Clinical Epidemiology and Global Health 7 (2019) 171–176

Contents lists available at ScienceDirect

Clinical Epidemiology and Global Health

journal homepage: www.elsevier.com/locate/cegh

Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance

Sai Prasad Potharaju*, M. Sreedevi

Dept of CSE, K L University, Guntur, AP, India

ARTICLE INFO

Keywords:
Microarray
Feature selection
Classification
High dimensionality

ABSTRACT

Objective: The objective of this research article is to present a novel feature selection strategy for improving the classification performance over high dimensional data sets. Curse of dimensionality is the most serious downside of microarray data as it has more number of genes(features). This leads to discouraged computational stability. In microarray data analytics, identifying more relevant features required full attention. Most of the researchers applied two stage strategy for gene expression data analysis. In first stage, feature selection or feature extraction is employed as a preprocessing step to pinpoint more prominent features. In second stage, classification is applied using selected subset of features.

Method: In this research also we followed the same strategy. But, we tried to introduce a distributed feature selection(dfs) strategy using Symmetrical Uncertainty(SU) and Multi Layer Perceptron(MLP) by distributing across the multiple clusters. Each cluster is equipped with finite number of features in it. MLP is employed over each cluster, and based on the highest accuracy and lowest Root Mean Square error rate(RMS) dominant cluster is nominated.

Result: Classification accuracy with Ridor, Simple Cart (SC), KNN, SVM are measured by considering dominant cluster's features. The performance of this cluster is compared with the traditional filter based ranking techniques like Information Gain(IG), Gain Ratio Attribute Evaluator(GRAE), Chi-Squared Attribute Evaluator (Chi). The proposed method is recorded approximately 57% success rate, 18% competitive rate against traditional methods after applying it over 7 well high dimensional and one lower dimension dataset.

Conclusion: The proposed methodology applied over very high dimensional microarry datasets. Using this method memory consumption will be reduced and classification performance can be improved.

1. Introduction

Feature Selection (FS) is routinely used as a preprocessing step for high-dimensional data. Any technique for high-dimensional data should deal with curse of dimensionality. Few hundreds to thousands of variables of microarray dataset leading to very high dimensionality. The performance of technique demeaned as the number of features rises for the data analysis.[1] FS is a process of removing an insignificant and superfluous features and attempts to find a candidate set of features that can best describes the data.[2] Irrelevant and superfluous features change or bias the learning algorithm's correctness, as they do not provide surplus knowledge to prediction and also it may puzzle the algorithm during learning and classification phase. Hence, it is an obligatory to drop insignificant and redundant features. Because of the high dimensionality the classification task has become more difficult.

As all the genes in a gene expression microarray data can't influence the classification model, it is now hour of need to derive the key genes by which classification task can become easier and accurate. For this issue, a possible choice is applying FS technique over high dimensional dataset. As a result of FS methods, computing time will be saved as well as prediction performance can be accelerated. FS can be called as attribute selection or variable selection in statistics. Because of the high dimensional property and lower instances rate of microarray data, a great computational approach is required to handle the difficulties of data analysis. It is strongly accepted that in majority of microarray data only few genes play a critical role in classification task and remaining genes are considered as irrelevant. So, FS concept is an important in preprocessing before proceeding to classification task.

To lift this subject there are 3 different types of FS approaches are existed in literature. Those are namely Filter, Wrapper, Embedded.[3] Filter mode is used to establish the subdivision of input features that have the significant predictive capacity. By nominating the righteous features, we can potentially upgrade the efficiency of classification. It uses distinct statistical tests to select the subdivision of features with the

* Corresponding author.
E-mail addresses: psaiprasadcse@gmail.com (S.P. Potharaju), msreedevi_27@kluniversity.in (M. Sreedevi).

https://doi.org/10.1016/j.cegh.2018.04.001
Received 9 March 2018; Accepted 26 April 2018
Available online 27 April 2018

excessive predictive power. For this, decide a statistical measure to apply, then compute the score for every feature. The features are then ordered by the score then features with the top scores are considered in learning model generation, while others are ignored. There by memory consumption can be pruned and computation time can be die down. Information gain (IG), Gain Ratio Attribute Evaluator (GRAE), Chi-Squared Attribute Evaluator (Chi), Symmetric Uncertainty (SU), and ReliefF (REL) are some of the popular filter based methods.

