Genes & Genomics (2019) 41:1301–1313
https://doi.org/10.1007/s13258-019-00859-x

Online ISSN 2092-9293
Print ISSN 1976-9571

RESEARCH ARTICLE

Improving classification accuracy of cancer types using parallel hybrid feature selection on microarray gene expression data

Lokeswari Venkataramana[1] · Shomona Gracia Jacob[2] · Rajavel Ramadoss[3] · Dodda Saisuma[1] · Dommaraju Haritha[1] · Kunthipuram Manoja[1]

Received: 27 February 2019 / Accepted: 2 August 2019 / Published online: 19 August 2019

Abstract

Background Data mining techniques are used to mine unknown knowledge from huge data. Microarray gene expression (MGE) data plays a major role in predicting type of cancer. But as MGE data is huge in volume, applying traditional data mining approaches is time consuming. Hence parallel programming frameworks like Hadoop, Spark and Mahout are necessary to ease the task of computation.

Objective Not all the gene expressions are necessary in prediction, it is very essential to select important genes for improving classification accuracy. So feature selection algorithms are parallelized and executed on Spark framework to eliminate unnecessary genes and identify only predictive genes in very less time without affecting prediction accuracy.

Methods Parallelized hybrid feature selection (HFS) method is proposed to serve the purpose. This method includes parallelized correlation feature subset selection followed by rank-based feature selection methods. The selected subset of genes is evaluated using parallel classification algorithms. The accuracy values obtained are compared with existing rank-weight feature selection, parallelized recursive feature selection methods and also with the values obtained by executing parallelized HFS on DistributedWekaSpark.

Results The classification accuracy obtained with the proposed parallelized HFS method is 97% and 79% for gastric cancer and childhood leukemia respectively. The proposed parallelized HFS method produced ~4% to ~15% improvement in classification accuracy when compared with previous methods.

Conclusion The results reveal the fact that the proposed parallelized feature selection algorithm is scalable to growing medical data and predicts cancer sub-types in lesser time with higher accuracy.

Keywords Parallelized hybrid feature selection · Correlation feature subset selection · Rank-based methods · Parallel classification · Spark · DistributedWekaSpark

✉ Lokeswari Venkataramana
lokeswaricts@gmail.com

Shomona Gracia Jacob
graciarun@gmail.com

Rajavel Ramadoss
rajavelr@ssn.edu.in

Dodda Saisuma
saisuma.dodda@gmail.com

Dommaraju Haritha
harithadommaraju@gmail.com

Kunthipuram Manoja
manojasgc@gmail.com

[1] Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai, India

[2] Muscat, Oman

[3] Department of ECE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Chennai, India

Introduction

Data mining is the procedure of mining knowledge from data and deals with the kind of patterns that can be mined. It plays a very important role in detection of fraud, providing better medical treatments at reasonable price, prediction of diseases at early stages, intelligent health-care decision support systems, etc., Various data mining techniques are classification, association, sequencing and clustering. These are predominantly used in health-care domain for early diagnosis of disease, selecting an appropriate treatment for the identified disease (Gracia Jacob 2015) and to design a drug for identified disease (Kang and Hong 2011). Health-care data mainly contains all the information regarding patients such as gene, protein, DNA, RNA sequences and about the parties involved in health-care industries. Humans have trillions of cells and each cell contains a complete copy of genome which is encoded into deoxyribonucleic acid (DNA). Gene is a segment of DNA that specifies how to make a protein. It is the basic physical and functional unit of heredity. They vary in size from few hundreds to more than two million bases. Gene expression is the process by which information encoded in a gene is converted into a protein (Yu et al. 2015). In cellular organisms, expression of right genes in right order at right time is crucial particularly during embryonic development and cell differentiation. Gene sequencing involves defining the order of bases and nucleotide units such as T, G, C, A (thymine, guanine, cytosine and adenine). When there is any change or variation in gene sequences, it results in inflammation of cell that results in cancer (Alshamlan et al. 2013). Cancer is basically a disease of 'genes gone bad'. Many cells control the way cells grow, divide and die. When there is an error in this, cell division may go out of control. Different kinds of cancer are caused by different sets of genes. So for its treatment, it is essential to know which of the genes in a cancer cell are behaving abnormally (Heo et al. 2013). The DNA microarray is the latest breakthrough in molecular biology, which provides researchers with an approach to monitor genome-wide expression systematically. Its application in cancer study has proved to be successful in elucidating the pathological mechanism and ultimately contributing to the battle against cancer. However, the current hurdle is how to make use of tremendous amount and ever-growing microarray experimental data to better predict cancer. MGE data has lot of noisy or irrelevant genes and missing data which affects prediction accuracy. Microarray data is high dimensional (thousands of genes) and low sample dataset (Golub et al. 1999). Feature selection which is a very important phase in data mining is essential to mine significant patterns from microarray data for detecting carcinogenic mutations. Selecting important genes from high-dimensional MGE data is time consuming for screening the disease. Gene (feature) selection reduces dimensionality of data and hence computation time also gets reduced. So, gene selection plays a major role in selecting predictive genes that are prominent in identifying the sub-types of cancer. The dimensions of MGE data is in thousands, all are not driver genes to identify the cancer sub-type. Hence, parallel gene (feature) selection methods select the required number of optimal genes from microarray data to improve classification of cancer sub-types in very less time.

As MGE data is high-dimensional, it can be run in parallel programming frameworks like Hadoop and Spark. Hadoop Map Reduce greatly simplified the big data analysis using large clusters of commodity hardware. But as data grows bigger, more complex, multi stage algorithms that need iterative processing and data sharing between stages are not supported by Hadoop (Peralta et al. 2015). So Apache Spark is used which is an open source big data processing framework that provides memory abstraction and efficiently shares data across the different stages of a map-reduce job with the help of in-memory data processing (Ryza et al. 2017). Every Spark application consists of a driver program that runs the users main function and executes various parallel operations on the worker or processing nodes of the cluster. The main memory abstraction that Spark provides is a resilient distributed dataset (RDD), which is a collection of elements partitioned across the nodes of the cluster that can be operated in parallel. Hence, in order to scale for growing health-care data, parallel programming frameworks are to be explored to parallelize data mining algorithms and extract hidden knowledge from huge health-care data.

