Available online at www.ijpe-online.com

International Journal of Performability Engineering

Vol. 13, No. 4, July 2017, pp. 383-389
DOI: 10.23940/ijpe.17.04.p5.383389

Clustering-Based Feature Selection Framework for Microarray Data

Smita Chormunge[a,*] and Sudarson Jena[b]

[a]*Research Scholar, Department of Computer Science and Engineering, GITAM University, Hyderabad, INDIA*
[b]*Department of Information Technology, GITAM University, Hyderabad, INDIA*

Abstract

Gene's expression data contains hundreds to thousands of features. It is challenging for machine learning algorithms to find the relevant information from such huge and correlated data. Irrelevant and redundant features are computationally costly and decrease the accuracy of machine learning algorithms. Feature selection plays important role to solve the problem of dimensionality. But most of the traditional feature selection algorithms fail to scale on high dimensionality problems. In this paper Clustering based Feature Selection Framework named as (CFSF) is proposed. CFSF produces optimal feature subset by eliminating irrelevant features using clustering algorithm and redundant features by applying filter measure on each cluster. Extensive experiments are carried out to compare proposed framework and other representative methods with respect to two classifiers namely Naive Bayes and Instance Based on microarray datasets. The empirical study demonstrates that the proposed framework is very efficient and effective for producing optimal feature subset and improves classifier performance.

Keywords: Feature Selection; Information Gain; Clustering; Classifiers; Filter measure.

(Submitted on December 4, 2016; Revised on May 7, 2017; Accepted on June 18, 2017)

1. Introduction

Recent technological developments and applications such as text categorization, information retrieval, image retrieval and Deoxyribo Nucleic Acid (DNA) microarray contain vast amounts of multivariate data. Unfortunately, the enlargement of data volume far outpaces human's ability to understand and handle it. Data mining deals with such large data by discovering patterns in large data sets. Applications with hundreds to thousands of attributes make it challenging for machine learning to find useful information from gigantic data streams, these larger and more complex data accumulate at an unprecedented speed. To represent data, large numbers of candidate features are collected because of unfamiliar relevant features. Irrelevant features do not change the target concept learnt through machine learning and redundant features do not include anything new to the target concept [1], but these irrelevant and redundant features significantly increase the computational cost of a learning process and decrease accuracy.

Because of this dimensionality problem most of the traditional learning algorithms fail to scale on large data with high features. In addition, the existence of noisy features degrades the performance of learning algorithms. Feature selection techniques are helps improving efficiency in solving the dimensionality problems in machine learning by selecting relevant and non-redundant features [2,3]. Selecting useful data makes learning task faster and more accurate. Machine learning algorithms become more scalable, reliable and accurate with assistance of feature selection. Feature selection is also helpful for prediction in data analysis process by selecting closed and related features. To produce a good feature subset, a feature selection algorithm can be use evaluation measures and search techniques. Most of the feature selection algorithms [4,5,6,7] have been proposed for classification techniques. Numerous of feature selection algorithms use statistical measures i.e. mutual

* Corresponding author.
E-mail address: smita2728@rediffmail.com

information, correlation and information gain measure. Based on evaluation measures there are three general approaches to feature selection that are Filters, Wrappers and Embedded methods [8].

Recent research works used feature clustering to improve the performance of learning algorithms and it demonstrated that application of cluster analysis has been more effective than traditional feature selection algorithms. Krier [9] has presented a methodology combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information. Van Dijck [10] proposed same methodology as Krier except that the former forces every cluster to contain consecutive features only. Qinbao Song [11] proposed a fast clustering based feature selection algorithm (FAST) based on the Minimum Spanning Tree (MST) method. Feature selection can simplify the calculation and help to get an accurate data model in data clustering [12]. Dario and Raul [13] has presented an algorithm that aims at finding subsets of features which are coherent with the Laplacian and, at the same time, present low redundancy between them. Study of this paper is an extension of feature selection algorithm based on spectral graph theory. Its focus is on solving redundancy problem for microarray data. Redundancy is measured by taking a geometric approach using some selected eigenvectors from the Laplacian matrix. This connects with ideas from spectral clustering theory. Before applying spectral clustering theory, pre-processing step is applied and checked for meaningfulness of the selected features from a biological point of view. Proposed algorithm yields in 5 top features. Better accuracy may not be achieved consistently by selecting best 5 features. Gouchol [14] has proposed efficient feature selection framework. It is appropriate to extract class-specific properties as a small number of genes from two-dimensional microarray data. Their feature selection method addresses the key issue in feature selection by removal of irrelevance and redundancy. Gouchol [14] has considered the genes and samples at the same time in selecting the most influential genes. This way of selection may lead to increase in computational time.

