Research Article

Shemim Begum*, Ram Sarkar, Debasis Chakraborty, and Ujjwal Maulik

Identification of Biomarker on Biological and Gene Expression data using Fuzzy Preference Based Rough Set

https://doi.org/10.1515/jisys-2019-0034
Received Jan 29, 2019; accepted Oct 09, 2019

Abstract: Cancer is fast becoming an alarming cause of human death. However, it has been reported that if the disease is detected at an early stage, diagnosed, treated appropriately, the patient has better chances of survival long life. Machine learning technique with feature-selection contributes greatly to the detecting of cancer, because an efficient feature-selection method can remove redundant features. In this paper, a Fuzzy Preference-Based Rough Set (FPRS) blended with Support Vector Machine (SVM) has been applied in order to predict cancer biomarkers for biological and gene expression datasets. Biomarkers are determined by deploying three models of FPRS, namely, Fuzzy Upward Consistency (FUC), Fuzzy Downward Consistency (FLC), and Fuzzy Global Consistency (FGC). The efficiency of the three models with SVM on five datasets is exhibited, and the biomarkers that have been identified from FUC models have been reported.

Keywords: FUC, FLC, FGC, Biomarkers, FPRS

1 Introduction

Classification of cancer, based on gene expression data, has become an attractive research area in the field of bioinformatics. Currently, diagnosis through the recent technique, Fine Needle Aspiration Cytology (FNAC) [1], is not up to the mark, because it has been reported that it does not possess high-quality diagnostic capability. Numerous studies about cancer classification exist. The methods cover Principle Component analysis (PCA) [2, 3], relief, mutual information and information gain [4], FPRS, and the like. From among these, FPRS is the most commonly used feature-selection (FS) method, because its reasoning methods are simpler than a computationally precise system. This results in the conservation of computational power, which is an interesting feature in real-time systems. Classification is performed by choosing the most compelling genes in order to construct a good classification model. In addition, identifying significant genes greatly reduces the run time in the designing of a good classification model. Besides, extracting significant genes from thousands of genes is a critical issue.

A complete review of the FS method has been described in [5]. Depending on the interaction of genes with the classification model, the FS method can be classified into three classes: filter, wrapper, and embedded methods. Filter methods [6] evaluate the relevance of features by observing the intrinsic properties of the data. These methods are fast and computationally less expensive than the wrapper method. The wrapper method

***Corresponding Author: Shemim Begum:** Govt College of Engg. & Textile Technology, Dept. of CSE, Berhampore, Murshidabad, West Bengal, India; Email: shemim_begum@yahoo.com
Ram Sarkar: Jadavpur University Ringgold standard institution – Computer Science and Engineering, India; Email: ramjucse@gmail.com
Debasis Chakraborty: MCET – ECE, Berhampore, West Bengal, India; Email: debasismcet@yahoo.in
Ujjwal Maulik: Jadavpur University Ringgold standard institution – CSE, Kolkata, West Bengal, India; Email: ujjwal_maulik@yahoo.com

[7] selects a subset of features by observing the performance of a classification model. However, this model has a great chance of over fitting and is computationally more expensive than the filter method. In contrast, the internal parameter of the classification model has great use in the case of the embedded method [8], which greatly reduces the computational cost of the FS method. In this context, the FS technique and the supervised classifier play an important role in selecting significant gene markers for cancer diagnosis. These techniques avoid the error that is reported by FNAC, because they examine the dataset in a less amount of time.

Several researches to identify genes have been carried out. In [9], researchers tried to introduce the distributed FS method by using symmetrical uncertainty and the Multi-Layer-Perceptron (MLP) classifier. Features are distributed across multiple clusters. The classifier MLP is applicable in each cluster and the cluster with the highest classification accuracy and lowest root means square error is nominated. Local search methods, sequential backward selection, and maximization of mutual information are applied on gene expression datasets to obtain the most relevant and non-redundant genes, and it is presented in [10]. Thereafter, global optimization techniques such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO) [11] are used to improve the classification accuracy and computational time. Also in [12] imperialist Competition algorithm, wavelet least squares support vector regression and imperialist competition algorithm is applied for assessing the incipient faults of transformer polymer Insulation. In [13], Binary Differential Evolution (BDE) algorithm with a rank-based filter FS method, BDEX_{Rank} and BDEX_{Rankf}, are applied in gene expression datasets. Both algorithms join the filter with the wrapper in a two-stage algorithm. In [14], Binary Particle Swarm Optimization (BPSO) encoding gene that classes sensitivity is used to extract the important features from five microarray datasets. Here, Extreme Learning Machine (ELM) functions as the classifier. Lately, bi-clustering has been very popular in the identifying of relevant genes in gene expression data. In [15], the bi-clusters algorithm with the use of formal concept analysis has shown to be an efficient methodology for bi-clustering binary data.

