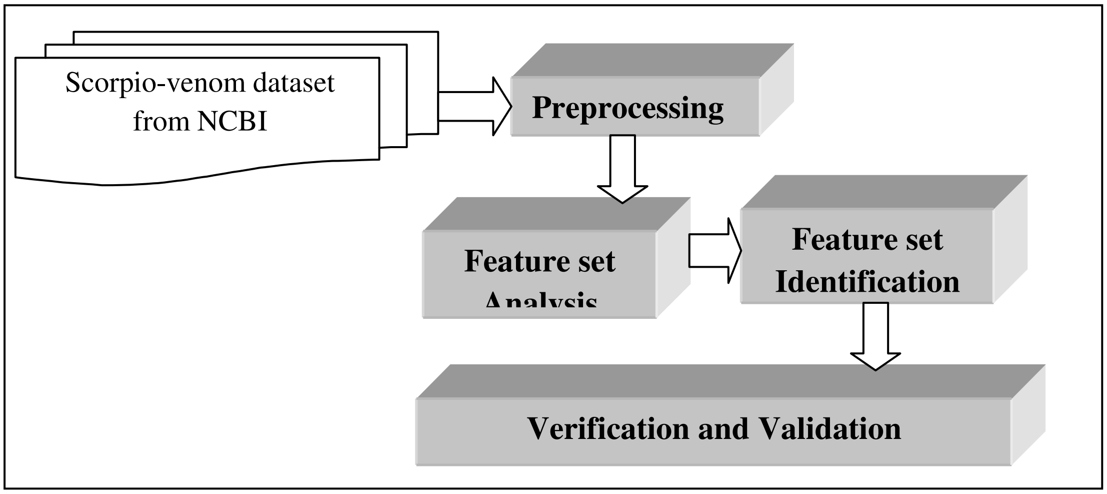

Scorpio-venom dataset from NCBI
Preprocessing
Feature set Analysis
Feature set Identification
Verification and Validation

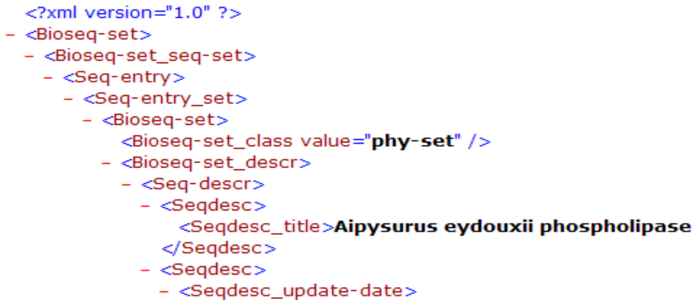
Figure 1. Genomic Feature set Selection (GFS).

The proposed framework consists of four phases which includes preprocessing, Feature set Analysis, Feature set Identification and verification and validation phase.

3.1 Dataset

The scorpio-venom dataset is used in the proposed work. It is downloaded from the NCBI which is in the XML format. The NCBI dataset is the integrated, text-based search and retrieval system used at the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxanomy, and others. The scorpio -venom dataset contains sequence information in XML format.

The genomic data in XML format has more than 3500 tags to represent the functional descriptions about the sequences like accession no, taxonomy, organism, lineage, sequence title, sequence descriptions, alternate name,gene name, author details, and identifiers related to other databases like GO, KEGG, PUBMED. The scorpio-venom specis in XML Format is shown in the Figure 2.The dataset consist of 107 documents in XML format for the scorpio -venom species.



```
  <?xml version="1.0" ?>
- <Bioseq-set>
  - <Bioseq-set_seq-set>
    - <Seq-entry>
      - <Seq-entry_set>
        - <Bioseq-set>
            <Bioseq-set_class value="phy-set" />
          - <Bioseq-set_descr>
            - <Seq-descr>
              - <Seqdesc>
                  <Seqdesc_title>Aipysurus eydouxii phospholipase
                </Seqdesc>
              - <Seqdesc>
                - <Seqdesc_update-date>
```

