Optimal Feature Selection using Fuzzy Combination of Feature Subset for Transcriptome Data

Vikas Singh
Dept. of
Electrical Engineering
IIT Kanpur, India
Email: vikkyk@iitk.ac.in

Harsh Vardhan
Dept. of
Electrical Engineering
IIT Kanpur, India
Email: harshv@iitk.ac.in

Nishchal K Verma
Dept. of
Electrical Engineering
IIT Kanpur, India
Email: nishchal@iitk.ac.in

Yan Cui
Dept. of
Genetics, Genomics and Informatics
University of Tennessee, USA
Email: ycui2@uthsc.edu

Abstract—**Applying machine learning algorithms directly on high dimensional datasets, like those encountered in transcriptome analysis, may lead to high time complexity and low performance of learning models, especially when the number of samples is small compared to the dimensionality. Selecting the optimal set of features then becomes an essential task for such datasets. Filter methods are one of the main class of techniques used for feature selection wherein a score is assigned to features based on criteria such as information gain, statistical measures or similarity based measures and then selects the best scored features. Using filter methods on the complete dataset results in features that have good performance over the dataset but might perform poorly in certain regions of the data, which affects accuracy for data points of those regions. To overcome this degradation in performance, we propose two novel methods to assign a robust score by using the fuzzy combination of the region-specific optimal feature subsets obtained using a standard feature selection algorithm (we use *mRMR* for this paper).We compare the result with state-of-the-art feature selection algorithm, *mRMR* (Minimum Redundancy Maximum Relevance) in the terms of accuracy on certain standard datasets.**

I. Introduction

With the rapid advancement of transcriptomic profiling technologies such as microarray and RNA-sequencing, the simultaneous monitoring of tens of thousands of genes expression levels becomes tractable [1]. Transcriptome provides valuable information for understanding how biological systems work in health and disease. However, in the analysis of transcriptome data, the power and accuracy of machine learning algorithms are hampered by the limited number of samples. This is a well-known small n, large p problem, where the number of genes (n) is much larger than the number of samples (p) in the dataset. Feature selection is often used to overcome this problem. This process involves selecting the optimal subset of the input features to extract the output from the input using a learning model. It has many benefits in terms of coping with dimensionality reduction, reducing variance of the data to prevent overfitting, discovering hidden structure in the data and increasing interpretability of the models. Feature selection has become mostly a necessary procedure in application of machine learning or pattern recognition algorithms to bio-medical engineering, text processing and image processing applications. Problems in these fields, especially medical, contain very few examples (in order of hundreds) and have huge number of feature (in order of tens of hundreds of thousands) [4], [5] which makes feature selection a necessary procedure for obtaining good performance of the trained models in reasonable time.

Filters, wrappers and embedded methods are the three general approaches are used for the feature selection [6]–[8]. The filter methods are generally based on ranking or space search and they used data pre-processing or filtering by which features are selected based on their intrinsic characteristics which determine their relevance with regard to the target classes [9], [10]. Mutual information (MI) [11], [12], Test feature selection [13], Correlation based feature selection [14], Bayesian networks [15], [16] and Information gain (IG) [17] are shown to be most effective filter feature selection method. The feature selection in filter methods are uncorrelated with learning methods, therefore it has good generalization. In the literature various feature scoring function are proposed which are similarity-based, information-theoretic based or some statistical measures based. Statistical measures are based on correlation i.e. chi-squared score [29], f-statistics [30] are very prevalent but they are not invariant under variable transformation. Similarity based algorithms use scores like Pearson correlation coefficients, Fisher criterion score [31], and the Kolmogorov-Smirnov test [32]. In the wrapper method feature selection is wrapped with learning model by which the usefulness of the feature is directly judged by the estimated accuracy of the learning method. Genetic algorithm (GA) [18]–[20] and sequential search [21] are most effective wrapper methods. Embedded techniques tend to do better computationally than wrappers but they make classifier dependent selections that might not work with any other classifier. Kleftogiannis *et al.* [22] have combined the support vector machine (SVM) with GA which increases both efficiency and robustness of the feature selection method. In [23]–[27] authors have presented Entropy-based Recursive Feature Elimination (E-RFE), recursive cluster elimination, an entropy based variant of support vector machine recursive feature elimination (SVM-RFE) and deep learning based methods by which irrelevant features are eliminated from data. Maulik *et al.* [28] have also proposed a novel feature selection which is based on forward greedy search algorithm and transductive support vector machine for classification.