This paper proposes Clustering based Feature Selection Framework named as CFSF to find most relevant feature subset in lesser time with better accuracy for microarray datasets. Proposed framework uses k-means clustering algorithm to form the clusters and remove irrelevant features then applies the filter measure to eliminate the redundant features and produce representative feature subset. Well known microarray datasets are used for experimental study. Proposed work is compared with other renowned feature selection algorithms such as Relief [15], IG [16] and Chi-squared [17] with respect to Naive Bayes and IB1 classifiers. Accuracy of proposed framework is tested with respect to all classifiers.

The section 2 of this paper presents a proposed framework. The empirical study and results are discussed in section 3. Finally, section 4 of this paper presents the conclusion.

2. Proposed Framework

The real world tasks often use many features, only few of which are relevant to the target concept. Number of candidate features introduced as relevant features are unknown for that task. Irrelevant, noisy and redundant features degrade the performance of learning algorithms, due to high dimensionality has impact on speed and accuracy of the result. To deal with such high dimensional genes data we proposed the Comprehensive feature selection framework as shown in figure 1. It is based on clustering algorithm and filter measures named as CFSF. Proposed framework is a combination of Clustering algorithm and information based filter measure.

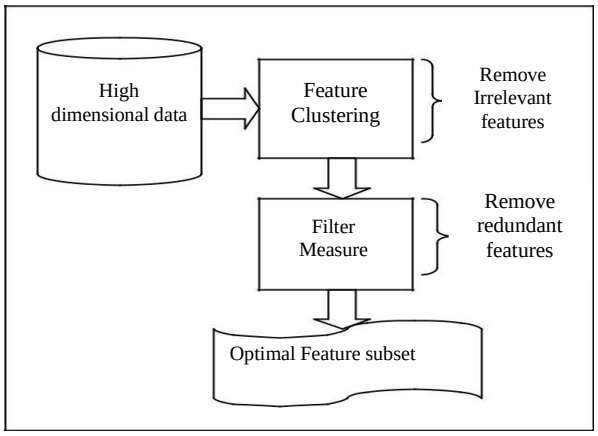

Figure 1. Proposed Framework (CFSF)

First we formed groups of features by using k-means clustering algorithm. K-means algorithm is very commonly used clustering algorithm. K-means is popular because of its linear complexity. It is also popular because of ease of implementation and speed of convergence and adaptability to sparse data [18]. Before forming cluster, user should specify number of clusters [19]. In K-means formulation, one is given an integer k and a set of n data points. k is the number of cluster centers. The goal is to choose k centers C to minimize the sum of the squared distances between each point and its closest center [20].

$$\varphi = \sum_{\mathbf{x}\in\chi} \min_{\mathbf{c}\in C} \|\mathbf{x} - \mathbf{c}\|^2 \quad (1)$$

It is standard practice to choose the initial centers uniformly at random from χ. Euclidean distance function is used to measure the similarity between features. It is a distance between two points in Euclidean space [21]. To denote the distance between vectors X (X_1, X_2...,X_n) and a point Y (Y_1, Y_2...,Y_n) notation $d_{x,y}$ is used. The Euclidean distance between two n dimensional vectors is written as shown in Eq. (2) [22]:

$$d_{x,y} = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (2)$$

The features which do not comply to any cluster after forming groups are considered as irrelevant features and are eliminated from the calculation. Once the clusters are formed the filter measure is applied on each cluster to remove redundant features.

CFSF Algorithm:

Input: Dataset D {F_1…..F_n}
Θ-Threshold value
Output: O-Optimal Feature Subset

//Forming Clusters

1) Arbitrarily choose k initial centers $C = \mathbf{c}_1, \ldots, \mathbf{c}_k$,
2) For each $I \in \{1, \ldots, k\}$, set the cluster c_i to be the set of points in χ that are closer to c_i than they are to c_j for all $j \neq i$
3) For each $i \in \{1, \ldots, k\}$, set c_i to be the center of the mass of all points in a set C_i of cluster i. $C_i = \frac{1}{|c_i|}\sum_{x\in c_i} X$
4) Repeat Steps 2) and 3) until $\mathbf{c}_i$ no longer changes.