In this paper, we introduced a distributed feature selection (dfs) strategy using SU, Correlation-based Feature Subset Selection (CFS), and MLP by distributing across multiple clusters. As a result of this distribution process, each cluster is loaded with small sized features. To know the strong cluster among the formed clusters, MLP is employed on each cluster. Based on its accuracy and lowest RMS error rate, strong cluster is documented. This strong cluster of features is compared with existing filter based ranking methods like IG, GRAE, Chi.

The idea of this method is inspired from the project allocation strategy applied in academics. Generally, students in final year need to undergo group academic project. Project coordinator need to form the groups(each group consists of 'N' students). There are few possibilities and limitations to form the groups. Those are : students can voluntarily form their own group as per the defined group size, it is similar to selection of features randomly. If all toppers are in the same group, their project result can be better than other groups, it is similar to selecting top ranked features from the list. If all poor students are in one group, their outcome of project can be poor. Alternatively, if toppers, average, and poor students are mixed in a proper structure, every group can be balanced. So, average and poor students can learn some useful knowledge from toppers, it is the process of applying ensemble approach which is widely accepted in classification. Here, performance of group (cluster) is depending on the members(features) of that group. In this work, we tried to present a structure which can form the balanced group to give the adequate knowledge. The proposed structure to form the groups will be explained later. For this research, rank of each features is computed using SU, and the minimum number of features to be included in each cluster is defined by the result CFS.

For testing the strength of the proposed method, 7 well known microarray datasets which has features in the range 2000–12582, one lower dimensional dataset are used. After deriving the features by existed methods and proposed method, well known classification techniques such as Ridor(Rule), SC (Tree), KNN(lazy), SVM(Support Vector Machine) are employed, then respective performance is analyzed.

In the next portion of this article, some of the existing literature over microarray dataset is presented. Distribution of feature selection among the cluster is presented in methodology section. Dataset description and experimental setup is given in Section 4, finally results and appropriate discussion is produced in Section 5. Article is concluded with possible suggestions.

2. Literature

Our main objective of this current work is to establish a DFS framework to reduce or select the best subset of features from the high dimensional data set. There are some existing review reports accessible to direct these feature selection. As discussed in the introduction, there are 3 approaches of feature selection (Filter, Wrapper, Embedded), we considered filter based approach. As filter method designates the rank to each feature based on the statistical score of a feature, we chosen one of the filter method namely Symmetric Uncertainty (SU). Little description about SU is given in the methodology division.

In addition to SU, there are few more filter based modes(IG, GR, CHI, REL) available for selection of features and designating the rank to each feature. Based on the rank assigned, and depending on the number of features to be selected for classification, top ranked features can be selected for analysis. SU is applied in FAST (Fast clustering bAsed feature Selection algoriThm) clustering [4] and improved FAST clustering techniques for dimensionality reduction.[5] Both these are based on the graph theory concept. In FAST, prim's algorithm is employed to select the subset, whereas in improved FAST, kruskal's minimum spanning tree is employed to derive the features. Information Gain based clustering frameworks have been proposed to find the best features of kidney and voting data sets.[6] A new feature selector using minimum variance method based on information gain is implemented and tested on 9 datasets. It displayed greater than the traditional feature selection algorithms.[7] Ensemble based multi filter feature selection framework is proposed for DDoS detection in cloud computing by combining the features extracted by IG, GR, CHI, REL. Authors considered the feature which are greater than the defined threshold value from the ensembled features.[8] This method reduced the number of features from 41 to 13.