The said algorithm is applied to recognize groups of genes that are relevant under a subset of samples. The article [16] demonstrated a deterministic initialization algorithm for the kmeans algorithm by identifying a set of clusters through the bi-partitioning approach. The results are promising in terms of computational time and stable convergence. In this paper, SVM [17] coupled with FPRS [18] is employed for informative gene selection from the datasets, and to diagnose the cancer. The kernel implicitly in SVM contains a nonlinear transformation, no assumptions about the functional form of the transformation, which makes data linearly separable. In this way SVM extends itself in order to classify the linearly inseparable data. Choosing an appropriate value for the parameters, SVMs can be robust, even when the training sample has some bias. Rough set model is designed based on the equivalence relations. The equivalence relations have one of the main drawbacks, when applying this model on complex decision making problem. Whereas, fuzzy preference relation focused the degree of preference quantitavely making it more powerful to collect information from fuzzy data than equivalence relation. This inspire us to use FPRS technique for gene selection. The details of three models of the FPRS method are explained in [19]. In this literature, we have considered Wisconsin Breast Cancer Dataset (WBCD), leukaemia, prostate, Diffuse Large B-cell Lymphomas (DLBCL), and Mixed-Lineage Leukaemia (MLL) datasets to conduct the experiment. Cancer research is biological. The observed results show that our proposed method selects significant genes for cancer classification. The statistical analysis of any given dataset is a new concept, which enlightens cancer research further. Research on the application of a data mining technique to diagnose the outcome of the disease is encouraging news recently [20].

2 Background

2.1 Support Vector Machine

SVM is a well-known, high-performance, supervised, learning algorithm [21], which follows the concept of the statistical learning theory [22]. Currently, it is being applied in many fields including remote sensing [23], biological data analysis [24], natural language processing, marine appliances, and so on. In the case of lin-

early separable binary cases, the goal is to design a hyper plane that will classify all training vectors in two classes. The best selection will be the hyper plane that maintains the maximum margin from both classes. When the data are not linearly separable [25], SVM performs a mapping to transfer the data from an input space to a higher dimensional feature space. The upper bound of the generalization error can be reduced by providing the largest possible distance between the separating hyper plane and the sample on either side of it. The decision function of SVM is denoted by $f(x) = (\omega, \varphi(x) + b)$.

Here, $\varphi(x)$ denotes the mapping of the sample, x, from an input space to a higher dimensional feature space. The SVM finds an optimal separating hyper plane by maximizing the separating margin between the two classes of data.

The following optimization can be solved to provide optimal value of ω and b.

Minimize

$$g(w, \xi_i) = \frac{1}{2} ||w| |^2 + c \sum_{i=1}^{i=N} \xi_i \tag{1}$$

and

$$y\left(w, \phi(x_i)\right) \geq 1 - \xi_i, \quad \text{where } \xi_i \geq 0.(1) \tag{2}$$

Here, C is the regularization parameter, and ξ_i is the ith slack variable. The details of SVM have been elaborated in [25].

2.2 Feature-selection

Feature-selection methods can be applied to maximize the performance of the classifier on cancer data analysis [26]. It is a useful technique in dealing with dimensionality reduction. In classification, it is used to find an optimal subset of relevant features so that the overall accuracy is increased, while the size of the dataset is reduced. When a classification problem is defined by features, the number of features can be quite large, many of which can be irrelevant. A relevant feature can increase the performance of a classifier, while an irrelevant feature can deteriorate it. Therefore, in order to select the relevant features, it is necessary to measure the appropriateness of selected features by using a feature-selection criterion. For example, the clinical classification accuracy of a dermatologist in the diagnosis of malignant melanomas is between 65% and 85%; whereas, after the application of the feature-selection algorithm, the accuracy increases to more than 95% in automated skin tumor identification systems [27]. Normally, the feature values are real and categorical. Hence, feature-selection algorithms such as the traditional Rough Set (RS) theory encounter problems. This problem can be efficiently tackled by using a fuzzy rough set theory, in which the membership values are set in the range [0, 1]. This allows for a higher degree of elasticity than that in the crisp rough set theory, which deals with only one and zero membership values.

Another kind of feature-selection algorithm deals with continuous decision output to handle regression problems. FPRS can also be applied to classification as well as to regression problems.

Hence, in our proposed methodology, we have exploited FPRS for feature (gene) selection. Finding informative genes by using FS [28] algorithms is an important aspect of diagnosing cancer. Initially, the dimension is reduced in order to reduce the computational cost [29]. Overall, noise is reduced to improve prediction accuracy.

3 Methodology and metrics

3.1 Fuzzy Preference Based Rough Set

Analyzing preference is a challenging task in the decision-making process. When fuzzy preference relation is blended with the rough set, it is known as a fuzzy rough set. In this paper, we have used the FPRS method

to identify the preference relation in order to aggregate the features [30]. To evaluate the robustness of our proposed methodology, we have conducted our experiments on WBCD, leukaemia, prostate, DLBCL and MLL datasets, respectively. Noteworthy genes are selected by the three selectors (FUC, FLC, and FGC) of the FPRS feature-selection method. Initially, the datasets are partitioned into 50% for training and 50% for testing. Gene selection from three models of FPRS is performed on the 50% training set. Subsequently, whole datasets are split into three training test set partitions, which are 80–20%, 70–30%, and 50–50%, respectively. SVM blended with three models of FPRS is employed on each partition of five datasets to justify the performance of our proposed technique.