978-1-5090-6020-7/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on December 01,2020 at 00:58:40 UTC from IEEE Xplore. Restrictions apply.

In this paper, we have proposed two novel methods to assign a robust score by using the fuzzy combination of the region-specific optimal feature subsets obtained using a standard feature selection algorithm (we use *mRMR* for this paper). Since running the algorithms on complete dataset results in features which have good performance over the dataset but might perform poorly in certain regions of the data, which affects accuracy for data points of those regions. To overcome this degradation in performance we use a state-of-the-art feature subset selection (*mRMR*) algorithm to select the optimal feature subsets for each cluster and represent the optimal feature subset for the complete dataset as the fuzzy combination of each of these subsets. Among the two methods proposed, the first method selects only *one*-feature subset from each cluster while the second method selects top t-feature subsets per cluster.

The rest of the paper is presented as follows; Section II describes basic definition of mutual information and fuzzy set and the functions which we will be using in the paper. Section III describes the proposed methodologies, Section IV describes the result and discussion, Finally, Section V concludes the paper.

II. Preliminaries

A. Mutual Information

The mutual information [11], [12] between two continuous random variables X and Y given as:

$$I(X;Y) = \iint_{x,y} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} dxdy \qquad (1)$$

where, $p_{X,Y}(x,y)$ is the joint probability density function while $p_X(x)$ and $p_Y(y)$ are the marginal probability distributions of the random variables X and Y.

B. Fuzzy sets

If X is a collection of attributes denoted by x, then a fuzzy set A in X is defined as a set of ordered pairs:

$$A = \{(x, \ \ \mu_A(x)) \mid \ x \in X\} \qquad (2)$$

where, $\mu_A(x)$ is called membership function(MF) of x for the fuzzy set A, which maps each element of X to a membership grade between 0 and 1 [33]–[37].

1) Fuzzy Union: We have constructed a new notion of fuzzy union which we utilise in our method to obtain closed form expressions for union of fuzzy sets having sigmoidal membership functions. Let S_1, S_2 and S_2 be 3 fuzzy sets and x be an element in them. We will use the symbol $\vee_{j=1}^{3} S_j$ to denote the fuzzy union of these sets. For x having membership function values $\sigma(x_1)$ in set S_1, $\sigma(x_2)$ in set S_2 and $\sigma(x_3)$ in set S_3 respectively, we define the membership function value of x in $\vee_{j=1}^{3} S_j$ as $\sigma(x_1 + x_2 + x_3)$ where, x_i are positive real numbers for $i \in \{1, 2, 3\}$. This satisfies the properties of the union over sets for positive and real x_i, $\sigma(x_i) \leq \sigma(x_1+x_2+x_3)$, since $\sigma(x)$ is an increasing function and its value is always less than 1.

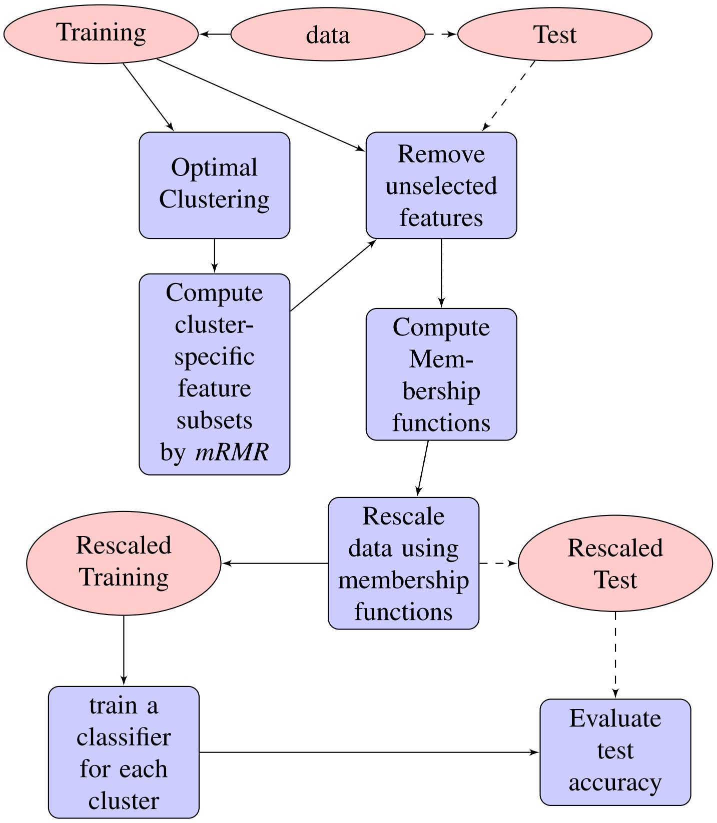
Training
data
Test
Optimal Clustering
Remove unselected features
Compute cluster-specific feature subsets by *mRMR*
Compute Membership functions
Rescale data using membership functions
Rescaled Training
Rescaled Test
train a classifier for each cluster
Evaluate test accuracy