There have been some literature existed in the past which was conducted over various microarray dataset for addressing the high dimensionality issue using various FS approaches. Distributed FS mechanism is applied over 8 microarray datasets by partitioning the dataset vertically. From the each partition strong features are selected by applying various FS methods like IG, CFS, Consistency based filters and ReliefF. Researchers employed c4.5, Naive Bayes(NB), KNN and SVM classifiers for testing their proposed method.[9] KNN is applied on 5 microarray datasets using mapreduce and hadoop framework by executing over multiple cluster nodes. Depending on the size of the features, researchers used 2 and 18 cluster nodes.[10] CFS method is applied over lymphoblastic leukemia dataset. With this approach, researchers have drawn 16 top quality features. After employing KNN, MLP and SVM, 92.5% average accuracy is recorded.[11] Random Forest(RF), KNN, MLP and SVM is employed over colon and leukemia dataset after preprocessing using Partial Least Squared(PLS), Quadratic Programming FS (QPFS), Max Relevance(Max Rel), and minimum redundancy and maximum relevance(mRMR) methods. Final experimental results indicated that mRMR has recorded best performance among all. [12] Leukemia dataset is analyzed with SVM classifiers and ReliefF,CFS. Experimental analysis recorded above 90% accuracy.[13] For Classifying multiple disease states associated with cancer based on gene expression profiles a novel method is proposed by the researchers using SVM and Random Forest(RF).[14]

As colon dataset is suffering with imbalanced class issue, to balance the instance rate, SMOTE has been applied. After applying this method, ensembling approaches are employed with J48, RF, REPTree. Researchers also considered IG ranking based filter approach to select the prominent features.[15] SMOTE(Synthetic Minority Over-sampling Technique) and ensembling methods are used over kidney disease dataset by few researchers.[16],[17] Distance based FS is applied over colon and leukemia datasets. SVM is employed over selected features by the researchers. As per their analysis of research B/SVM is selected 6.36% features and 9.5% misclassification rate over colon dataset. Also, 9.54% of features and 3.08% misclassification is recorded over leukemia dataset.[18] Artificial Neural Networks (ANN), SVM, Bayesian Networks, Decision Trees are applied over cancer dataset in the mini review report published by the authors.[19]

3. Methodology

Proposed method is majorly based three important components: 1. Symmetric Uncertainty (SU), 2. Correlation-based Feature Subset Selection (CFS), 3. Multi Layer Perceptron(MLP).

3.1. SU

Symmetrical Uncertainty is a statistical measure which records the SU score of a feature then designates the rank to each feature. SU score with high value has top rank and less value has least rank. For feature selection, top ranked features can be selected depending on the requirement and type of the problem on which it will be applied. It can be defined as below.

$(SU) = 2*IG/(H(F1)+H(F2)$

IG is Information Gain; H(F1) is Entropy of F1; H(F2) is Entropy of F2

SU score will be in the coverage [0,1]. SU score 1 shows one feature can predict entirely others, 0 shows two features are uncorrelated. For introduced framework, feature which has SU score 0 is ignored as it can't influence the learning model.

3.2. CFS

It evaluates the strength of a subset of attributes by taking into account of the individual predictive ability of each attribute along with the degree of redundancy between them. Subsets of features that are strongly correlated with the class while having low intercorrelation are preferred. For drawing the subset, CFS employes the searching technique. In this paper, we used greedy stepwise searching technique is used. Instead of selecting the subset of features, we taken number of features drawn by this method to decide the minimum number of features to be included in each cluster.

3.3. MLP

MLP is one of the favored classifier applied in many domains for high accurate results, which is on the basis of neurons and perceptrons. As current methodology dissect the initial feature area into set of clusters(groups) to know the prominent cluster, primarily MLP is employed on each cluster formed by proposed approach then recorded its accuracy and RMS error rate.