Figure 2.scorpio-venomspecis in XML Format

3.2 Preprocessing

The preprocessing phase of the proposed work consists of two main process which includes extraction of keyword from XML document and construction of text matrix for feature selection.The dataset in XML format is stored into DB2 data base, where each XML transaction data file is uniquely identified with transaction identifiers. The XML documents are stored as xml files in the db2 database and the features are extracted from the DB2 database using XQuery features which offers easy access to the entire XML data or part of XML data without converting to any conventional format. The key filter interface developed in DB2 XQuery provides the way to extract the necessary fields like gene name and associated go terms, keywords in every biological XML file. The 358 keywords are extracted from original Dataset which contain 511 keywords.

The document-term matrix contains rows corresponding to the documents and columns corresponding to the terms. After extraction of the keywords from the XML document, the term Matrix is constructed which is the input for further processing. The term matrix is represented as binary encoded format. The values are encoded as zero and one, in the presence of keyword one is entered and in absence of a keyword zero is entered into the corresponding place, since in the xml database all the elements have the same domain values so no string edit measures are needed. Table 1 shows the snapshot of the binary coded term matrix for the dataset chosen.

Table 1. Binary coded Term Matrix

Keywords/ Documents	acidic phospholipase A2	acidic phospholipase A2 precursor	ammodytes	………	Viridovipera stejnegeri
d1	1	0	1	………	1
d2	0	1	1	………	0
:				………	
d107	0	0	1	………	0

3.3 Feature Set Analysis

The phase 2 consists of two main processes which includes clustering of the documents to assign class label and analyzing the features using Filter based feature selection approaches using supervised methods like $\boldsymbol{x}^2$ statistics, CHIR statistics . Document clustering is the act of collecting similar documents into bins, where similarity is some function on a document. The term matrix constructed is given as input for the clustering phase. The documents are initially clustered for analyzing the features using hierarchical clustering algorithm. The proposed work we have considered 107 documents with 358 extracted keywords. On clustering the 107 documents 30 clusters are generated. From the generated cluster it is found that single document is found in many clusters and maximum documents are found in 9 clusters. So we have taken the cluster which contains highest number of documents to analyze the feature attributes and find the term relevance using filter based approach.

Feature Analysis Based On x^2 Statistics

The $\boldsymbol{x}^2$ Statistics can be used to measure the independence between the keyword and the category[14]. This can be done by comparing observed frequency in the 2-way contingency table with the expected frequency when they are assumed to be independent. For the $\boldsymbol{x}^2$ keyword-category independency, we consider the null hypothesis and alternative hypothesis. The null hypothesis is that the keyword and category are independent. On the other hand, the alternative hypothesis is that the keyword and category are not independent. To test the null hypothesis, we compare the observed frequency with the expected frequency calculated under the assumption that the null hypothesis is true. The expected frequency E(i,j) can be calculated as :

$$E(i,j) = \frac{\sum_{a\in\{\omega,-\omega\}} O(i,j) \sum_{b\in\{c,-c\}} O(i,j)}{n} \quad \dots eq\ (1)$$

The $\boldsymbol{x}^2$ statistics value is defined as:

$$x^2{}_{w,c} = \sum_{i\in\{\omega,-\omega\}} \sum_{j\in\{c.-c\}} \frac{\left(O(i,j) - E(i,j)\right)^2}{E(i,j)} \quad \dots eq(2)$$

The degrees of freedom for the x^2 Statistics is calculated by using the equation 3.

$$DF = (r-1) * (c-1) \qquad \dots \dots \dots eq(3)$$

Here, r and c are the number of row and column. In the fifth step, Look up the x^2 Critical value and Conclude the result **.**Looking up the x^2 distribution table, if the critical value is much smaller than the x^2 Statistics values, then the null hypothesis is rejected. This can be explained that it is significant and there is some dependency between the keyword and the category. The statistics e relationship is analyzed between the keyword and category for the 9 Clusters which contain maximum number of documents. Among 358 keywords, 156 keywords are found to be dependent to the category.