// Applying Filter measure on each cluster

5) Compute IG = H(C) −H(C/A) by Eq. (3) and (4),
6) Repeat steps for all features of Clusters
7) Set Threshold value Θ.
   For all relevant features {f_1…. f_n} If (IG > Θ)
8) O= IG {f_1…. f_k}

Feature selection algorithms use statistical measures based on distance, information and error rate. Here we used information based filter measure on each cluster. Information based measure is a simplest attribute ranking method, this measure determine the information gain from a feature. It calculates the difference between entropy of the distribution before the split and the entropy of the distribution after the split. Attributes are ranked based on the information gain value. It is widely used in text categorization applications and the recent studies found on microarray data analysis and image data analysis [16]. If A is an attribute and C is the class, Eq. (3) and (4) show the entropy of the class before and after observing the attribute [16]. Ranker search method is used for ranking the features.

$$H(C) = -\sum_{c\in C} P(c)\log_2 P(c) \quad (3)$$

$$H(C|A) = -\sum_{a\in A} P(a)\sum_{c\in C} p(c|a)\log_2 p(c|a)$$

Each attribute A_i is assigned a score based on the information gain between itself and the class:

$$IG_i = H(C) - H(C|A_i) \quad (4)$$

$$= H(A_i) - H(A_i|C)$$

$$= H(A_i) + H(C) - H(A_i|C)$$

For example consider two variables X and Y. We can see they are dependent or independent to each other by calculating information gain about X after observing Y and the amount of information gained about Y after observing X. If value is the same, it ensures they have equal information. Which indicates these are redundant features. Likewise, we find all redundant features from each cluster and eliminated those features. Remaining relevant features are ranked by decreasing order based on their information values calculated by equation (3) and (4). To get the smallest subset threshold value is applied. If the features computed value is less than threshold value, then those features are eliminated from the list and remaining relevant features are ranked.

Computational time is calculated to build dataset for proposed framework and other methods. To verify the accuracy of proposed work with respect to classifiers, F-measure metric is used. F-measure is one of the external metric which measures the effectiveness of clustering algorithms. It computes the score by calculating two factors Precision and Recall where 1 is the best value and 0 is worst value of score. Precision is the number of correct positive results divided by the number of all positive results. Recall is the number of correct positive results divided by the number of positive results that should have been returned. It computes the weighted average of Precision and Recall [23] as shown in Eq. (5). The F-measure metric is calculated as given by

$$F=2 * \frac{Precision.recall}{Precision+recall} \tag{5}$$

3. Result and Analysis

3.1 Empirical study

In this section we have presented a summary of well-known Microarray datasets used for empirical study which features ranges from 2000 to more than 7000. Colon Cancer dataset is a microarray dataset which enclose 62 samples collected from patients. Out of 62, 40 samples are tumor biopsies labeled as "negative" and 22 samples are normal labeled as "positive" biopsies are from healthy parts of the colons of the same patients, the total 2000 number of genes to be tested. Small Round Blue Cell Tumors (SRBCT) dataset is Gene expression data which contains 2308 genes for 83 samples from the microarray experiments of SRBCT. In 83 samples 63 samples are training samples and 25 are test samples. Lymphoma is a broad term encompassing a variety of cancers of the lymphatic system. It contains 4026 number of genes the number of samples is 62. There are all together three types of lymphomas. The first type Chronic Lymphocytic Lymphoma, the second type Follicular Lymphoma and the third type Diffuse Large B-cell Lymphoma.

The Leukemia data set include 7129 genes taken over 72 samples. Two variants of leukemia sample (AML, 25 samples, or ALL, 47 samples). CNS microarray dataset is for study of heterogeneous group of tumors. It contains 7129 number of genes and 42 numbers of samples (heterogeneous group of tumors). It contains 7129 number of genes and 42 numbers of samples [25]. Table 1 represents the summary of datasets. Data mining tool Weka [24] is used for analyzing the results.