The proposed methodology is as per the flowchart given in Chart 1 .

3.4. DFS algorithm

Input: D, C, N, TF, ListTF
D: Balanced data set
C: Number of clusters
N: Minimum number of features to be in each cluster
TF: Total number of features (SU score > 0)
ListTF : List of features whose SU score > 0
Output: Cluster of features

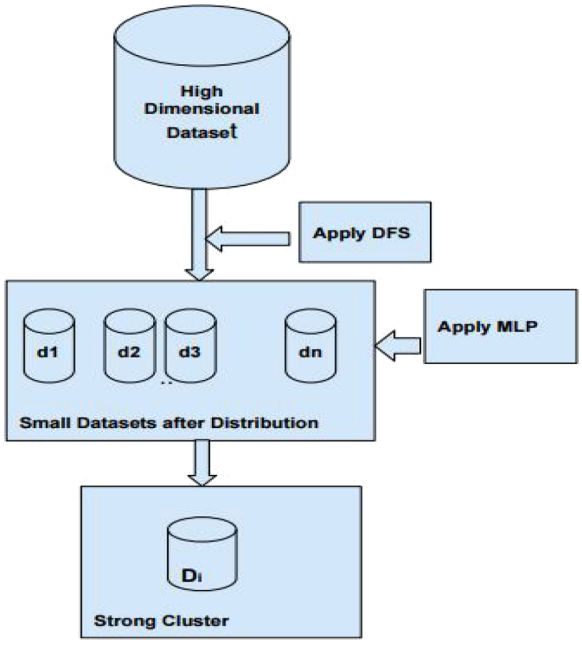


Chart 1. Distribution of Features among clusters.

Table 1
Cluster of features.

I[st] Order (C1)	2[nd] Order (C2)	3[rd] Order (C3)	Direction
f1	f2	f3	Left to Right
f6	f5	f4	Right to Left
f7	f8	f9	Left to Right
		f10	Right to Left

Cluster 1 : f1, f6, f7.
Cluster 2: f2, f5, f8.
Cluster 3: f3, f4, f9 (Note : f10 will be discarded as per step 9).

Table 2
Dataset description.[a]

Dataset	#F	#I	#C
COLON	2000	62	2
LEUKEMIA	7129	72	2
LEUKEMIA_3C	7129	72	3
LEUKEMIA_4C	7129	72	4
LYMPHOMA	4026	66	3
SRBT	2308	83	4
MLL	12,582	72	3
MUSK	166	476	2

[a] http://csse.szu.edu.cn/staff/zhuzx/Datasets.html.

Table 3
MLP performance over clusters formed by DFS.

Dataset	#NC	BCID	Accuracy	RMS	SF
COLON	5	3	83.87	0.3753	27
LEUKEMIA	12	11	98.61	0.1246	85
LEUKEMIA_3C	8	5	98.61	0.0778	114
LEUKEMIA_4C	6	5	97	0.1183	133
LYMPHOMA	12	6	100	0.032	182
SRBT	5	2	100	0.0257	133
MLL	36	16	100	0.044	145
MUSK	4	3	89.91	0.2908	36

1 Apply SU on D and define TF, store all features in ListTF in descending order of its rank (Rank 1 has high priority).
2 Apply CFS on D and define N
3 Define C

$C = (TF/N)$

4 Store first(next) 'C' number of features from ListTF in left to right direction, such that first feature would be placed in first cluster, second feature would be in second cluster and so on. Get next 'C' number of feature from ListTF for next iteration.
5 Store next 'C' number of features from ListTF in right to left direction, such that first feature would be placed in last cluster, second feature would be in second last cluster and so on. Get next 'C' number of feature from ListTF for next iteration.
6 Repeat step 4 and 5 until all features are organized.
7 Group, all vertically first order features into first cluster, second order features into second cluster, and so on.
8 If all clusters are structured with equal number of features, then stop. Otherwise remove the last feature from the cluster which has an extra feature.(As it can't influence the learning model).
9 Apply MLP on each cluster, based on the highest accuracy and minimum RMS error rate nominate the dominant cluster

3.5. Example

Assume T = 12 (total # features)
TF = 10 (# features whose SU value is greater than 0)
N = 3 (# features in each cluster)

Table 4
Result analysis of traditional and proposed DFS method.