Our proposed methodology is depicted in Figure 1.

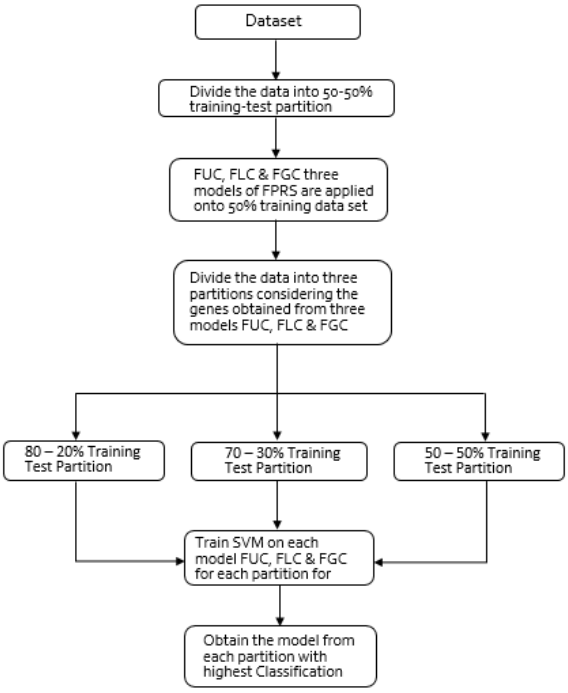

Figure 1: Overall Procedure of the proposed technique

Fuzzy preference relation can be explained as: A fuzzy product set, $v \times v$, which is described by a membership function, $\mu_r : v \in [0, 1]$. The fuzzy preference relation can be expressed by an $m \times m$ matrix $(r_{ij})_{m\times m}$, in which r_{ij} represents the preference of r_i over r_j. $r_{ij} = 1/2$ demonstrates that r_i and r_j are the same, $r_{ij} > 1/2$ demonstrates that r_i is preferred to r_j, and $r_{ij} = 1$ demonstrates that r_i is preferred to r_j. Meanwhile, $r_{ij} > 1/2$ indicates that r_j is preferred tor_i. Here, the preference matrix $r_{ij} + r_{ji} = 1$, $\forall i, j \in \{1 \dots m\}$, in which the cardinality of v is finite. We consider v as being a universe of finite numbers of the object, $v = \{a_1, a_2 \dots a_m\}$. The feature value of any object, a, is denoted by $f(a, I)$, where I is the feature of the object. The upward and downward fuzzy preference relations over v are as follows:

$$r_{ij} \geq \frac{1}{1 + e^{-\beta(f(a_i,l)-f(a_j,\ l))}}$$

and

$$r_{ij} \leq \frac{1}{1 + e^{-\beta(f(a_i,l)-f(a_j,\ l))}}$$

Here β is a positive constant. In FPRS, we are required to know the decision or conclusion with respect to available criteria. We can also know which criteria are important to decision-making, as well as the criteria that are unwanted or redundant. We can get an idea about, which criteria are important for decision making; also which criteria are unwanted or redundant.

Currently, researchers mostly use the RS theory as a numerical tool to cope with the problem of uncertainty and incompleteness. In [31], the RS model is defined by the concept of fuzzy relation and fuzzy operators (max and min).

The RS theory is superior to multiple regressions, because it does not require any previous knowledge about the data that is considered. Fuzzy preference is a special case of fuzzy relations. The rough set FS method collects information from both the distance metric and lower approximation dependency value. The FS method also considers the total number of objects in the boundary region and the distance of those objects from the lower approximation. Though it is very efficient in mining technique, the rough set suffers from huge computation of either discernibility function or the positive region to find the attribute reduction. To overcome the problem rough set need to be blend with fuzzy as feature selection. This fuzzy preference relation, when blended with the rough set, is known as FPRS, and this relation has been modeled to measure the fuzzy preference.

Let an information system be (U, F), where $U = \{a_1 \dots a_n\}$ is a nonempty finite set of objects, and $F = \{F_1 \dots F_N\}$ is a finite set of features to classify the objects.

A decision table is defined by $DT = (U, C, D)$, where the set of features are grouped into condition (C) and decision (D). The conditions (features) are assumed, and the task is to classify the object. The decision about the object of U is then predicted approximately.

Let N decision class labels can be represented as $d_1, d_2, \dots, d_N$, where $d_1 < d_2 < \dots \dots d_N$.