Related work

Data mining in clinical datasets: Chuang et al. (2011) discussed on how important it is to select a proper number of relevant genes that directly affect classification accuracy and proposed a hybrid feature selection (HFS) method in microarray data analysis. The proposed method was implemented in two stages: In the first stage, a filter method called correlation-based feature subset selection (CFS) was used to calculate correlation-feature weight for each feature which helps in finding relevant features. In the second stage, a wrapper method called Taguchi-genetic algorithm (TGA) was applied on the features obtained in first stage to test them and thus find optimal feature subsets. These subsets were classified using K-nearest neighbor (KNN) method with leave-one-out cross validation (LOOCV) based on Euclidean distance calculations. Apart from these, genetic algorithms were used with randomness for global search

over entire search space. Lu et al. (2017) discussed that HFS algorithm provides highest classification accuracy when compared to conventional feature selection algorithms. They proposed MIMAGA-selection method that combines mutual information maximization (MIM) and adaptive genetic algorithm (AGA). The hybrid approach usually emphasizes on the advantages of the sub-algorithms and therefore is more robust when compared to traditional approaches. Li and Liu (2017) discussed various challenges of feature selection for big data analytics and ways to overcome them. They presented an open source feature selection repository called scikit-repository that assists researchers to achieve more reliable evaluation in developing new feature selection algorithms. Bolón-Canedo et al. (2015) suggested distributed feature selection on microarray data due to which execution time could be reduced unlike existing algorithms that work in centralized fashion. It discussed vertical partitioning of data (i.e., by features) as there are large number of features when compared to samples. Partitioning was done using two methods: (1) performing a random partition and (2) ranking the original features before generating the subsets. The second method was used to improve the performance of the first one. Then filter methods like information gain (IG) and ReliefF were applied on each of them and each returns a subset. Finally a merging procedure was used to combine the results. Merging method was used such that it eliminates redundant features in the subsets. The proposed method can also be executed in parallel as all the tasks are independent. This reduced execution time to a greater extent and classification accuracy is same or higher when compared to the existing algorithms. Hall (2000) proposed correlation-based feature selection (CFS) for discrete and numeric class machine learning. The algorithm is based on the hypothesis that good feature subsets contain features highly correlated with class and uncorrelated with each other. The author had demonstrated how CFS can be applied on both regression and classification problems of machine learning. CFS results in greater dimensionality reduction when compared to ReliefF. CFS can be a practical feature selector for machine learning algorithms. Ramani and Jacob (2013) presented a novel cancer prediction framework for gene expression datasets to improve accuracy. They proposed rank-weight feature selection (RWFS) method that identifies the features commonly reported by various feature selection algorithms. This method generated higher predictive performance with minimal set of features. Eiras-Franco et al. (2016) discussed multithreaded and Spark parallelization of feature selection filters. They explored new implementations of four feature selection algorithms i.e., ReliefF, IG, CFS, support vector machine recursive feature elimination (SVM-RFE). There is a significant improvement in execution time for ReliefF algorithm and scaled well in number of nodes for Spark. A new distributed CFS implementation in Spark obtained considerable improvement than that of existing multithreaded versions included in Weka, a new multithreaded IG implementation was found to be less cluster relevant i.e., it yields better results when implemented on a single computer, a new SVM-RFE multithreaded implementation processes multiclass datasets four times faster than that of sequential processing in Weka. Singh and Sivabalakrishnan (2015) reviewed about feature selection of gene expression data for cancer classification. Authors discussed that the feature selection methods consistently improves the performance. It is infeasible to use single algorithm for different datasets as each algorithm has different behavior. Feature selection plays a major role in accurately classifying large datasets like gene and protein expressions. So, proper cancer classification can be achieved using feature selection algorithms. Based on the existing literature, it was identified that feature selection on microarray gene expression data greatly improves classification accuracy and parallelization of algorithms on Spark reduces execution time.

The parallel recursive feature selection (RFS) method introduced in previous work Venkataramana et al. (2018) that utilized CFS approach to select prominent genes for accurate prediction of cancer sub-types. The parallel RFS method performed better than rank weight feature selection (RWFS) method in literature (Ramani and Jacob 2013) for brain cancer, glioblastoma and lung cancer but not for gastric cancer and childhood leukemia. RWFS method yielded classification accuracy of 93% and 65%, whereas parallel RFS method yielded classification accuracy of 92% and 64% for gastric cancer and childhood leukemia, respectively. Hence, the MGE data of these two cancer sub-types were further investigated to select optimal number of genes and improve classification accuracy. So a parallelized HFS method was proposed for MGE data which is detailed in the following section.

Proposed parallelized hybrid feature selection for classification of cancer sub-types

In MGE data, some of the features are irrelevant (i.e., all the features may not be necessary to classify or predict cancer sub-type) and redundant. A feature is potentially redundant when the information contributed by the feature is more or less same as one or more other features (Ali and Shahzad 2012). Among a group of potentially redundant features, a small number of features can be selected as a part of the final feature subset without causing any negative impact to learning model accuracy.

Hybrid approaches combine two or more well-studied algorithms to form a new strategy for solving a particular problem. The hybrid approach usually capitalizes on the