Table 1. Summary of Datasets used for empirical study

Datasets	Features	Instances	Class
Colon Cancer	2000	62	2
SRBCT	2308	83	4
Lymphoma	4026	62	3
Leukemia	7129	72	2
CNS	7129	60	2

Table 2. Evaluation of Feature selection methods for Naïve Bayes classifier

Datasets	CFSF		Relief		IG		Chi-squared	
	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy
Colon Cancer	0.3	0.92	1.94	0.55	0.52	0.55	0.48	0.55
SRBCT	0.34	0.93	2.98	0.98	0.54	0.98	0.49	0.98
Lymphoma	0.53	0.95	3.59	0.93	0.72	0.93	0.73	0.93
Leukemia	1.05	0.93	6.72	0.97	0.99	0.97	1.2	0.97
CNS	0.91	0.88	4.23	0.62	0.99	0.62	1.02	0.62

Table 3. Evaluation of Feature selection methods for IB1 classifier

Datasets	CFSF		Relief		IG		Chi-squared	
	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy
Colon Cancer	0.48	0.93	1.29	0.77	0.23	0.77	0.21	0.77
SRBCT	0.25	0.95	2.87	0.84	0.33	0.84	0.54	0.84
Lymphoma	0.42	0.96	3.2	0.97	0.55	0.97	0.6	0.97
Leukemia	0.84	0.93	5.63	0.87	0.75	0.87	0.91	0.87
CNS	0.78	0.70	3.85	0.57	0.67	0.57	0.88	0.57

3.2 Analysis

We evaluated the computational time in seconds and accuracy for each dataset using proposed framework and compared results with other representative feature selection methods such as Relief, IG and Chi-squared. The computational time and accuracy of proposed framework and comparative methods are analyzed with respect to different classifiers NB and IB1on each microarray dataset. A 10 fold Cross-validation strategy is used for evaluating accuracy where each dataset is divided into ten partitions. Nine are used as training set and one is used as test set. This process is repeated for 9 times. We set the relevance threshold value as -1.79 which is default value given in Weka [24]. Evaluation of proposed framework and other methods for Na ïve Bayes classifier results are represented in table 2. We have shown better performance of each dataset with CFSF in boldface. Evaluation of Colon Cancer dataset with Na ïve Bayes classifiers by using proposed framework takes 0.3 sec time to build model and accuracy is 0.92% whereas comparative feature selection methods Relief, IG and Chi-squared takes more computational time 1.94 sec,0.52 sec and 0.48 sec respectively and accuracy is very low 0.55%. For SRBCT dataset execution time is less for proposed framework as comparative with other methods but accuracy little bit decrease. By observing results with Na ïve Bayes classifier Proposed Framework performed well with Colon cancer, SRBCT, Lymphoma and CNS dataset for time and for accuracy point of view, it fits/ suits well for Colon Cancer, Lymphoma and CNS datasets.

The results obtained for feature selection methods with IB1 classifier is shown in table 3. The datasets SRBCT, Lymphoma performed well in time for proposed framework and good in accuracy except Lymphoma. Relief, IG and Chi-squared methods accuracy is almost same but varies in time.

Runtime to build the datasets for different feature selection algorithms Relief, Chi-squared and IG and CFSF with respect to different classifiers NB and IB1 is graphically represented in Figure 2 and 3 respectively. Relief feature selection algorithm takes more time to execute datasets as compared with other methods for all classifiers. CFSF works well with IB1 classifier than other classifiers. Accuracy of the Feature Selection algorithms with Na ïve Bayes and IB1 Classifier is represented in fig. 4 and 5 respectively. Accuracy point of view CFSF performs well with IB1 classifier than Na ïve Bayes classifier. By observing all results we found that proposed framework is fastest and good in accuracy though in most of dataset.