Dataset	Method	Ridor	SC	KNN	SVM	Dataset	Method	Ridor	SC	KNN	SVM
Colon	IG	70.96	79.03	80.64	88.7	Lymphoma	IG	96.96	96.96	96.96	98.48
	GR	72.58	82.25	80.64	85.48		GR	100	100	96.96	100
	CHI	72.58	79.03	85.48	87.09		CHI	93.93	98.98	95.45	98.48
	DFS*	79.03	82.25	82.25	87.09		**DFS***	98.48	96.96	98.48	98.48
leukemia	IG	91.66	86.11	98.61	97.22	MLL	IG	84.72	86.11	95.83	97.22
	GR	90.27	84.72	95.82	97.22		GR	93.05	90.27	94.44	95.83
	CHI	91.66	86.11	95.83	97.22		CHI	84.72	86.11	95.83	98.61
	DFS*	93.05	94.44	95.83	98.61		**DFS***	87.5	88.88	93.05	100
leukemia _3c	IG	87.5	84.72	94.44	97.22	SRBT	IG	81.92	86.74	100	100
	GR	84.72	84.72	97.22	94.44		GR	81.92	81.92	100	100
	CHI	88.88	86.11	93.05	95.83		CHI	81.92	85.54	100	100
	DFS*	90.27	87.5	95.83	97.22		**DFS***	85.54	81.92	97.55	100
leukemia_4C	IG	88.88	84.72	97.22	95.83	MUSK	IG	74.15	81.51	84.03	78.15
	GR	91.66	88.88	93.05	91.66		GR	71.42	74.57	82.98	71.42
	CHI	90.27	87.5	93.05	91.66		CHI	73.1	79.2	82.98	78.78
	DFS*	81.94	83.33	95.83	97.22		**DFS***	73.1	74.78	84.66	78.99

* Proposed method.

Table 5
Win/Draw/Loss rate of proposed method.

	Ridor	SC	KNN	SVM	AVERAGE
WIN	70.83%	45.83%	45.83%	62.50%	56.25%
Draw	4.17%	19.05%	8.33%	29.17%	15.18%
Loss	25.00%	37.50%	45.83%	8.33%	29.17%

Graphical representation of all result analysis given in below Table 6.

C = 10/3 = 3 (# clusters)
ListTf= {f1, f2, f3, f4, f5, f6, f7, f8, f9, f10}
As per the algorithm proposed, features in each cluster will be formed as below Table 1.

4. Experiment

In this portion, description of datasets selected for testing the proposed distribution method is given, as well as the filter method and ranker approaches used for comparing the proposed method. For testing the accuracy level of proposed method and existing methods, 4 classifiers which are based on different conceptual theory are selected. The complete experiment is carried out using WEKA tool, with all default settings. The system configuration includes, Intel® Xeon(R) CPU E31220 @ 3.10 GHz × 4 processor with 8 GB Ram and 64 Bit, Ubuntu 16.04 operating system. The summary of datasets are given in below Table 2, which consists of total number of features (F), instances(I), Classes(C)

Datasets are divided into ⅔ and ⅓ ration respectively for training and testing purpose. As per the Table 2 statistics, there are 7 high dimensional datasets with minimum 2000 and maximum 12,582 features in it. To check the proposed method strength a lower dimensional set which has 166 features is considered.

After applying our proposed DFS method over all selected datasets, features are distributed across multiple clusters. For testing the strong clusters, MLP is employed over all those formed clusters and respective accuracy and RMS error rate is recorded. Table 3 shows the number of clusters formed(NC), best cluster id(BCID), accuracy of MLP and RMS error rate on it, number of features in best cluster(SF).