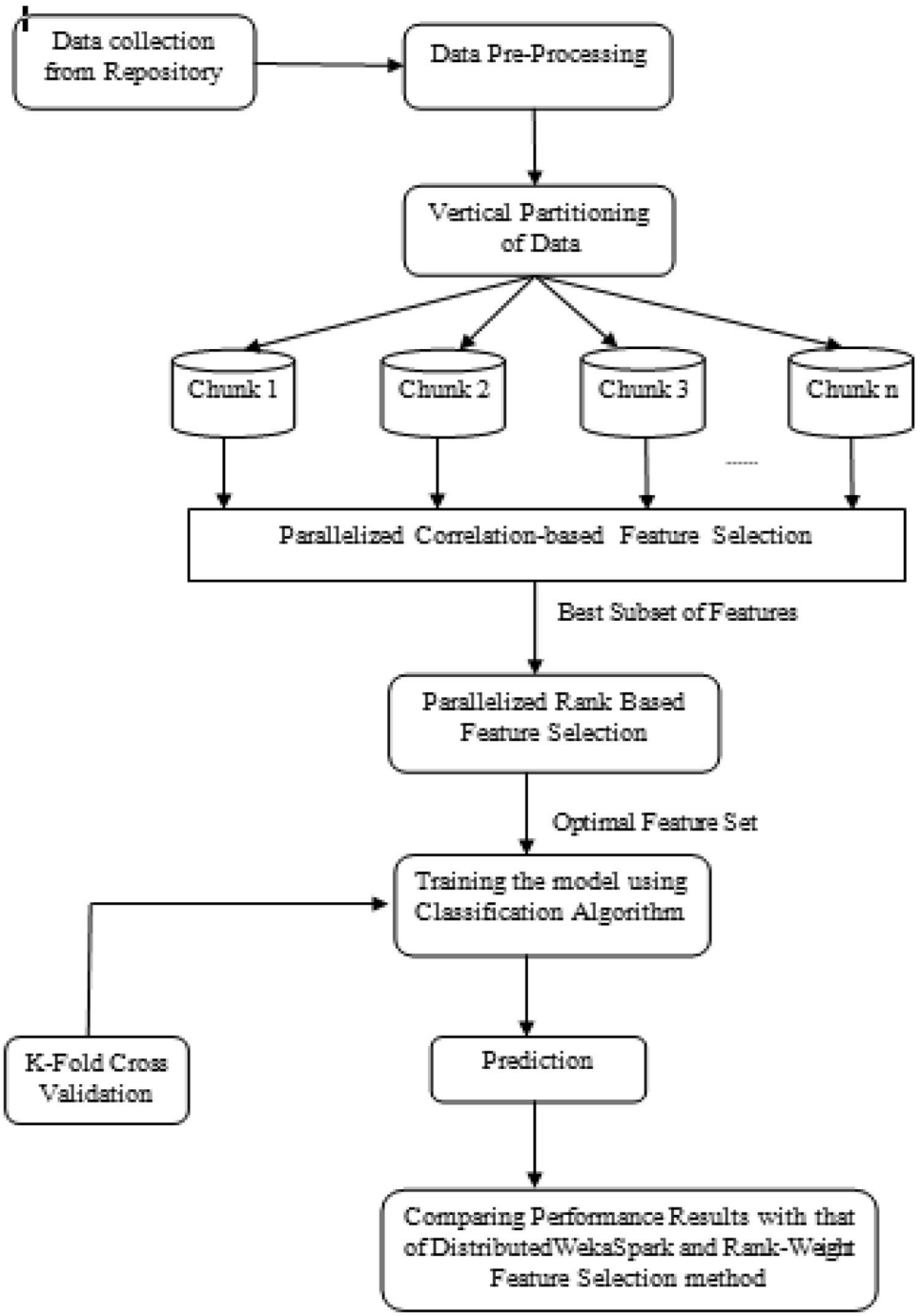

Fig. 1 Parallelized hybrid feature selection and classification on Spark

advantages from the sub-algorithms and therefore is more robust comparing with traditional approaches. The proposed parallelized HFS algorithm includes parallelized CFS and rank-based feature selection (RFS) methods.

Figure 1 depicts the proposed computational methods for classifying cancer types from MGE data. The input file is given for data pre-processing during which the input comma separated values (CSV) file is partitioned vertically and the obtained dataset is split into chunks of data. The proposed parallelized HFS algorithm (parallelized CFS and RFS methods) is applied to obtain optimal and relevant features. Parallelized tree models namely decision tree (DT) and random forest (RF) are constructed on Spark to evaluate selected features.

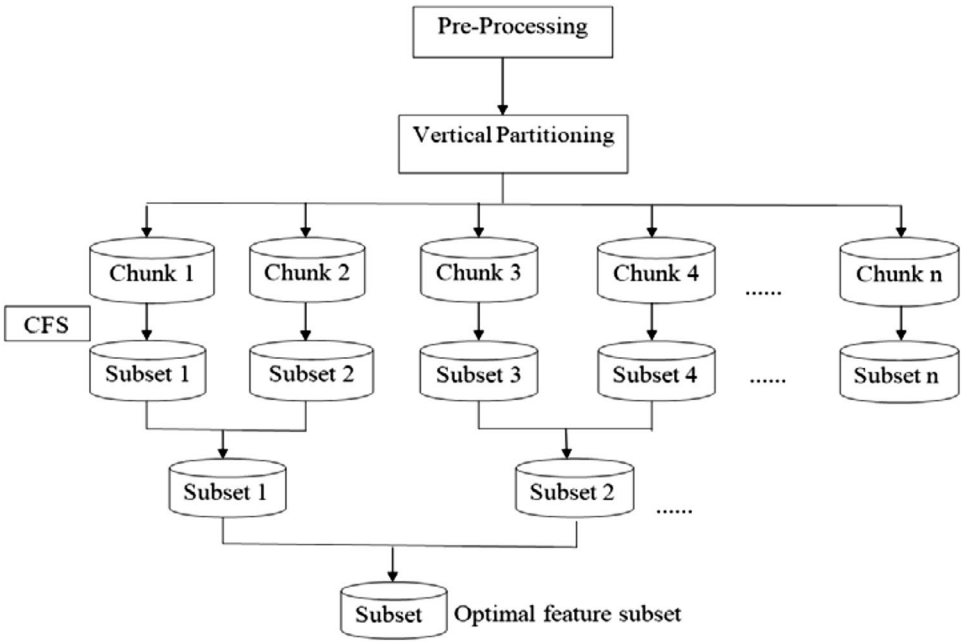

Fig. 2 Parallelized correlation-based feature subset selection

K-fold cross validation is applied to build and evaluate the constructed classifier. RF and DT are used for parallelized classification as they provide accurate results when compared to the other classification algorithms and tree models could be easily parallelized compared to other classification algorithms (Ryza et al. 2017; Spark Release 2.2.1 2019). The obtained results are compared with previously reported results from RWFS method (Ramani and Jacob 2013) as well as with results obtained from DistrbutedWekaSpark (DWS) (WEKA 2019). Parallelized HFS method includes parallelized CFS and parallelized RFS methods. The proposed parallelized HFS method was considered to be an embedded feature selection method as it combines filter (CFS) and wrapper (rank-based) methods. These methods are described in the following subsections.

Correlation-based feature subset selection (CFS)

The core idea of using feature selection is to remove the redundant or irrelevant features present in the data without incurring loss of information. Redundant or irrelevant features are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated. The CFS hypothesis (Hall 2000) suggested that the most predictive features needed to be highly correlated to the target class and least relevant to other predictor attributes. The following equation dictated the merit of a feature subset S that consisted of k features:

$$\text{Merit}\, S_k = \frac{\overline{k_{rcf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{1}$$

where, $\overline{r_{cf}}$ is the average value of all feature–class correlations, and $\overline{r_{ff}}$ is the average value of all feature–feature correlations.

The CFS criterion defined as follows

$$\text{CFS} = \text{MAX}\, S_k \frac{r_{cf1} + r_{cf2} + \cdots + r_{cfk}}{\sqrt{k + 2(r_{f1f2} + r_{f2f3} + \cdots + r_{fkf1})}}. \tag{2}$$

where, $\overline{r_{cfi}}$ and $\overline{r_{fifj}}$ variables are referred to as correlations. The attributes that portrayed a high correlation to the target class and least relevance to each other were chosen as the best subset of attributes (Yuan et al. 2017).