Feature Analysis Based On $\boldsymbol{CHIR}$ Statistics $(R_{\omega,c})$

CHIR is a supervised learning algorithm based on $\boldsymbol{x^2}$statistics, which determines the dependency between a keyword and a category and also the type of dependency [13]. Type of dependency indicates whether the feature is a positive or negative dependency for the category. There are two steps to evaluate the dependency of a keyword *w*, to category c. First step to build a table of the Observed Frequency and the Expected Frequency. Second step to calculate the $R_{\omega,c}$ by using the equation 4.

$$R_{w,c} = \frac{o_{w,c}}{E_{w,c}} \qquad \dots\dots\dots eq(4)$$

If there is no dependency between the term w and the category c, then the value of $R_{\omega,c}$, is close to 1. If there is a positive dependency then the observed frequency is larger than the expected frequency, hence value of $R_{\omega,c}$,is larger than 1 and when there is a negative dependency $R_{\omega,c}$, is smaller than 1. By calculating the $R_{\omega,c}$, it resulted all the 358 keywords are positively dependent to atleast one cluster because the values are greater than one.

3.4 Feature Set Identification

The Third phase is the Feature set Identification Phase. In this phase the Features are Ranked based on $x^2_{\max}$, x^2_{avg} and rx^2 Statistics. The highly ranked Features are used for analyzing the term relevance. The Feature sets are identified from ranking are Clustered with respect to documents.

Ranking Based On $\boldsymbol{x^2_{\max}}$, $\boldsymbol{x^2_{\text{avg}}}$ And $\boldsymbol{rx^2}$ Statistics

The best keyword are selected for finding the keyword-goodness .The Supervised feature selection method uses $\boldsymbol{x}^2_{\max}$ or $\boldsymbol{x}^2_{\text{avg}}$ to select the best keywords from *m* categories. The $\boldsymbol{x^2}_{\mathbf{max}}$ and $\boldsymbol{x}^2_{\text{avg}}$ are calculated by using the equation 8 and equation 9.

$$\boldsymbol{x}^2_{\max}(\omega) = \max_{\mathrm{j}}\left\{\boldsymbol{x}^{\,2}_{\omega,\mathrm{c_j}}\right\} \qquad \dots eq(8)$$

$$\boldsymbol{x}^2_{\text{avg}}(\omega) = \sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{p(c_j)}\boldsymbol{x}^{\,2}_{\omega,\mathrm{c_j}} \qquad \dots eq(9)$$

Here, $\mathrm{p(c_j)}$ is the probability of the documents to be in the category $\mathrm{c_j}$,then keyword whose keyword-goodness measure is lower than a certain threshold would be removed from the feature space. In other words, x^2 selects terms having strong dependency on categories. The supervised feature selection method CHIR uses $\boldsymbol{rx^2}$ to measure the Keyword -goodness and prove that $\boldsymbol{rx^2}$ values represent only the positive keyword-category dependency. The following are the steps to select the *n* keywords. There are three steps to calculate the rx^2 Statistic value. In the first step the $\boldsymbol{rx^2}$ Statistic value is calculated by using the equation 10.

$$\boldsymbol{rx^{\,2}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} \boldsymbol{p(\ R_{\omega,c_j})\, x^2}_{\boldsymbol{\omega c_j,}} \quad \boldsymbol{w]ith} \quad \boldsymbol{R_{\omega,c_j}} > 1 \quad \dots \boldsymbol{eq(10)}$$