In the RS theory, the fundamental operations are lower approximation and upper approximation. It is assumed that $R >$ and $R <$ are the fuzzy preference relation that is generated by $P \subseteq C$. The fuzzy preference approximation qualities [32] of the decision D in terms of P are denoted as:

$$\alpha_p^{>}\left(D^{\geq}\right) = \frac{\sum_j \sum_{x \epsilon d_j^{\geq}} R^{>} d_j^{\geq}(x)}{\sum_j ||d_j^{\geq}||} \tag{3}$$

$$\alpha_p^{<}\left(D^{\leq}\right) = \frac{\sum_j \sum_{x \epsilon d_j^{\leq}} R^{<} d_j^{\leq}(x)}{\sum_j ||d_j^{\leq}||} \tag{4}$$

$$\alpha_p\left(D\right) = \frac{\sum_j \sum_{x \epsilon d_j^{\leq}} R^{<} d_j^{\leq}(x) + \sum_j \sum_{x \epsilon d_j^{\geq}} R^{>} d_j^{\geq}(x) A = \pi r^2}{\sum_j (||d_j^{\leq}|| + ||d_j^{\geq}||)} \tag{5}$$

$$0 \leq \alpha_p^{>}(D^{\geq}) \leq 1, \quad 0 \leq \alpha_p^{<}(D^{\leq}) \leq 1 \quad \text{and } 0 \leq \alpha_p(D) \leq 1.$$

It can be stated that D is global or upward or downward consistent if $\alpha_p\ (D) = 1$, $\alpha_p > (D \geq) = 1$ or $\alpha_p < (D \leq) = 1$, respectively.

4 Experimental results and discussion

In our experiments, we have first selected relevant features by using the FUC, FLC, and FGC models of the FPRS method. Then classifier SVM is applied on each partition of all the datasets considering the genes from each model of FPRS method. The characteristics of cancer datasets are small sample size with huge number of attributes (often with hundreds or thousands of dimensions). Hence, the dataset contain huge inconsistency due to redundancy and noise of the data. The rough set contains an effective tool to deal with inconsistency data.

4.1 Dataset Description

Here, we have used three publicly available bi-class datasets, one multiclass, and one biological dataset to conduct the experiment. The datasets are available in [33]. The details of each dataset are described below and presented in Table 1.

Table 1: Brief Description of the Datasets Used in the Proposed Work

Data set	No. of features	No. of samples	No. of classes
WBCD	9	699	2
Leukaemia	5147	72	2
Prostate	12533	102	2
DLBCL	7070	77	2
MLL	12533	72	3

WBCD

This kind of abnormal growth cell can lead to the formation of a lump or tumor. Breast cancer begins with the abnormal growth of breast cells in human body. There are two types of tumors (benign and malignant). Benign tumors do not spread outside the lobules (the onset of breast cancer). They look like normal cells, and they grow very slowly. It can lead to non-invasive breast cancer. In contrast, malignant tumors break out into the surrounding breast tissues. The dataset consists of 699 instances from needle aspirates of a patient's breast. The dataset contains 16 instances with missing values, which are removed from it, resulting in 683 samples. After removing the noisy samples, the dataset now contains 444 and 239 samples that belong to benign and malignant classes, respectively.

Leukemia

The dataset consists of 72 samples and 5147 genes. Out of these, 47 samples are from Acute Myeloid Leukemia (AML) patients and 25 samples are from Acute Lymphocytic Leukemia (ALL) patients.

Prostate

This is an affymatrix Human Genome 95AV2 array set. The dataset consists of 102 samples with 12533 attributes. Among the 102 samples, 52 are from prostate tumor samples and 50 samples are from non-tumor prostate samples.

DLBCL

Diffuse Large B-cell Lymphomas (DLBCL) and Follicular Lymphoma (FL) are the two B-cell lineage malignancies. DLBCL consists of 77 samples and 7070 genes. Among the 77 samples, 58 samples are from DLBCL class and 19 samples from FL classes. Out of 58 samples, 26 samples are from a patient with fatal disease and 32 samples are from a cured patient.

MLL

The dataset consists of 72 samples and 12533 genes with 3 classes. The three classes are ALL (24), (MLL) (20) and AML (28), respectively.

5 Results and Discussion

The results that are obtained from the proposed approaches are presented in this section. The three models of the FPRS method are compared to assess the effects of FUC along with FLC and FGC models. To identify the best approaches, all the approaches are compared in Table 2 in terms of classification accuracy. Results are compared with different partitions: 80–20%, 70–30%, and 50–50% on five datasets. The performance of

Table 2: Classification accuracy and the number of genes for FUC, FLC and FGC model of the FPRS method with different training-test partitions.

Dataset	Model	No. of Genes	Classification Accuracy (%)		
			80-20%	*70-30%*	*50-50%*
WBCD	**FUC**	3	**100**	**99.02**	**98.24**
	FLC	5	97.81	96.58	96.19
	FGC	6	97.08	95.32	94.13
Leukaemia	**FUC**	**9**	**96.19**	**95.02**	**90.23**
	FLC	16	90.12	90.00	87.04
	FGC	20	90.00	88.34	87.02
Prostate	**FUC**	**9**	**97.32**	**97.10**	**95.00**
	FLC	20	93.54	93.00	91.17
	FGC	22	89.14	87.62	85.90
DLBCL	FUC	14	95.33	94.32	90.13
	FLC	**12**	**96.14**	**95.67**	**90.39**
	FGC	20	90.34	87.22	85.12
MLL	**FUC**	**9**	**96.45**	**96.00**	**95.67**
	FLC	11	89.00	88.00	85.33
	FGC	15	90.13	87.00	82.00

FUC over FLC and FGC is outlined in this section. It is worth mentioning that FUC outperforms the other two methods over all the datasets. In contrast, FUC does not perform well only for the DLBCL dataset that has 95.33% accuracy. However, FLC exhibits 96.14% of accuracy for the same dataset.