Normally parallel programming framework divides data horizontally. But in our data set, the number of columns (features) outnumbers the number of rows (samples). Hence, data is split vertically and feature selection is applied to each of the split part. Figure 2 details steps involved in parallelized CFS. As the dataset is split along features, the features in one chunk may be relevant to features in other chunks. Hence, features across chunks have to be analyzed to obtain differentially expressed genes. So, the feature selection algorithm is iteratively applied on the obtained feature subset from initial chunks of data until only one chunk of optimal features are found. CFS is a fully automatic algorithm. It does not require the user to specify any thresholds or the number of features to be selected, although both are simple to incorporate if desired. Here CFS is a filter, and as such does not incur the high computational cost associated with repeatedly invoking a learning algorithm.

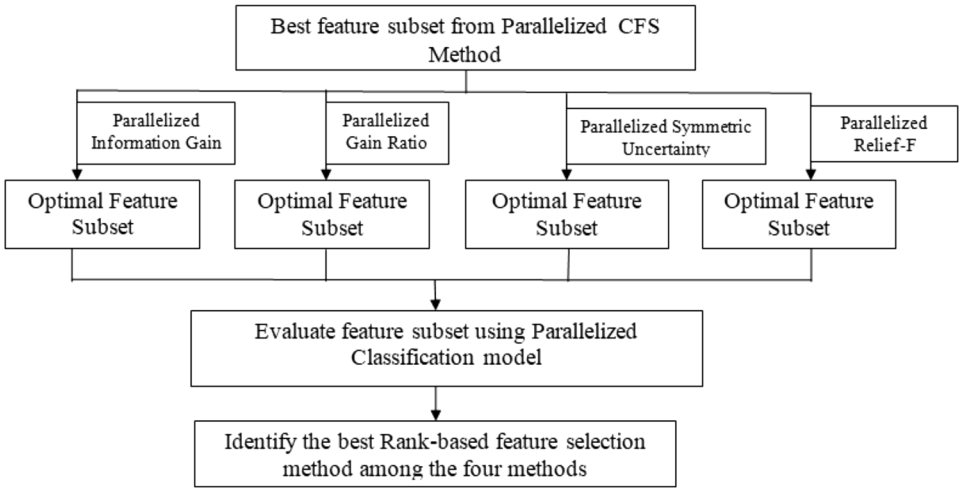

Fig. 3 Parallelized rank-based feature selection

Table 1 Microarray gene expression data description

Gene dataset	No. of genes	Total no. of samples	No. of Target classes	Class wise samples	Cancer sub-types
Gastric cancer (3 class)	4522	30	3	8 5 17	1. Normal gastric tissue 2. Diffuse gastric tumor 3. Intestinal gastric tumor
Childhood leukemia (acute lymphoblastic leukemia) (4 class)	8280	60	4	13 21 16 10	1. Mercaptopurine alone (MP) 2. High-dose methotrexate (HDMTX) 3. Mercaptopurine and low-dose methotrexate (LDMTX_MP) 4. Mercaptopurine and high-dose methotrexate (HDMTX_MP)

Parallelized rank-based feature selection algorithm

Diverse feature ranking and feature selection techniques have been proposed in the machine learning literature. The purpose of these techniques is to discard irrelevant or redundant features from a given feature vector. Rank-based method takes the output of CFS as input and four RFS algorithms are applied on it separately to produce four subsets. Four rank-based methods used are IG, gain ratio (GR), symmetric uncertainty (SU) (Ali and Shahzad 2012) and ReliefF (Wang et al. 2016), that resulted in four feature subsets as shown in Fig. 3. The best feature subset was obtained by evaluating the features using parallel classification algorithms. The parallel Rank-based method which yielded this best subset of features was considered as a best RFS method.

Materials

Microarray gene expression data for gastric cancer with three classes and childhood leukemia with four classes are extracted from Artificial Intelligence Biolabs (Bioinformatics Laboratory 2019). Table 1 portrays the description of dataset used in this research.

Algorithm: Parallelized Hybrid Feature Selection Algorithm

Input: Microarray Gene Expression Data (D) with *'m'* samples and *'n'* genes

Output: Optimal and Relevant Gene set

1. Split the dataset vertically along columns (feature-wise) such that each chunk as *'n/5'* number of genes.
2. Apply parallelized Correlation Feature Subset Selection (CFS) on each chunk of data.
   for i = chunk_1 to chunk_n do
   2.1 Merit $S_{ki} = \frac{\overline{k_{rcf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}$
   2.2 Best FS_i = MAX $S_k \frac{r_{cf1}+r_{cf2}+\ldots+r_{cfk}}{\sqrt{k+2(r_{f1f2}+r_{f2f3}+\ldots+r_{fkf1})}}$
3. Obtain best feature subset from each chunk of data.
4. Apply parallelized CFS recursively across chunks to capture inter-chunk correlation of feature subset until only one chunk of optimal and minimal features are obtained.
5. Following parallelized CFS, parallelized Rank-based feature selection methods are to be applied.
6. Apply parallelized Information Gain, Gain Ratio, Symmetric Uncertainty and ReliefF separately on the features obtained from parallelized CFS.
   6.1 **Information Gain:**

$$Information\ Gain\ (A) = Info_A - Info_A(D)$$
$$Info_D = -\sum_{i=1}^{m} p_i log_2 p_i$$

   6.2 **Gain Ratio:**

$$GainRatio_A = \frac{Gain_A}{SplitInfo_A}$$
$$SplitInfo_A(D) = --\sum_{j=1}^{v} \frac{|D_j|}{|D|} x log_2\left(\frac{|D_j|}{|D|}\right)$$

   6.3 **Symmetric Uncertainty:**

$$SU\ [X, Y] = 2\left[\frac{IG(X/Y)}{H(X)+H(Y)}\right]$$

   6.4 **ReliefF:**

   It evaluates the worth of a feature by repeatedly sampling an instance and considering the value of the given feature for the nearest instance of the same and different class. It evaluates the worth of a feature by repeatedly sampling an instance and assigns a weight to each feature based on the ability of the feature to distinguish among the classes.

7. Number of features to be selected is decided using embedded feature selection approach that combines wrapper and filter-based approaches.
8. Evaluate the feature set by constructing learning models namely parallelized Decision Tree and Random Forest.
9. Feature-set which offer highest prediction accuracy is to be chosen as Optimal and Relevant Feature Subset.
10. Return Optimal and Relevant Feature Subset.