Fig. 1: Complete flow diagram of proposed approach

C. Sigmoid

The sigmoid function σ is defined below.

$$\sigma(x) = \frac{1}{1 + exp(-x)} \qquad (3)$$

It is bounded in the range $(0, 1)$ with the limiting values achieved at $x \to \mp\infty$ respectively. Also, the function has a value of 0.5 at $x = 0$.

III. Proposed Methodologies

The main motivation behind using these methods are the observation that using the same feature selection algorithm on different clusters of the dataset yields different optimal feature sets. Thus selecting a single feature subset for the complete dataset should result in loss of accuracy. To make the feature selection for the whole dataset and more context-sensitive we propose a fuzzy combination of the optimal feature subsets, based on some existing feature selection method, of each cluster. The proposed methods comprise of three steps, namely optimal clustering, cluster-specific feature selection and membership function assignment of each features. The K-Means with K-Means++ initialization [38] is used as the clustering algorithm to find the optimal number of clusters for each dataset (here the variable N) by silhouette score analysis [39]. After partitioning the dataset into clusters, we perform *mRMR* feature selection on each cluster separately. For the

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on December 01,2020 at 00:58:40 UTC from IEEE Xplore. Restrictions apply.

subsequent steps, we propose slightly two different methods for membership function assignment i.e. Membership Assignment : *One*-feature subset per cluster(MA-1) and Membership Assignment : Top t-feature subsets per cluster (MA-t)

A. Membership Assignment: One-feature subset per cluster(MA-1)

For each optimal cluster we compute the k best features and their respective mutual information with that cluster's response vector using *mRMR*. Let $S_1, S_2, \cdots, S_j, \cdots, S_N$ be the sets of indices of optimal features for the N optimal clusters.

We first define the identity function $\mathbb{1}(x)$, which is 1 if its argument is true and 0 otherwise. We will use the $\cup$ symbol to represent union of sets (non-fuzzy) and the $\vee$ symbol to represent fuzzy union-like operation on sets (as defined above). X is the feature matrix and y is the response vector. $X(t,i)$ represents the value of the i^{th} feature for the t^{th} data point and y(i) represents the response of the t^{th} data point.

Having stored mutual information $I(x_{i,j}; y_j)$ between the responses y_j and the i^{th} input feature for the j^{th} cluster, we define a membership function $(f_j(i,t))$ of the i^{th} feature for the t^{th} data point in the j^{th} cluster as;

$$f_j(i,t) = \sigma(\frac{\mathbb{1}(i \in S_j) I(x_{i,j}; y_j)\lambda}{d_{t,j}}) \tag{4}$$

where, $d_{t,j}$ is the euclidean distance between the t^{th} data point and the j^{th} cluster. λ is the scaling constant and σ is the sigmoidal function and $x_{i,j}$ is the column vector of the i^{th} feature for the j^{th} cluster.

Now, we define the optimal feature set for the complete dataset as $\vee_{j=1}^{N} S_j$, where N is the total number of optimal clusters and the fuzzy union-like operation as defined in section II:

Then for t^{th} data point, we can define the total membership function of its i^{th} feature as the fuzzy union-like operation on the feature subsets selected in each of its N optimal clusters.

$$\begin{gathered} f(i,t) = \vee_{j=1}^{N} f_j(i,t) f(i,t) \\ = \vee_{j=1}^{N} \sigma(\frac{\mathbb{1}(i \in S_j) I(x_{i,j}; y_j)\lambda}{d_{t,j}}) \end{gathered} \tag{5}$$

By rearranging the eq. (4) and eq. (5) the total membership function is calculated as;

$$f(i,t) = \sigma(\sum_{j=1}^{N} \frac{\mathbb{1}(i \in S_j) I(x_{i,j}; y_j)\lambda}{d_{t,j}}) \tag{6}$$

The features which are not present in $\vee_{j=1}^{N} S_j$ are discarded. For the remaining features, using the above membership functions we scale the i^{th} feature of the t^{th} data point as;

$$X^{new}(t,i) = X(t,i) f(i,t) \tag{7}$$

This scaled dataset i.e. transformed dataset act as a new dataset with reduced features. The complete approach is described in Algorithm 1 and shown in Fig. 1 in a concise manner.