In this work, 3 well known feature ranking methods are used to compare the proposed study. Information Gain is one of the popular univariate practice of assessing the features. This assess the features as per their information worth and appraise a single attribute at a moment. It produces an orderly classification of all the attributes, and then a threshold is needed to choose a certain number of them as per the order recorded. The Gain Ratio is the non-symmetrical measure that is established to requite for the bias of the Information Gain.

In this work 4, well known classifiers are applied over the selected features. Those includes, K-Nearest Neighbor(KNN) is a classification algorithm which is an example of lazy lerner. Support vector machine (SVM) is another classification strategy a hyperplane in high -dimensional space. SC is tree based classification strategy and Ridor is rule based classification methods.

5. Results and discussion

After applying the proposed method, SF number of features are selected as per Table 3. To test the strength of the proposed approach, 3 different filter based ranking FS methods have been used with 4 different classifiers. From the traditional methods top ‘N” features are derived, where ‘N’ is equal to number of features(SF) in a strong cluster. For example, strong cluster of colon dataset has 27 features in it. So, top 27 features derived by traditional methods are chosen for evaluating the performance of proposed method. The performance evaluation of traditional and proposed method is produced in below Table 4.

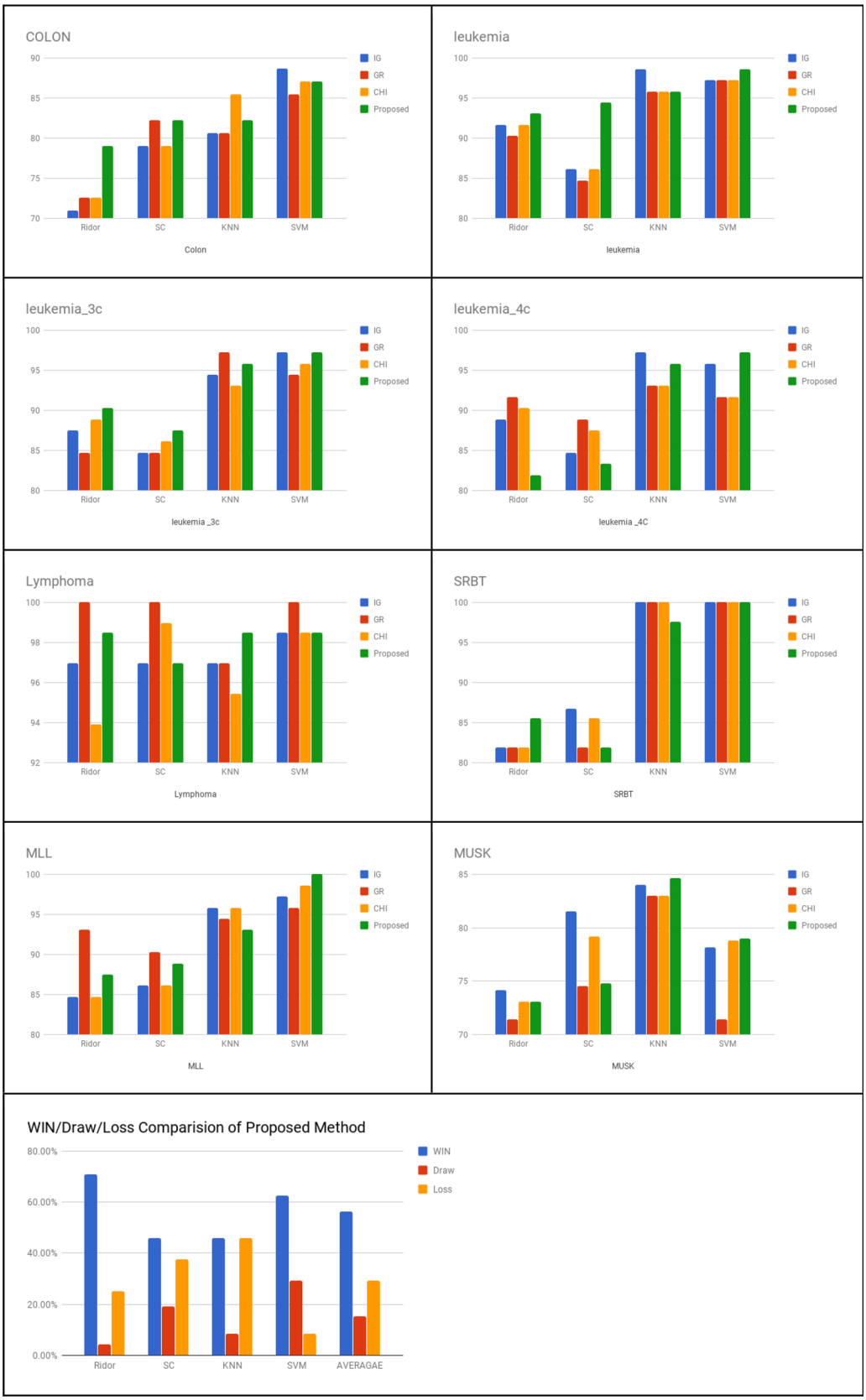
After detailed analysis of the results obtained, following noticeable points are gathered. Proposed method is recorded 75% success rate over IG and GR, 50% over chi with colon. Over leukemia, DFS is recorded 75% winning rate over traditional IG, GR, Chi. DFS recorded 75% success rate when compare with IG and GR, 100% success over Chi with leukemia _3c dataset. With leukemia_4c dataset, 50% success rate is obtained when compare with GR and Chi, it is 25% when compare with IG. In case of Lymphoma, DFS registered 25% winning rate when compare with IG and Chi, it is 50% with GR.Using SRBT, the proposed method is performed little poor, its success rate is 25% only. Using MLL dataset, success rate of dfs is 75% when compare with IG and Chi, it is 25% with GR. Dfs is performed 100% better when compared with GR, its performance is 50% with IG and Chi over Musk dataset.