Results

Parallelized CFS was executed on Spark that results in a subset of features. The resultant subset of features was given as input to rank-based methods (i.e., IG, GR, SU and ReliefF) separately. Number of features to be selected was specified as input to rank-based methods. This results in an optimal feature subset. For example, the number of features selected by CFS for childhood leukemia dataset is 17 in 44 s as shown in Table 2. The number of features given as input to rank-based methods was varied from 1 to 17 and the subset that gave highest accuracy when classified was taken as best feature subset. Similarly

Table 2 Parallelized CFS results on DWS and Spark

Dataset	No. of genes	No. of instances	Parallelized CFS on DWS		Parallelized CFS on Spark	
			No. of genes selected by CFS	Execution time (in s)	No. of genes selected by CFS	Execution time (in s)
Gastric cancer	4524	30	52	1080	25	35
Childhood leukemia	8281	60	39	2160	17	44

Table 3 Parallelized classification on DWS

Dataset	Ranking method	Cross-validation					
		Decision tree			Random forest		
		No. of features selected	Accuracy (in %)	Execution time (in s)	No. of features selected	Accuracy (in %)	Execution time (in s)
Childhood leukemia	Information gain	15	34.42	8	**15**	**85.25**	**9**
	Gain ratio	10	34.42	8	10	85	9
	Symmetric uncertainty	25	34.42	8	25	85.25	5
	ReliefF	25	34.42	8	25	85.25	5
Gastric cancer	Information gain	45	54.83	3	**45**	**96.77**	**5**
	Gain ratio	45	54.83	1	**45**	**96.77**	**6**
	Symmetric uncertainty	52	54.83	3	52	96.77	4
	ReliefF	52	54.83	3	52	95	5

for gastric cancer 25 features are selected as best feature subset in 35 s. Forward selection approach was followed to select optimal features from the result of parallel CFS method. Classification model was evaluated using percentage split and K-fold cross validation. In percentage split evaluation method, 70% of original data was used for training and remaining 30% as test data.

Tables 3 and 4 depict the parallelized classification accuracy when DT and RF are constructed using DistributedWekaSpark (DWS) and Spark, respectively. As shown in Table 3, the parallelized HFS approach with RF on DWS yielded classification accuracy of 85.25% and 96.77% for childhood leukemia and gastric cancer, respectively. The parallelized HFS approach with decision tree on DWS yielded classification accuracy of 34% and 54% for childhood leukemia and gastric cancer, respectively. The decision tree results were very poor compared to RF which is an ensemble classifier. The highest accuracy obtained from four rank-based feature selection methods are indicated with bold face in Tables 3 and 4.

Similarly on Spark framework, parallelized IG method was applied on the CFS result of childhood leukemia, it yielded maximum accuracy when the output number of features was set to 11 with respect to RF classifier. There was 79.14% classification accuracy with percentage split of 70% training data and 30% testing data when parallel ReliefF was used as feature selection algorithm. Its classification accuracy was 65.64% and execution time was 22.98 s when the classifier was build on Spark and evaluated using five-fold cross validation. Similarly, for Gastric cancer, parallelized ReliefF with RF produced classification accuracy of 97.22% in 30 s which was depicted in Table 4. The constructed RF was evaluated with nine-fold cross validation method. The low accuracy results for childhood leukemia were due to the fact that the distinctive characteristics between the different sub-types are very minimal. The sub-types of childhood leukemia were classified as only low and high dose mercaptopurine with methotrexate. Parallel decision tree classifier yielded low classification accuracy as ~ 49% to ~ 59% for childhood leukemia and it was ~ 88% to ~ 91% for gastric cancer. Table 5 lists the differentially expressed genes that are optimal and relevant obtained from parallelized HFS method and resulted in maximum classification accuracy.

Thus, it can be inferred that using parallelized HFS and RF method gave 65.64% accuracy for childhood leukemia and 97.22% for Gastric cancer with very less execution time compared to results from DistributedWekaSpark. Even though the class-wise samples are very few which is the basic nature of MGE data, the objective is to provide a parallelized framework that could work at the dimension-wise and classify cancer sub-types in very less time without any obstacle on prediction accuracy.

Figure 4a, b represent the receiver operating characteristics area under curve (ROC AUC) (Bang et al. 2017) for RF classifier obtained when applied on gastric cancer and

Table 4 Parallelized classification on Spark

Dataset	Ranking method	Percentage split						Cross-validation		
		Decision tree			Random forest			Random forest		
		No. of features selected	Accuracy (in %)	Execution time (in s)	No. of features selected	Accuracy (in %)	Execution time (in s)	No. of features selected	Accuracy (in %)	Execution time (in s)
Childhood leukemia	Information gain	8	54.03	4.53	13	51.97	7.17	**11**	**65.64**	**22.98**
	Gain ratio	6	59.59	4.86	7	63.95	6.95	12	60.76	23.22
	Symmetric uncertainty	6	59.09	4.83	8	52.43	7.08	16	60.76	23.9
	ReliefF	15	49.57	4.98	**11**	**79.14**	**7.84**	14	54.14	23.63
Gastric cancer	Information gain	23	91.6	3.73	21	94.87	5.24	22	93.52	31.02
	Gain ratio	20	88.8	3.8	21	96	5.81	19	97.22	30.57
	Symmetric uncertainty	19	88.42	4.16	17	96	5.77	19	97.22	30.63
	ReliefF	21	90.9	4.22	17	95.8	5.66	**19**	**97.22**	**30.1**

childhood leukemia datasets with the threshold values 0.6, 0.7, 0.8 and 0.9. ROC AUC is the curve obtained by plotting false positive rate (FPR) against true positive rate (TPR). The highest TPR obtained for gastric cancer and childhood leukemia was 0.9 and 0.83, respectively.

Comparison of results with that of DistributedWekaSpark (DWS), sequential rank-weight feature selection (RWFS) and parallelized recursive feature selection (RFS)

The results obtained by executing parallelized HFS on Spark are compared with that obtained by executing parallelized HFS (Das et al. 2017; Lee and Leu 2017) on DistributedWekaSpark (DWS) and RWFS method (Ramani and Jacob 2013). In RWFS method, all gene expression datasets contained absolute values. In order to identify the most relevant genes for classification, six feature selection algorithms viz, fuzzy rough set evaluator with best first search approach and attribute evaluators that ranked the features based on the IG, SU, Chi-Square co-efficient, ReliefF factor and the GR were utilized (Eiras-Franco et al. 2016). The minimal feature subset returned by all the six feature selection algorithms were then compared to determine the genes that were commonly reported by all the feature selection techniques. The selected features were evaluated using ten classifier models.

Figure 5 depicts the results of proposed parallelized HFS with RF compared with results from DWS. Figure 6 depicts the comparison of proposed method with the results of RWFS method (Ramani and Jacob 2013). In HFS, the value of K used for K-fold cross validation for childhood leukemia dataset was 5 whereas it was 9 for gastric cancer dataset.