Here, $p(R_{\omega,c_j})$ is the weight of $x^2{}_{\omega c_{j,}}$ in the corpus. In terms of $R_{\omega c_{j,}}$, $p(R_{\omega,c_j})$ is defined as

$$\boldsymbol{p(R_{\omega,c_j})} = \frac{\boldsymbol{R_{\omega,c_j}}}{\sum_{j=1}^{m} \boldsymbol{R_{\omega,c_j}}} \qquad \boldsymbol{wit]h} \;\; \boldsymbol{R_{\omega,c_j}} > 1 \;\; \dots \boldsymbol{eq(11)}$$

In the second step the keywords are sorted in descending order of their rx^2 Statistic. In the third step the top n keywords from the list are selected. The largest values of $\boldsymbol{rx^2}$ indicates that the keyword ω is more relevant to the category c. Keyword with top $\boldsymbol{rx^2}$ values are chosen as features. Let us assume the threshold value to remove the redundant and irrelevant features. Minimum threshold value 10 is taken 156 keywords are retrieved. For an average 50 is taken as threshold value 77 keywords are retrieved. For Maximum 85 is taken as threshold value 58 keywords are retrieved. Among 358 keywords 58 keywords are ranked high and these keywords are said to be more relevant and strongly dependent to the category.

Table 2. Relevant Keywords Retrieved

Threshold Value	**No. of Keywords Retrieved**
10	156
50	77
85	58

Clustering Based On Selected Feature Set

The documents are clustered based on the new feature set with {156, 77, 58} keywords. On clustering the 107 documents with respect to these feature set, 30 clusters are generated for each. On clustering the 107 documents with respect to 156 keywords13 clusters are obtained with more than one documents.

Among the 107 documents , it is founded that 12 documents are empty. Remaining 95 documents are judged to be of 156 keywords in entire hierarchy. On clustering the 107 documents for feature set with 77 keywords we obtained 6 clusters which contain more than one document. Among the 107 documents, it is founded that 82 documents are empty. Remaining 25 documents are judged to be of 77 keywords in entire hierarchy.

On clustering the 107 documents for feature set with 58 keywords we obtained 2 clusters which contain more than one document. Among the 107 documents, it is founded that 12 documents are judged to be of 58 keywords in entire hierarchy. Remaining documents are found to be empty.

3.5 Verification and Validation

The fourth phase is the verification and validation phase. The clusters are generated using the features set are evaluated by using the validation metrics. F-measure used to validate the clusters of Original Feature set, Feature sets with 156, 77 and 58 keywords. The clusters are validated for biological relevance by our experimental results is compared with existing BLAST tool.

3.5.1 Verifying the Terms Using F-measure

In our proposed work the keywords are verified using the F-Measure. F-Measure which is the harmonic mean of the precision and recall. The formula for the corresponding Precision, Recall and F-measure is shown in the eq.12, eq.13, and eq.14 respectively. For any Topic T and cluster X: N1: Number of documents judged to be of topic T in cluster X.N_2: Number of documents in cluster X.N_3: Number of documents to be judged to be of Topic T in entire hierarchy.

$$Precision = \frac{N_1}{N_2} \qquad \ldots eq(12)$$

$$Recall = \frac{N_1}{N_3} \qquad \ldots eq(13)$$

$$F - measure = \frac{2 * (Precision * \ Recall)}{(Precision + Recall)} \quad \ldots eq(14)$$

Verification of Original Feature Set

On clustering the 107 documents for original feature set with 358 keywords we obtain 9 clusters which contain more than one document. Among 9 clusters, 5 clusters {c3, c8, c11, c12, c24} contain maximum number of documents. It is found that 98 documents are found to be judged to be of Topic T in entire hierarchy. Remaining 9 documents are found to be empty. The clusters of feature set with 358 keywords are evaluated with three measures (Precision, Recall and F-measure) and the obtained result is given in Table 3.