In Table 3, we have presented the accuracy of the whole dataset (without FS) by using the SVM classifier along with 50–50% training test partition of the whole dataset. The observed results are very poor in terms of accuracy. The significant genes that are obtained from our proposed methods are listed in Table 4. In this literature the effectiveness of our proposed method is demonstrated on five datasets. Also it is clear from the table that some biomarkers (marked in bold faces) identified from our method are reported by other literature (described in the supplementary copy). Hence, those biomarkers have a great impact in causing cancer. Whereas, the rest are informative genes as we are getting good classification accuracy as well as good statistical measurement of the proposed method. It can be observed from Table 5 that there are some overlapping genes between different models of FPRS method. For example FUC, FLC and FGC three models of FPRS have selected two common genes from leukemia dataset.

Table 3: The Number of features of the proposed method on five datasets, and the accuracy using the SVM classifier on the whole dataset without using any FS method.

Dataset	No. of features in proposed model	Accuracy on selected gene (%)	Accuracy on whole dataset using SVM (%)
WBCD	3	100	**85.67**
Leukaemia	9	100	83.78
Prostate	9	100	62.75
DLBCL	12	100	76.92
MLL	9	100	68.57

An experimental study on five datasets is performed using statistical measurement, and the results are reported in Figure 2. In Figure 2(a–d) WBCD dataset explores equal substitution (100%) between specificity and sensitivity. In contrast, for the rest of the datasets, it explores ≥ 95% statistical measurement for leukaemia,

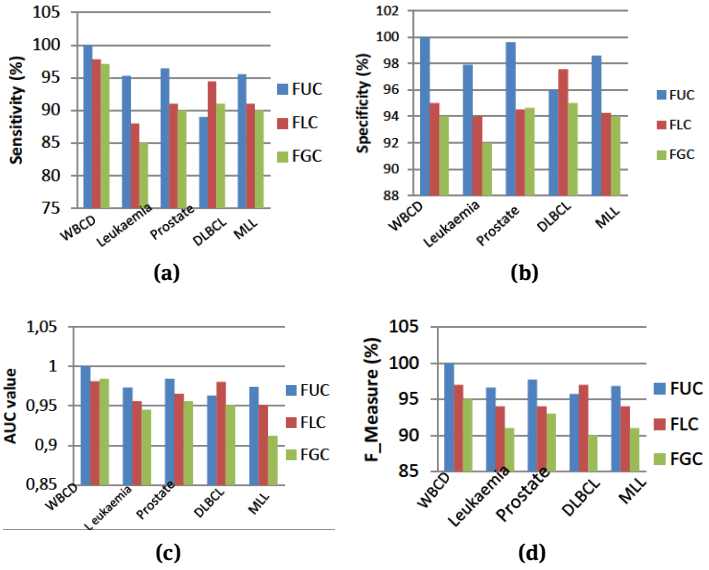

Figure 2: (a–d): The Performance (sensitivity, specificity, AUC, and measure) of the classifier SVM from the FUC, FLC and FGC methods on five datasets: WBCD, leukaemia, prostate, DLBCL and MLL respectively.

prostate, and MLL datasets. FUC obtains the Area under Curve (AUC) value (=1) for the WBCD dataset. However, for the other datasets, the AUC value almost reaches one (=1). It can be concluded from the figure that the FUC model provides the best statistical measurement when compared to the two FLC and FGC models for the four datasets from among the total five datasets. Moreover, the boxplot results for the three models are depicted in Figure 3(a–e). The training and test sets are randomly partitioned, and the classification by using