The main theme of the proposed work is to apply parallel computational methods and improve classification accuracy with high speed up. DisrtibutedWekaSpark (DWS) is a distributed computational tool that combines the complementary properties of Weka (Hall et al. 2009) and Spark into one unit. The reason behind comparison of proposed parallelized HFS and RF method on Spark with DistributedWekaSpark is due to the fact that the existing methods in the DWS tool performed inferior than the proposed parallelized HFS method for gastric cancer. DistributedWekaSpark took more execution time compared to proposed parallelized HFS method which was evident from results in Fig. 5. DWS yielded 95% accuracy in 1 h whereas the proposed parallelized HFS method yielded 97% in 60 s. This justifies the necessity of parallelized framework to select optimal number of genes and accurately classify the sub-types of cancer in very less time without any adverse effect on classification accuracy.

Parallelized HFS improved accuracy by 14% when compared to existing rank-weight feature selection method in

Table 5 Optimal and relevant feature set

MGE dataset	Parallelized feature selection method	No. of optimal features	Optimal and relevant geneset
Gastric cancer	Correlation feature selection and ReliefF	19	AC002077_at, U33286_at, D78134_at, U13737_at, Y10032_at, U96915_at, U50360_s_at, X99584_at,U46767_at, L27706_at, U05681_s_at, D90276_at, X76717_at, V00572_at, X83416_s_at, S85655_at, U79241_at, D55716_at, D87445_at
Childhood leukemia	Correlation feature selection and information gain	11	39867_at, 41168_at, 31506_s_at, 37529_at, 37888_at, 38555_at, 37553_at, 33670_at, 35727_at, 41159_at, 37721_at

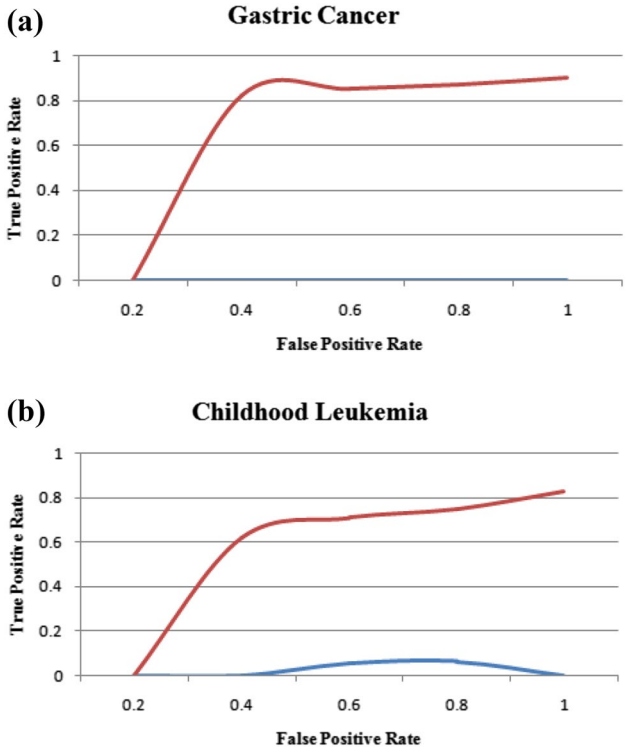

Fig. 4 ROC AUC for MGE data; **a** gastric cancer, **b** childhood leukemia

literature (Ramani and Jacob 2013) which was 65% and it was low when compared to results from DWS for childhood leukemia dataset. While for gastric cancer dataset, the parallelized HFS method improved accuracy by 2% with very minimal number of features when compared to that of DistributedWekaSpark. Parallelized HFS method improved accuracy by 4% when compared to RWFS. RWFS is a sequential computational method. On the other hand, there was drastic decrease in execution time. Parallelized HFS on DWS took 2160 s for gastric cancer and 1080 s for childhood leukemia, whereas Parallelized HFS on Spark took only 60 and 51 s, respectively. Comparing execution time of RWFS method, it took 1800 s for gastric cancer and 900 s for childhood leukemia which is higher than the execution time of parallelized HFS on Spark.

Table 6 displays the performance of proposed parallelized HFS method compared with parallelized RFS method in previous work (Venkataramana et al. 2018). The classification accuracy improved to ~ 5% and ~ 15% for gastric cancer and childhood leukemia respectively with parallelized HFS and RF when compared with parallelized RFS and RF. The execution time of proposed parallelized HFS was higher compared to previous work as it applies both parallelized CFS and RFS methods to select optimal number of genes.

Discussions

The parallel computational methods are necessary to speed up the computational task and produce results in less time. But this greatly affects the accuracy of results (Lokeswari et al. 2019). The proposed parallelized HFS method plays a vital role in selecting only the optimal number of predictive genes from high dimensional MGE data to accurately classify sub-types of cancer in very less time. The number of genes in MGE data is in thousands, this huge number of genes may mislead the classifier model due to presence of redundant and irrelevant genes. Hence, parallelized HFS method was proposed in the current research work to select differentially expressed genes and accurately classify cancer sub-types in less time. The classification accuracy obtained was 97% for gastric cancer and 79% for childhood leukemia in 60 and 51 s, respectively. The parallel feature selection (RFS) method in our previous work (Venkataramana et al. 2018) yielded 92% and 64% classification accuracy for gastric cancer and childhood leukemia respectively. The sequential feature selection (RWFS) method in the existing work (Ramani and Jacob 2013) yielded 93% and 65% classification accuracy. Although the existing parallel and sequential feature selection methods in literature produced better classification accuracy, sequential methods took ~ 3 and ~ 5 min for gastric cancer and childhood leukemia respectively which is high compared to time taken by