Table 3.Evaluation Metrics for the Original Dataset

Cluster (%)	Precision (%)	Recall (%)	F-measure (%)
c3	85	50	63
c8	80	45	58
c11	83	25	38
c12	77	20	32
c24	79	20	32

Verification of Feature set with 156 keywords

On clustering the 107 documents for feature set with 156 keywords we obtained 13 clusters {c1, c2, c3, c4, c5, c6, c7, c8, c9, c11, c12, c14, c30} which contain more than one document. Among 13 clusters, 4 clusters {c1, c4, c5, c9} contain maximum number of documents. It is founded that 95 documents are found to be judged to be of Topic T in entire hierarchy. Remaining 12 documents are found to be empty. The clusters of feature set with 156 keywords are evaluated with three measures (Precision, Recall and F-measure) and the obtained result is given in Table 4.

Table 4.Evaluation Metrics for the Feature set with 156 keywords.

Cluster(%)	Precision(%)	Recall(%)	F-measure(%)
c1	80	60	69
c4	83	43	57
c5	42	42	57
c9	87	45	59

Verification of Feature set with 77 keywords

On clustering the 107 documents for feature set with 77 keywords we obtained 6 clusters {c1, c3, c5, c10, c11, c30} which contain more than one document. Among 6 clusters, 2 clusters {c3, c11} contain maximum number of documents. It is founded that 25 documents are found to be judged to be of Topic T in entire hierarchy. Remaining documents are found to be empty. The clusters of feature set with 77 keywords are evaluated with three measures (Precision, Recall and F-measure) and the obtained result is given in Table 5.

Table 5.Evaluation Metrics for the Feature set with 77 keywords.

Cluster(%)	Precision(%)	Recall(%)	F-measure(%)
C3	82	75	78
C11	83	78	80

Verification of Feature set with 58 keywords

On clustering the 107 documents for feature set with 58 keywords we obtained 2 clusters {c1, c11} which contain more than one document. Among 2 clusters, clusters {c11} contain maximum number of documents. It is founded that 12 documents are found to be judged to be of Topic T in entire hierarchy. Remaining documents are found to be empty.

The clusters of feature set with 77 keywords are evaluated with three measures (Precision, Recall and F-measure) and the obtained result is given in Table 6.

Table 6.Evaluation Metrics for the Feature set with 58 keywords.

Cluster (%)	Precision (%)	Recall (%)	F-measure (%)
C11	84	95	89

Analyzing the Feature sets

In this section, we have compared the result obtained above. Using the Filter based Feature selection methods the feature set extracted from scorpio-venom dataset are evaluated using precision, recall and F-measure. The cross matrix for metrics are displayed in the Table 7.

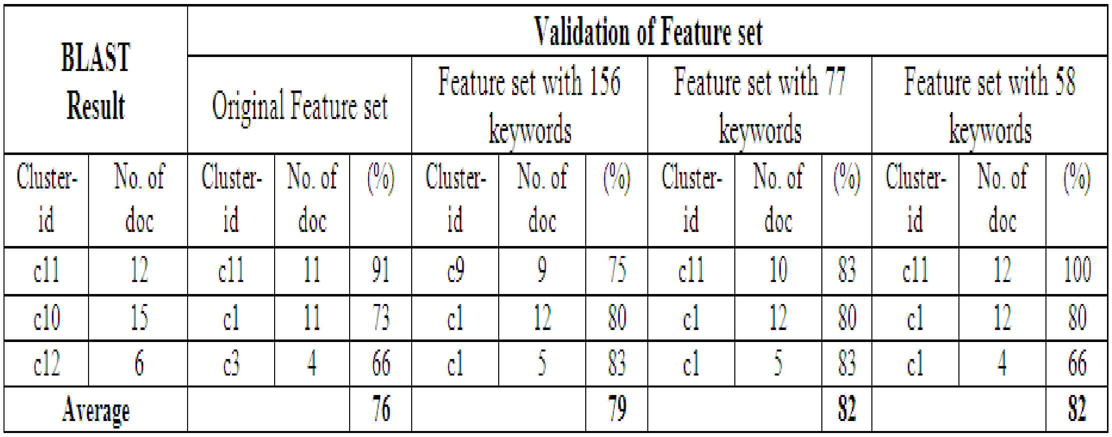
Table 7. Evaluation Metrics for Feature set