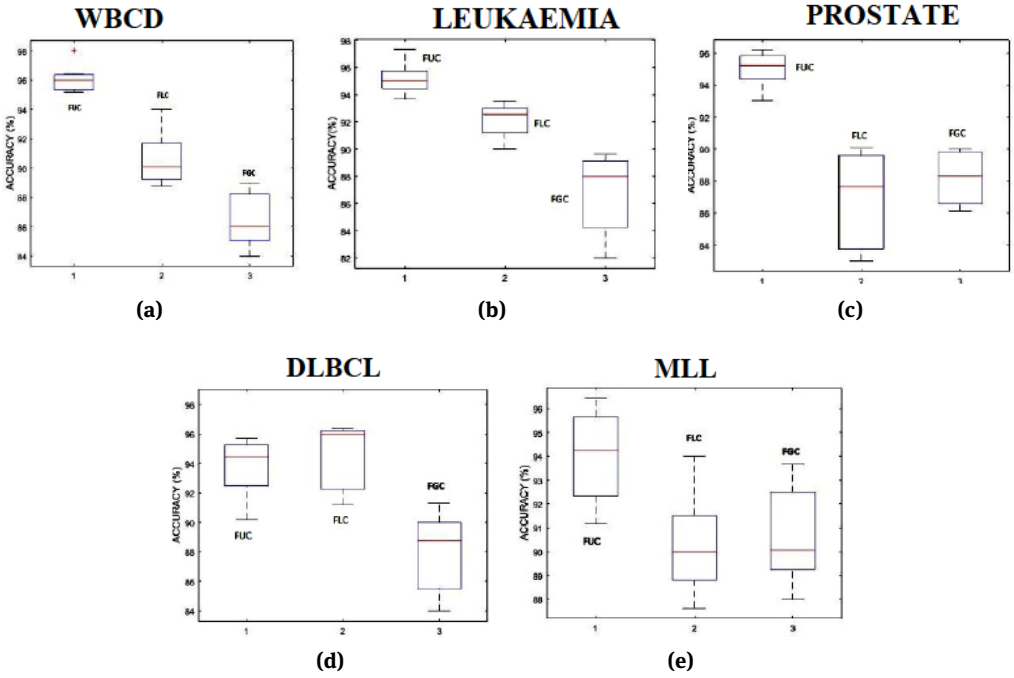

Figure 3: (a-e): The boxplot showing the accuracies produced by different models of the FPRS method over the best 7 runs for different training subsets of a size that is 50% of WBCD, leukaemia, prostate, DLBCL and MLL datasets.

SVM is performed to obtain the boxplot results. From the figure, it is clear that the FUC surpasses the other two models, because it is placed in the topmost position. The lists of gene that are selected from our proposed FUC models are reported in Table 4. The experiment is implemented in matlab by using the Lib-SVM software that was developed by Chang and Lin [34]. The experiment is performed on Intel Core i5 2430M CPU (2.46 GHZ) with 4GB of RAM.

Table 4: Biomarkers and informative genes obtained from the proposed model for the five datasets.

Dataset	WBCD	Leukaemia	Prostate	DLBCL	MLL
Informative Genes	**C1**	**CST3**	**HPN**	**POU6F1**	**MME**
	C2	**DNTT**	**TGFB3**	**TRIB2**	**SPTBN1**
	C6	**CD33**	**XBP1**	**HLA_A**	**LIG4**
		MPO	**MAF**	**GPCR**	**FLT3**
		CD79A	**CALM1**	CTSD	**CCNA1**
		CCND3	DF	PSMC1	**GRAF**
		ZYX	LMO3	GPR18	CTSD
		CD19	PTGDS	MELK	1894_f_at
		TCF3	P4HB	SEPP1	MBNL1
				KPNA2	
				TXNIP	
				EEF1A1	

Table 5: Number of genes overlapped between different methods of FPRS method.

Datasets	FUC+SVM	FLC+SVM	FGC+SVM
WBCD	2	2	2
Leukaemia	2	2	2
Prostate	3	3	3
DLBCL	2	2	2
MLL	0	0	0

6 Comparative study

The investigation of the performance of our proposed approach in terms of classification accuracy is summarized in Table 6. Our proposed method produces good results for all the datasets with good classification accuracy. However, it is difficult to further compare our method with those listed in the said table. The results are promising in most of the published works. As for the MLL dataset, it makes it difficult to search reference for comparison.

The classification accuracy is 95.67% with nine genes when our proposed method is used. It can be stated that the proposed method cannot outperform all the present methods. However, it can surpass some of the published articles.

Table 6: Classification accuracies (%) obtained from our method and other models obtained from literatures.

		WBCD	Leukaemia	Prostate	DLBCL	MLL
Proposed Method	**FUC + SVM**	**100**	**96.19**	**97.32**	**95.33**	**96.45**
	J4, MLP (PCA) [35]	97.56				
	Decision Tree [36]	96.14				
	SVM, artificial Neural network [37]	97.00				
	Decision tree with feature-selection [38]	97.85				
	SVM [39]	94.54				
	LEM2, Rough set reduction method [40]	96.40				
	F_score with SVM [41]	99.51				
	PCA + MLP [42]		94.42		91.60	
	GA-PCA & CCA [43]		88.23			
	VVRKFA [44]		94.81	93	88.97	
	DRF0+SVM/KNN [45]		94.12	97.06	94.67	
	DRF+IG+SVM [45]			97.06		

7 Conclusion

The FUC model is employed as a gene-selection method. The proposed model evaluates the significance of genes with respect to a certain criteria. By using the RS theory, the proposed method yields information about the criteria that are important for conclusion as well as about those that are redundant or unwanted. Four gene-expression datasets and one biological dataset are investigated by the proposed method. The biomarkers and informative genes are identified here. A comparative study can establish the superiority of the proposed model in conjunction some of the published articles.