Average winning, draw and loss rate of the proposed method after applying various classifiers is given in below Table 5. Success rate of ridor is 70.83%, SC is 45.83%, KNN is 45.83% and SVM is 62.50%. The average success rate of the proposed method when compare with traditional methods after applying 4 classifiers is 56.25%, draw rate of the dfs is 15.18%.

6. Conclusion

In this research attempt, we have proposed a distributed feature selection(dfs) approach that can be applied for complex high dimensional datasets for increasing the classification performance. As high dimensional dataset contains large number of redundant and irrelevant feature and only prominent features helps in predicting an unknown

Table 6
Graphical representation of result analysis.



instance, a novel method is needed for selecting the best features from the existed space. We used the Symmetrical Uncertainty, Correlation based filter selection and multi layer perceptron approaches for the proposed method. Using proposed approach, features are distributed among the various clusters without any repetition. Out of many clusters only one strong cluster based on the result of MLP is selected which has limited number of features. Proposed approach is compared with three well known filter based ranking feature selection methods and 4 various classification algorithms. The proposed method is recorded approximately 57% success rate, 18% competitive rate against the traditional methods after applying it over 7 well high dimensional and one lower dimension dataset. Finally, it is suggested that the analyzing each cluster with MLP in sequential order is time consuming, if large number of cluster are formed, it has to be analyzed parallely using any parallel programming concepts like hadoop which is our future study, so it could be considered as a generalized approach for distributing feature selection.

Acknowledgements

The authors would like to thank Department of Computer Engineering, SRES Sanjivani College of Engineering, Kopargaon, Maharashtra, India for providing necessary support for carrying the research work. The authors also like to thank to Research & Development team of K L University, India for their continuous support for carrying the research work.