Feature Set	Keywords	Precision (%)	Recall (%)	F-measure (%)
Original Feature set	358	80	32	45
Extracted Feature set	156	83	48	60
	77	82	77	80
	58	84	95	89

3.5.2 Validating with BLAST

In our proposed work the feature sets are extracted using Filter based Feature selection methods. The extracted feature set is compared with the Standard BLAST tool. The Clustering result produced by the BLAT tool is accurate and it is already proven. We have compared our result with BLAST to verify the biological relevance of the documents grouped based on the feature set. Here, we have taken three clusters {c11, c10, c12} from BLAST and compare it with our approach. The following Table 8 shows the results of Biological Validation Using BLAST.

Table 8. Biological Validation Using BLAST



BLAST Result		Validation of Feature set											
		Original Feature set			Feature set with 156 keywords			Feature set with 77 keywords			Feature set with 58 keywords		
Cluster-id	No. of doc	Cluster-id	No. of doc	(%)	Cluster-id	No. of doc	(%)	Cluster-id	No. of doc	(%)	Cluster-id	No. of doc	(%)
c11	12	c11	11	91	c9	9	75	c11	10	83	c11	12	100
c10	15	c1	11	73	c1	12	80	c1	12	80	c1	12	80
c12	6	c3	4	66	c1	5	83	c1	5	83	c1	4	66
Average				76			79			82			82

The documents clustered using the feature set is verified with the documents clustered using the BLAST. We have found that the feature set with 77 keywords and 58 keywords has grouped 82% of similar documents that are grouped by existing BLAST tool. In this study we have found that the feature set with 77 keywords and the feature set with 58 keywords are biologically relevant for grouping the documents. From the experimental result we have consider the Feature set with

77 keywords and Feature set with 58 keywords extracted based on supervised and unsupervised Filter based approach can be used for Clustering and Classification.

4. CONCLUSION

Mining biological data is an emerging area of intersection between bioinformatics and data mining. It is difficult to explore and utilize the features set from huge amount of data. Feature selection is a method for improving the efficiency and accuracy of text categorization algorithms by removing redundant and irrelevant terms from the corpus. The proposed work is to study and analyze the Filter based Feature Selection Approach for identifying Feature from Genomic Sequence Database. A framework "Genomic Feature set selection (GFS)" is designed to implement the proposed approach. The framework is processed using four different phases. The experimental results it is found that the Analysis of Feature set with 58 keywords is 89% accurate for grouping the documents which is evaluated and verified by using F-Measure. The implemented work is also validated with existing Standard BLAST tool. On validating, we have identified that 82% of documents grouped by Feature set with 77 and 58 keywords are similar to the documents grouped by the BLAST tool.

REFERENCE

[1] Asha Gowda Karegowda, M.A.Jayaram, A.S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", 2010 International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 7.

[2] Balaji Krishnapuram, Alexander J. Hartemink, Lawrence Carin, Mario A.T. Figueiredo, "A Bayesian Approach to Joint Feature Selection and Classifier Design " , IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 26, NO. 9, SEPTEMBER 2004.

[3] Bassam Al-Salemi ., Mohd. Juzaiddin Ab Aziz , "Statistical Bayesian Learning For Automatic Arabic Text Categorization" , In Journal of Computer Science 7 (1): 39-45, 2011.

[4] Bharati M. Ramageri , "Data Mining Techniques And Applications", In Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 , 2010.

[5] Daniel T. Larose, " *Discovering Knowledge in Data: an Introduction to Data Mining,* ISBN 0-471-66657-2, 2005.