The final goals of our proposed method are the accurate tumor identification and diagnose cancer, which is possible by finding biomarkers. It is apprehensible as our proposed method is capable of extracting some biomarkers as it is reported by some literatures. But there are some limitations in FPRS method during fuzzy preference based approximation. Firstly, a function is used to compute the preferences degrees of objects. But there are lot of functions (and choices of their parameters) to measure the preferences. Secondly, user should provide information whether monotonous relations persists between attributes and decision. However, it is difficult for the user to provide such information in application and an algorithm is require to provide such information. Finally, the membership of any consistent sample to the upper approximation may be less than that of lower approximation, which is not compatible with the definition of fuzzy approximation. Moreover, the limitations of SVM is the lack of transparency of results.

Hence, the method can be used as an alternative means in the diagnosis of cancer. For future work it might be interesting to exploit imperialist competition algorithm for measuring the efficiency of our model.

References

[1] S. K Thazha , H. Fernandez , C. P. Cruz , J. P. Cruz, Role of Fine needle aspiration cytology in the diagnosis of palpable breast lesions and its correlation with histopathology Basis, International Journal of Health Sciences & Research, vol.8, issue. 10, 2018.

[2] Y. KY, R. WL, Principal component analysis for clustering gene expression data, Bioinformatics, vol. 17, issue. 9, pp.763–74, 2001.
[3] L. JJ, C. WS and S. XG, Cancer classification based on microarray gene-expression data using a principal component accumulation method, Sci China Chem, vol. 54, issue. 5, pp. 802–11, 2011.
[4] X. Wang and O. Gotoh, A robust gene selection method for microarray-based cancer classification, Cancer Informatics, vol. 9, pp. 15-30, 2010.
[5] Y. Saeys, I. Inza and P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics, vol. 23, issue. 19, pp. 2507–17, 2007.
[6] Y. Ej, R. Me, S. Sa, W. Wk, Patel D, Mahfouz R, et al., Classification, subtype discovery and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, Cancer cell, vol. 1, issue. 2, pp. 133–43, 2002.
[7] O. CH, T. P, Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, Bioinformatics, vol. 19, issue.1, pp. 37–44, 2003.
[8] R. Díaz-Uriarte, S. Alvarez de Andrés, Gene selection and classification of microarray data using random forest, BMC Bioinformatics, 2003.
[9] S. P. Potharaju, M. Sreedevi, Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance, Clinical Epidemiology and Global Health, 2018.
[10] S. Tiwari, B. Singh and M. Kaur, An approach for feature selection using local searching and global optimization techniques, the natural computing applications forum, vol. 8, issue. 10, pp. 2915-2930. 2017.
[11] M. Panda, Performance comparison of genetic algorithm, particle swarm optimisation and simulated annealing applied to TSP, International Journal of Applied Engineering Research, vol. 13, and issue. 9, pp. 6808-6816, 2018.
[12] J. Liu, H. Zheng, Y. Zhang, X. Li, J. Fang, Y. Liu, C. Liao, C. Liao, Y. Li and J. Zhao, Imperialist competition algorithm, dissolved gases forecasting based on wavelet least Squares Support vector regression and imperialist competition algorithm for assessing incipient faults of transformer polymer insulation. Polymers, vol. 11, issue. 1, 2019.
[13] J. Apolloni, G. Leguizamon and E. Alba, Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments, Applied Soft Computing, vol. 38, issue. c, pp. 922-932, 2018.
[14] F. Han, C. Yang, Y. Wu, J. S. Zhu, Q. H. Ling, Y. Q. Song and D. S. Huang, A gene selection method for microarray data based on binary pso encoding gene-to-class sensitivity information, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, issue. 1, pp. 85-96, 2017.
[15] A. Houari, W. Ayadi and S. B. Yahia, A new FCA-based method for identifying bi-clusters in gene expression data, International Journal of Machine Learning and Cybernetics, vol. 9, issue. 11, pp. 1879-1893, 2018.
[16] R. Jothi, S. K. Mohanty and A. Ojha, DK-means: a deterministic K-means clustering algorithm for gene expression analysis, Pattern Analysis and Applications, pp. 1-19, 2017.
[17] V. Vapnik, The nature of statistical learning theory, vol. 8, Issue. 6, New York, 1995.
[18] U. Maulik, D. Chakra borty, Fuzzy preference based feature selection and semi supervised SVM for cancer classification, IEEE Transactions on Nano Bioscience, vol.13, issue. 2, pp. 52-160, 2014.
[19] A. Saxena, L. K. Gavel, M. M. Shrivas, Rough set for feature selection and classification: An overview with application, IJRTE, vol-3, Issue. 5, 2014.
[20] American Cancer Society: Breast Cancer Overview, January, 2016.
[21] R. Caruana and A. N. Mizil, An empirical comparison of supervised learning algorithms, 23^{rd} International Conference on Machine Learning, Pittsburgh, P.A, 2006.
[22] V. N. Vapnik, An overview of structural learning Theory, IEEE Transactions of Neural Networks,vol.10,no.5, 1999.
[23] Prasad S. Thenkabail, Remote Sensing Open Acces Journal: Increasing impact through Quality publications, Remote Sensing, vol. 6, pp. 7463-7468, 2014.
[24] B. M. Gayathri,C. P. Sumathi and T. Santhanam, Breast cancer diagnosis using machine learning algorithm A survey, International Journal of Distributed and Parallel Systems (IJDPS) vol.4, issue. 3, May 2013.
[25] P. Chen, C. Lin and B. Scholkorf, A tutorial on support vector machines: Applied Stochastic Models in Business and Industry, vol. 21, issue. 2, pp. 111- 136, 2005.
[26] V. Kumar and S. Minz, Feature Selection: A Literature Review, Smart Computing Review, vol. 4, issue. 3, 2014.
[27] A, Michael, MD. Marchetti, Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to Dermatologists for the diagnosis of melanoma from dermoscopic images, Journal of the American Academy of Dermatology, Elsevier, vol. 78, issue. 2, pp. 270-277, 2017.
[28] I. Guyon and A. Elisseeû, An Introduction to variable and feature selection, Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.
[29] Xindong. Wu...D. Steinberg, Top 10 algorithms in data mining, Springer-Verlag London Limited, vol. 14, pp. 1-37, 2007.
[30] D. Chakraborty and U. Maulik, Identifying cancer biomarkers from microarray data using feature selection and semi supervised learning, IEEE Journal of Translational Engineering in Health and Medicine: vol. 2, 2014.
[31] Z. Pawlak, Rough set theory and its applications, Journal of Telecommunication and Information Technology, vol. 3, issue. 3, pp. 7-10, 2002.
[32] Q. Hu, D. U and M. Gua, Fuzzy Preference Based Rough Set, Information sciences, vol. 180, issue. 10, pp. 2003-2022, 2010.