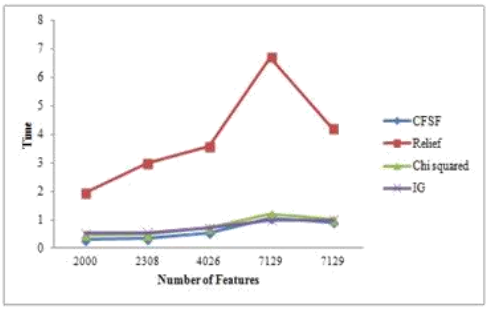

Figure 2. Runtime of the Feature Selection Algorithms with Naive Bayes Classifier

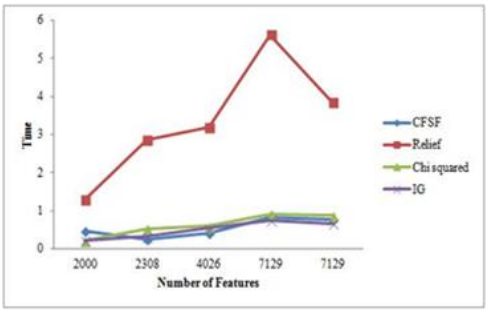

Figure 3. Runtime of the Feature Selection Algorithms with IB1 Classifier

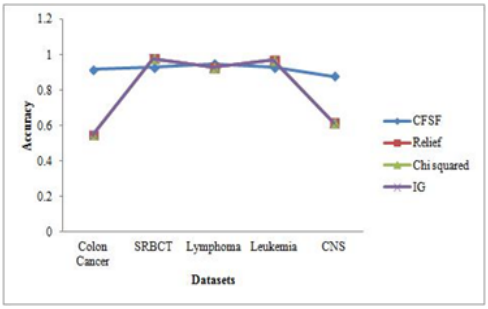

Figure 4. Accuracy of the Feature Selection Algorithms with Naïve Bayes Classifier

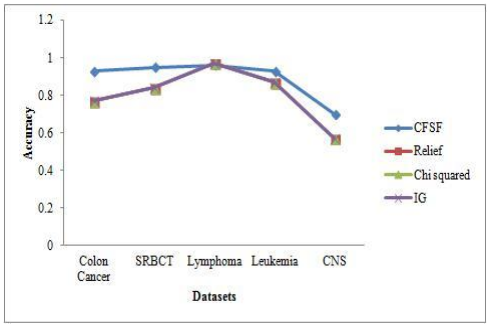

Figure 5. Accuracy of the Feature Selection Algorithms with IB1 Classifier

4. Conclusion

Gene's data challenges the traditional feature selection algorithms because of high dimensionality problem which reduces the performance of learning algorithms. In this paper Clustering based Feature Selection Framework (CFSF) is proposed for Microarray data. Proposed Framework produce good feature subset by removing irrelevant and redundant features by using clustering algorithm and filter measure. Computational time and accuracy are evaluated for proposed framework and compared these with other renowned feature selection methods with respect to Naïve Bayes and Instance based classifiers on each dataset. The results demonstrate that proposed framework is not only efficient and effective to produce feature subset but also improves classifier performance. Further, we plan to test this proposed framework on different datasets and to explore combining different measures to get more interesting results.

References

1. John, G. H., Kohavi, R. and Pfleger, K., "Irrelevant features and the subset selection problem" In *Proc. the Eleventh International Conference on Machine Learning*, 121-129, 1994.
2. M. Dash and H. Liu, 1997, "Feature Selection for Classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131 156.
3. Liu, H. and Yu, L, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491-502, 2005.
4. D. K. Bhattacharyya, J. K. Kalita, "Network Anomaly Detection: A Machine Learning Perspective," *CRC Press*, 2013.
5. H. Frohlich, O. Chapelle, B. Scholkopf, "Feature selection for support vector machines by means of genetic algorithm, in: Tools with Artificial Intelligence," *Proceedings 15th IEEE International Conference on, IEEE*, pp. 142–148,2003.
6. S.-W. Lin, K.-C. Ying, C.-Y. Lee and Z.-J. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection," *Applied Soft Computing*, 12 (10) 3285–3290,2012.
7. L. Yu, H. Liu, "Redundancy based feature selection for microarray data," *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2004, pp. 737–742,2004.
8. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.
9. C. Krier, D. Francois, F. Rossi, and M. Verleysen, "Feature Clustering and Mutual Information for the Selection of Variables in Spectral Data," *Proc. European Symp. Artificial Neural Networks Advances in Computational Intelligence and Learning*, pp. 157-162, 2007.
10. G. Van Dijck and M.M. Van Hulle, "Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis," *Proc. Int'l Conf. Artificial Neural Networks*, 2006.
11. Qinbao Song, Jingjie Ni, and Guangtao Wang A, "Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, No. 1, January 2013
12. Yu-MengXu , Chang-DongWang and Jian-HuangL, "Weighted Multi-view Clustering with Feature Selection," *Pattern Recognition*,53,pp-25-35,2016.
13. Darío Garcá-Garcá and Raúl Santos-Rodríguez, "Spectral Clustering and Feature Selection for Microarray Data," *Machine Learning and Applications, Fourth International Conference on* (2009), Miami Beach, Florida, Dec. 13, 2009 to Dec. 15, 2009, pp: 425-428, ISBN: 978-0-7695-3926-3: DOI http://doi.ieeecomputersociety.org/10.1109/ ICMLA. 2009.86
14. Gouchol Pok, Jyh-Charn Steve Liu, and Keun Ho Ryu, "Effective feature selection framework for cluster analysis of microarray data," *Bioinformation*. 2010; 4(8): 385–389. PMCID: PMC2951666.
15. K. Kira and L.A. Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm," *Proc. 10th Nat'l Conf. .Artificial Intelligence*, pp. 129-134, 1992.
16. Mark A. Hall, Geoff rey Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, NO. 3,2003.
17. G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," *J. Machine Learning Research*, vol. 3, pp. 1289-1305,2003.
18. Dhillon I. and Modha D, "Concept Decomposition for Large Sparse Text Data Using Clustering. Machine Learning," 42, pp.143-