References

[1]. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Feature selection for high-dimensional data. *Prog Artif Intell.* 2016;5(2):65–75.
[2]. Peralta D, del Río S, Ramírez-Gallego S, Triguero I, Benitez JM, Herrera F. Evolutionary feature selection for big data classification: a mapreduce approach. *Math Prob Eng.* 2015;2015.
[3]. Panthong R, Srivihok A. Wrapper feature subset selection for dimension reduction based on ensemble learning algorithm. *Proc Comp Sci.* 2015;72:162–169.
[4]. Song Q, Ni J, Wang G. A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE Trans Knowl Data Eng.* 2013;25(1):1–14.
[5]. Malji P, Sakhare S. *January. Significance of entropy correlation coefficient over symmetric uncertainty on FAST clustering feature selection algorithm. 2017 11th International Conference on Intelligent Systems and Control (ISCO)*. IEEE; 2017:457–463.
[6]. Potharaju SP, Sreedevi M. A novel cluster of feature selection method based on information gain. *Int J Control Theory Appl.* 2017;10(14):10–16.
[7]. Venkataraman S, Sivakumar S, Selvaraj R. A novel clustering based feature subset selection framework for effective data classification. *Indian J Sci Technol.* 2016;9(4).
[8]. Osanaiye O, Cai H, Choo KKR, Dehghantanha A, Xu Z, Dlodlo M. Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. *EURASIP J Wirel Commun Networking.* 2016;2016(1):130.
[9]. Bolón-Canedo Verónica, Sánchez-Maroño Noelia, Alonso-Betanzos Amparo. Distributed feature selection: an application to microarray data classification. *Appl Soft Comput.* 2015;30:136–150.
[10]. Kumar Mukesh, Kumar Rath Nitish, Swain Amitav, Kumar Rath Santanu. Feature selection and classification of microarray data using MapReduce based ANOVA and K-Nearest neighbor. *Proc Comp Sci.* 2015;54:301–310.
[11]. Singhal Vanika, Singh Preety. *Correlation based feature selection for diagnosis of acute lymphoblastic leukemia. Proceedings of the Third International Symposium on Women in Computing and Informatics*. 2015; 2015:5–9 ACM.
[12]. Sun Jing, Passi Kalpdrum, Kumar Jain Chakresh. *Improved microarray data analysis using feature selection methods with machine learning methods. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2016; 2016:1527–1534 IEEE.
[13]. Yang Sitan, Naiman Daniel Q. Multiclass cancer classification based on gene expression comparison. *Stat Appl Genet Mol Biol.* 2014;13(no. 4):477–496.
[14]. TAşçi A, İnce T, GüZELış C. A comparison of feature selection algorithms for cancer classification through gene expression data: leukemia case. *2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*. Bursa, Turkey2017; 2017:1352–1354.
[15]. Al-Bahrani Reda, Agrawal Ankit, Choudhary Alok. *Colon cancer survival prediction using ensemble data mining on SEER data. 2013 IEEE International Conference on Big Data*. 2013; 2013:9–16 IEEE.
[16]. Potharaju Sai Prasad, Sreedevi M. Ensembled rule based classification algorithms for predicting imbalanced kidney disease data. *J Eng Sci Technol Rev.* 2016;9(no. 5):201–207.
[17]. Potharaju Sai Prasad, Sreedevi M. An improved prediction of kidney disease using SMOTE. *Indian J Sci Technol.* 2016;9(no. 31).
[18]. Wenyan Z, Xuewen L, Jingjing W. Feature selection for cancer classification using microarray gene expression data. *Biostat 03 Biometrics Open Acc J.* 2017;1(2):555557. http://dx.doi.org/10.19080/BBOAJ.2017.01.555557.
[19]. Kourou Konstantina, Exarchos Themis P, Exarchos Konstantinos P, Karamouzis Michalis V, Fotiadis Dimitrios I. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J.* 2015;13:8–17.