[33] [Online]. Available: http://www.biolab.si/supp/bi-cancer/projections/
[34] C-C Chang, C-J Lin, LIBSVM: A library for support vector Machine, ACM Transaction on Intelligent Systems and Technology, vol. 2, issue. 3, no. 27, 2011.
[35] G. I. Salama, M. B. Abdelhalim, and M. Abd-elghany Zeid, Experimental comparison of classifiers for breast cancer diagnosis, Seventh International Conference on Computer Engineering & Systems (ICCES), 2012.
[36] Y. Li, Z. Chen, Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction, Applied and Computational Mathematics, vol. 7, issue. 4, pp. 212-216, 2018.
[37] S. S. Shrivastava , A. Sant , R. P. Aharwal, An overview on data mining approach on breast cancer data, International Journal of Advanced Computer Research, vol. 3, , issue. 13, no. 4, 2013.
[38] D. Sudhir, A. A. Ghatol Ashok., Pande Amol P., Neural Network aided breast cancer detection and diagnosis,7th WSEAS International Conference on Neural Networks, 2006.
[39] L. Bhambu, Dr. D. Kumar, A novel approach for classification on breast cancer data set, vol. 5, issue. 7, International Journal of Advanced Research in Computer Science and Software Engineering, 2015.
[40] D. Lavanya and Dr. K.Usha Rani, Ensemble decision making system for breast cancer data, International Journal of Computer Applications, vol.51, no.17, pp.19-23, 2012.
[41] M. F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis, Expert Systems with Applications, vol. 36, issue. 2, pp. 3240–3247, 2009.
[42] R. O. Vega, G. S. Ante, M. A. de Luna, R. Vega, L. E. F. Morales and H. Sossa, Improving pattern classification of DNA microarray data by using PCA and logistic regression, Intelligent Data Analysis, vol. 20, pp. S53–S67, IOS Press, 2016.
[43] S. J. Susmi, H. K. Nehemiah and A. Kannan, Hybrid dimension reduction techniques with genetic algorithm and neural network for classifying leukemia gene expression data, Indian Journal of Science and Technology, vol. 9, 2016.
[44] S. Ghorai, A. Mukherjee, P. K. Dutta, Gene expression data classification by VVRKFA, Procedia Technology, vol. 4, pp. 330 – 335, 2012.
[45] P. Jaganathan, N. Rajkumar, and R. Kuppuchamy, A comparative study of improved f-score with support vector machine and RBF network for breast cancer Classification, International Journal of Machine Learning and Computing, vol. 2, issue. 6, 2012.
[46] http://docs.lib.purdue.edu/ecetr
[47] C. Chu, A. Hsu, K. Chou, P. Bandettini and C. Lin, Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images, Neuroimage, vol. 60, and issue. 1, pp. 59-70. 2012.
[48] X. Y. Xu, Z. Xiaoshu Zhu, L. Quan, O. S. Gilbert, W. Jianxin, Cluster-Mine: a knowledge-integrated clustering approach based on expression profiles of gene sets Hong-Dong L, Biorxiv preprint first posted online Jan. 29, 2018.