175, 2001.
19. LeiWuy, Rong Jinz, Steven C.H. Hoiy, Jianke Zhu, and Nenghai Yu, "Learning Bregman Distance Functions and Its Application for Semi-Supervised Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 3, pp. 478-491, 2010.
20. Onoda, T., Sakai, M., "Independent component analysis based seeding method for k-means clustering," In: *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* 2011. doi:10.1109/WI-IAT.2011.29.
21. Smita Chormunge, Sudarson Jena, "Metric based Performance Analysis of Clustering Algorithms for High Dimensional Data," Proc International Conf on IEEE, doi 10.1109/CSNT CSNT, pp 1060-1064,.2015.127,2015.
22. Michael Greenacre,Raul Primicerio, "Measures of Distance between Samples: Euclidean.. Fundacion," *BBVA publication*, ISBN: 978-84-92937-50-9 pp-47-59,2013.
23. Bourennani F,Ken Q. Pu,Ying Zhu, "Visualization and Integration of Databases Using Self-Organizing Map," *IEEE International Conference on Advances in Databases, Knowledge, and Data Applications*, pp-155-160, 2009,DOI 10.1109/DBKDA.2009.30.
24. Remco R. Bouckaert, Eibe Frank, Peter Reutemann, Mark Hall, Richard Kirkby, Alex Seewald and David Scuse, "WEKA Manual for Version 3-7-10", 2013.
25. Datasets can be downloaded from https://archive.ics.uci.edu/ml/datasets/DBWorld+emails, http://csse.szu.edu.cn/staff/zhuzx/Datasets.html, http://repository.seasr.org/Datasets/UCI/arff/.

Smita Chormunge is a Ph. D. student of Computer Science and Engineering at GITAM University, Hyderabad, INDIA. She obtained her Master degree and Bachelor Degree in Computer Science and Engineering from SRTMU and PUNE University respectively. Her research interests are in data mining and High Dimensional data. She has published research works in journals and top conference proceedings such as ACM, IEEE and Springer.

Sudarson Jena is currently working as Associate Professor in the Dept. of Information Technology, GITAM University, Hyderabad. He received M. Tech degree in Computer Science and Engineering from JNTU- Hyderabad and Ph.D. degree in Computer Science from Sambalpur University, in 2008. Dr. Jena has published more than 65 Technical papers in referred Journals and Conference proceedings. One Ph. D scholar has successfully completed under his guidance at JNTU-Hyderabad in 2015 and 10 other Ph.D. research scholars are working under his supervision in different universities in India. Dr. Jena was the Joint-Editor of the Journal, The Chanakya during 2005-2007 and Mentor of Interscience Research Network (IRNet) since 2012. His research interests include Parallel and Distributed System, Data Mining, Reliability and Performance Evaluation of Interconnection Networks, Grid Computing, Cluster Computing and Soft Computing.