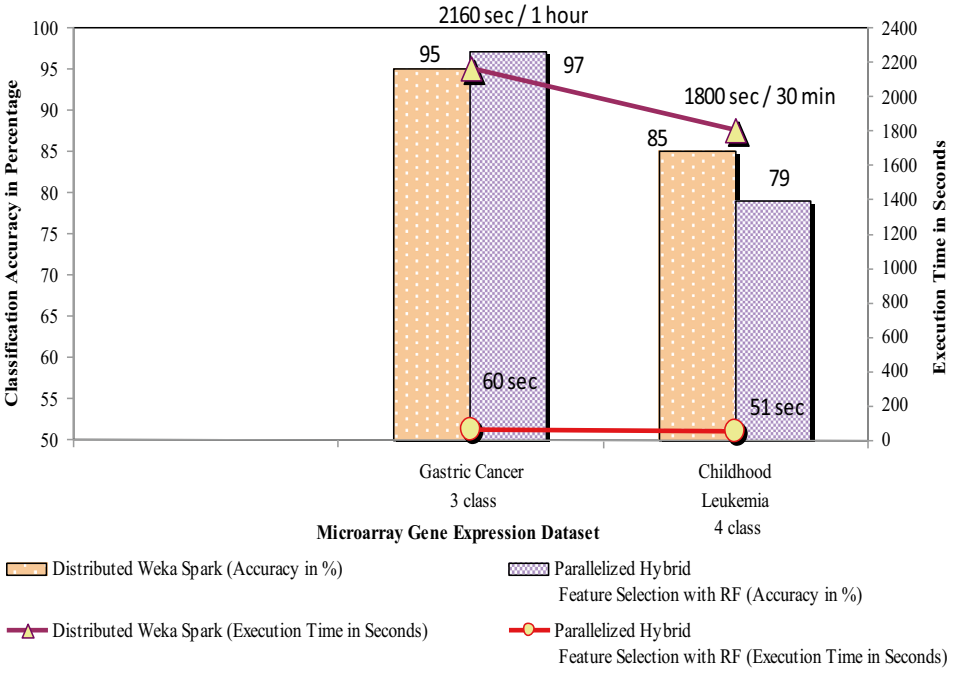

Fig. 5 Comparison of results of HFS on DWS and Spark

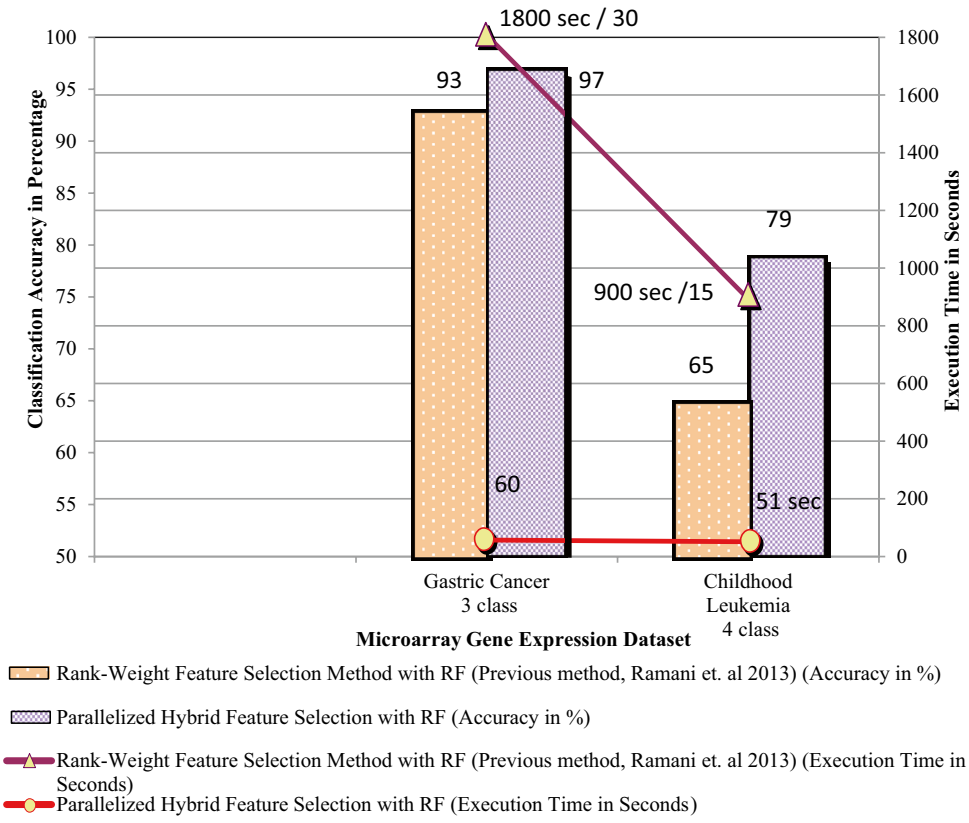

Fig. 6 Comparison of results of RWFS and HFS on Spark

parallelized methods. Compared with previous feature selection methods parallelized HFS performed better in terms of accuracy and speed up.

Moreover, genomic data are ever increasing and hence scalable parallelized computational methods are essential to analyze voluminous data in order to unearth significant genes that play a pivotal role in classification of disease.

Table 6 Comparison of parallelized HFS and random forest with previous work

MGE dataset	Parallelized recursive feature selection with RF (Venkataramana et al. 2018)		Parallelized hybrid feature selection with RF	
	Accuracy in %	Execution time (in s)	Accuracy in %	Execution time (in s)
Gastric cancer (3 class)	92	35	97	60
Childhood leukemia (4 class)	64	44	79	51

The objective of parallel feature selection methods is to choose less number of predictive genes (features) and improve classification accuracy. The challenge with microarray data is dealing with low-sample and high-dimensional nature of it. Several authors in literature have applied feature selection on microarray data which was detailed as follows.

The colon cancer dataset with 62 × 2000 dimensions yielded 88% classification accuracy with Chi square test and decision tree. The classification accuracy obtained was 92% with SU and Naïve Bayes (Wang and Gotoh 2010). The number genes selected was 100.

The lung cancer dataset with 203 × 12,600 dimensions yielded 93% accuracy with 700 genes from parallel Chi square test and RF classifier (Zhang et al. 2014). The Glioblastoma dataset with 50 × 12,625 dimensions yielded 82% accuracy with 700 genes from parallel Chi square test and RF classifier. It resulted in 96% accuracy with 66 genes from parallel RFS and RF classifier (Venkataramana et al. 2018).

Never the less, when parallelized HFS method was applied on Lung Cancer and Glioblastoma datasets, the classification accuracy was improved by ~2% to ~3%. Despite, the less number of samples and dimensions in the considered dataset (30 × 4522 and 60 × 8280), the proposed method could be applied for any growing samples and dimensions of any health-care data. So, it is emphasized that the proposed parallelized HFS method does not result in over fitting for the given dataset. It was identified from the investigations on microarray data that accuracy increases with growing number of samples and dimensions. It is also inferred that accuracy decreases with decrease in number of samples and dimensions. The authors in the current research have chosen 30 to 60 samples, which is the worst case condition; still the classification accuracy is maintained or improved with the proposed parallelized HFS method. This proves the robust nature of the proposed method. The proposed parallelized HFS method could scale for even large sized gene, protein expressions and RNA sequence (TCGA) data. Hence, the proposed method is the generalized approach for any sized dataset.