[6] Feng Tan, Xuezheng Fu, Hao Wang, Yanqing Zhang, and Anu Bourgeois, "A Hybrid Feature Selection Approach for Microarray Gene Expression Data", V.N. Alexandrov et al. (Eds.): ICCS 2006, Part II, LNCS 3992, 2006.

[7] Feng Tan, Xuezheng Fu, Hao Wang, Yanqing Zhang, and Anu Bourgeois, "A Hybrid Feature Selection Approach for Microarray Gene Expression Data", V.N. Alexandrov et al. (Eds.): ICCS 2006, Part II, LNCS 3992, 2006.

[8] George H.John, Ron Kohavi, Karl Pfleger,"Irrelevant Features And The Subset Selection Problem",Machine Learning : Proceedings of the Eleventh International Conference,Morgan Kaufmann Publishers,San Francisco,CA,1994.

[9] Gouchol Pok, Jyh-Charn Steve Liu, Keun Ho Ryu, "Effective feature selection framework for cluster analysis of microarray data", ISSN 0973-2063 (online) 0973-8894 ,Bioinformation 4(8): 385-389,2010.

[10] Guyon, Isabelle, and Andr´e ELISSEEFF, "An introduction to variable and feature selection". Journal of Machine Learning Research, 2003 1157–

[11] Lin Sun, Jiucheng Xu, Zhan'ao Xue, Lingjun Zhang, "Rough Entropy-based Feature Selection and Its Application", Journal of Information & Computational Science 8: 9 (2011) 1525–1532.

[12] Pabitra Mitra C.A.,Murthy, SanKar K.Pal, “Unsupervised Feature Selection Using Feature Similarity”, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, NO. 4, April 2002.

[13] Tao Liu, Shengping Liu, Zheng Chen, “An Evaluation on Feature Selection for Text Clustering”, Proceedings of the Twentieth International Conference On Machine Learning(ICML-2003).

[14] Xiubo Geng, Tie-Yan Liu, Tao Qin, Hang Li, “ Feature Selection for Ranking”,2007.

[15] Yiming Yang and Xiu Liu, “A re-examinations of text categorization methods”,1992.

[16] Yiming Yang, “A Comparitive study of Feature selection in Text Categorization”,1997.

[17] Yiming Yang, “Noise Reduction in a Statistical Approach to Text Categorization”,1995.

[18] YongSeog Kim, W. Nick Street and Filippo Menczer, “Feature Selection in Unsupervised Learning via Evolutionary Search”, 2000.

[19] YongSeog Kim, W. Nick Street, and Filippo Menczer, “Feature Selection in Data Mining”.

[20] Yvan Saeys, Inaki Inza2 and Pedro Larranaga, “A review of feature selection techniques in bioinformatics”,Vol.23 no. 19 2007, pages 2507–2517 doi:10.1093/bioinformatics/btm344.

[21] Zheng Zhao, Lei Wang and Huan Liu, “Efficient Spectral Feature Selection with Minimum Redundancy”,2010.

Authors

Ms V Bhuvaneswari received her Bachelors Degree (B.Sc.) in Computer technology from Bharathiar University, India 1997 , Masters Degree (MCA) in Computer Applications from IGNOU, India . and M.Phil in Computer Science in 2003 from Bharathiar University, India . She has qualified JRF , UGC-NET, for Lectureship in the year 2003 She is currently pursuing her doctoral research in School of Computer Science and Engineering at Bharathiar University in the area of Data mining . Her research interests include Bioinformatics, Soft computing and Databases. She is currently working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, India. She has for her credit publications in journals, International/ National Conferences.

Ms. K. Poongodi received her B.Sc degree in Mathematics with Computer Applications from Bharathiar University, Master Degree(MCA) in Computer Applications from Anna University, M.Phil in Computer Science in 2007, 2010, and 2011 respectively, from Bharathiar University, Coimbatore, India. She has attended National conferences and Presented a paper. Her area of interest includes Data Mining, Data Structures and Bioinformatics.