Conclusion

Parallelized HFS algorithm was used in order to improve classification accuracy of cancer types. HFS as in detail involves parallelized CFS that selects optimal features and parallelized RFS methods that rank the features and select only the important features. Parallelized classifier model was build and selected features were evaluated using percentage split and k-fold cross validation methods. This improved accuracy by a great margin and at the same time greatly reduces execution time. For childhood leukemia dataset, parallelized HFS gave accuracy of 79% which was ~ 14% higher when compared to existing RWFS method which produced 65% and it was low when compared to results from DWS. While for Gastric cancer dataset, there was ~ 2% improvement in accuracy by HFS when compared to that of DWS and it also selected very minimal number of features as predictive features. Parallelized HFS method improved accuracy by ~ 4% when compared to RWFS. Parallelized HFS yielded ~ 5% and ~ 15% improvement in classification accuracy when compared with parallelized RFS method. As part of our future work, deep neural network algorithms can be explored on MGE and Next Generation Sequencing datasets for cancer prediction.

Acknowledgements This research work is part of project work funded by Science and Engineering Research Board (SERB), Department of Science and Technology (DST) funded project under Young Scientist Scheme—Early Start-up Research Grant- titled "Investigation on the effect of Gene and Protein Mutants in the onset of Neuro-Degenerative Brain Disorders (Alzheimer's and Parkinson's disease): A Computational Study" with Reference no-SERB—YSS/2015/000737/ES.

Compliance with ethical standards

Conflict of interest Lokeswari Venkataramana, Shomona Gracia Jacob, Rajavel Ramadoss, Dodda Saisuma, Dommaraju Haritha and Kunthipuram Manoja declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent is not necessary as this article does not involve human or animal participants.

References

Ali SI, Shahzad W (2012) A feature subset selection method based on symmetric uncertainty and ant colony optimization. In: IEEE international conference on technologies (ICET), pp 1–6

Alshamlan HM, Badr GH, Alohali Y (2013) A study of cancer microarray gene expression prole: objectives and approaches. In: Proceedings of the world congress on engineering, vol 2, pp 1–6

Bang MS, Kang K, Lee JJ, Lee YJ, Choi JE, Ban JY, Oh CH (2017) Transcriptome analysis of non-small cell lung cancer and genetically matched adjacent normal tissues identifies novel prognostic marker genes. Genes Genom 39(3):277–284

Bioinformatics Laboratory (2019). http://www.biolab.si/supp/bi-ancer/projections/info/ALLGSE412_poterapiji.html. Accessed 20 July 2019

Bolón-CanedoV V, Sánchez-Maroño N, Alonso-Betanzos A (2015) Distributed feature selection: an application to microarray data classification. Appl Soft Comput 30:136–150

Chuang LY, Yang CH, Wu KC, Yang CH (2011) A hybrid feature selection method for DNA microarray data. Comput Biol Med 41(4):228–237

Das AK, Goswami S, Chakrabarti A, Chakraborty B (2017) A new hybrid feature selection approach using feature association map for supervised and unsupervised classification. Expert Syst Appl 88:81–94

Eiras-Franco C, Bolón-Canedo V, Ramos S, González-Domínguez J, Alonso-Betanzos A, Touriño J (2016) Multithreaded and Spark parallelization of feature selection filters. J Comput Sci 17:609–619

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Gracia Jacob S (2015) Discovery of novel oncogenic patterns using hybrid feature selection and rule mining. Ph.D. Thesis. Anna University. India

Hall MA (2000) Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of the seventeenth international conference on machine learning, pp 359–366

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11(1):10–18

Heo J, Lee JS, Leem SH (2013) Distinct gene expression signatures during development of distant metastasis. Genes Genom 35(4):511–522

Kang S, Hong S (2011) Prediction of personalized drugs based on genetic variations provided by DNA sequencing technologies. Genes Genom 33(6):591–603

Lee CP, Leu Y (2017) A novel hybrid feature selection method for microarray data analysis. Appl Soft Comput 11(1):208–213

Li J, Liu H (2017) Challenges of feature selection for big data analytics. IEEE Intell Syst 32(2):9–15

Lokeswari YV, Jacob SG, Ramadoss R (2019) Parallel prediction algorithms for heterogeneous data: a case study with real-time big datasets. In: Peter JD, Alavi AH, Javadi B (eds) Advances in big data and cloud computing. Springer, Singapore, pp 529–538

Lu H, Chen J, Yan K, Jin Q, Xue Y, Gao Z (2017) A hybrid feature selection algorithm for gene expression data classification. Neurocomputing 256:56–62

Peralta D, del Río S, Ramírez-Gallego S, Triguero I, Benitez JM (2015) Herrera F (2015) Evolutionary feature selection for big data classification: a Mapreduce approach. Math Probl Eng 2015(246139):1–11

Ramani RG, Jacob SG (2013) Benchmarking classification models for cancer prediction from gene expression data: a novel approach and new findings. Stud Inform Control 22(2):134–143

Ryza S, Laserson U, Owen S, Wills J (2017) Advanced analytics with Spark: patterns for learning from data at scale. O'Reilly Media Inc., Northern California, USA

Singh RK, Sivabalakrishnan M (2015) Feature selection of gene expression data for cancer classification: a review. Procedia Comput Sci 50:52–57

Spark Release 2.2.1—Apache Spark (2019). https://spark.apache.org/releases/spark-release-2-2-1.html. Accessed 25 July 2019

Venkataramana L, Jacob SG, Ramadoss R (2018) Parallelized classification of cancer sub-types from gene expression profiles using recursive gene selection. Stud Inform Control 27(1):215–224

Waikato Environment for Knowledge Analysis (WEKA) (2019). http://weka.sourceforge.net/packageMetaData/distributedWekaSpark/index.html. Accessed 26 July 2019

Wang X, Gotoh O (2010) A robust gene selection method for microarray-based cancer classification. Cancer Inform 9:CIN-S3794

Wang Z, Zhang Y, Chen Z, Yang H, Sun Y, Kang J, Yang Y, Liang X (2016) Application of ReliefF algorithm to selecting feature sets for classification of high resolution remote sensing image. In: 2016 IEEE international geoscience and remote sensing symposium (IGARSS), pp 755–758

Yu JF, Guo J, Liu QB, Hou Y, Xiao K, Chen QL, Wang JH, Sun X (2015) A hybrid strategy for comprehensive annotation of the protein coding genes in prokaryotic genome. Genes Genom 37(4):347–355

Yuan M, Yang Z, Huang G, Ji G (2017) Feature selection by maximizing correlation information for integrated high-dimensional protein data. Pattern Recognit Lett 92:17–24

Zhang H, Li L, Luo C, Sun C, Chen Y, Dai Z, Yuan Z (2014) Informative gene selection and direct classification of tumor based on chi square test of pairwise gene interactions. Biomed Res Int 2014(589290):1–9