This design of membership function offers us several advantages. The numerator contains a mutual information between the feature and the response which increases membership values for more relevant features. The distance term in the denominator is to account for the fact that each data point should select the most relevant features according to its position (i.e, the best features according to its cluster). A sigmoidal membership function takes care of the singularities, keeps the membership function between 0 and 1 and enables us to get a 'nice'looking closed form total membership function for each feature.

Algorithm 1 : Membership Assignment: *One*-feature subset per cluster(MA-1)

```
Begin
Input: $X$,$y$
Output:$X^{new}$
Perform silhouette analysis to obtain optimal number of
  clusters ($N$)
Cluster the data into $N$ cluster using K-Means with K-
  Means++ initialization. Let $X_i, y_i$ be the subset of $X, y$
  present in the $i^{th}$ cluster
for j = 1 to N do                          //(for each cluster)
    $S_j$ = mRMR($X_j, y_j$)
end for
for t = 1 to n do                       //(for each data point)
    for i $\in \cup_{j=1}^{N} S_j$ do          //(for each selected feature)
        $X^{new}$(t,i) = X(t,i)$\sigma(\sum_{j=1}^{N} \frac{\mathbb{1}(i \in S_j) I(x_{i,j}; y_j)\lambda}{d_{t,j}})$
    end for
end for
return $X^{new}$
End
```

B. Membership Assignment: Top t-feature subsets per cluster (MA-t)

The previous method as discussed selects the locally most relevant features according to the data point. However, since it uses mutual information of the individual features, it cannot consider the relevance of a feature subset. *mRMR* being a subset-based method tries to estimate relevance and redundancy of a feature subset to the response. Even though the features have been selected by a feature subset selection algorithm, they are being weighed by their individual mutual informations. Such an approach will give consistently low membership values to complementary features even if their combination is an exceedingly good feature subset. To tackle this problem, we try to extend our fuzzy combination approach to directly the feature subsets. Hence, instead of selecting the top feature subset of each cluster, we compute the t best k feature subsets of each cluster and take their fuzzy combination (like fuzzy union) as the optimal feature subset. The features selected in previous algorithm are replaced by feature subsets and the mutual information term between individual feature and response is replaced by the estimated mutual information between the feature subset and response. This makes our

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on December 01,2020 at 00:58:40 UTC from IEEE Xplore. Restrictions apply.

complete algorithm operate on feature subsets handling the difficulties faced due to complementarity of features. Now to compute the membership functions of the individual features we simply take product of membership functions of each feature subset in which it appears. The mutual information term in the numerator gives higher membership value to features belonging to good feature subsets and the distance term in the denominator ensures giving higher membership value to features of 'locally'good feature subsets. The details of this method have been explained below.

In this method, we select the L best feature subsets according to *mRMR* for each cluster. Let the L best feature subsets for the j^{th} cluster be $S_{j,1}, S_{j,2}, .., S_{j,L}$ and let $I(S_{j,1}; y_j), I(S_{j,2}; y_j), ..., I(S_{j,L}; y_j)$ be their respective mutual information with the response as estimated by *mRMR*. Then we can now define the membership function for i^{th} feature of t^{th} data point in the feature subset $S_{j,l}$ as;

$$f_{j,l}(i,t) = \sigma(\frac{\mathbb{1}(i \in S_{j,l}) I(S_{j,l} y_j)\lambda}{d_{t,j}})i \tag{8}$$

Algorithm 2 : Membership Assignment: Top t-feature subsets per cluster(MA-t)

```
Begin
Input: $X$,$y$
Output:$X^{new}$
Perform silhouette analysis to obtain optimal number of clusters (N)
Cluster the data into $N$ cluster using K-Means with K-Means++ initialization. Let $X_i, y_i$ be the subset of $X, y$ present in the $i^{th}$ cluster
for j = 1 to N do                                  //(for each cluster)
    Obtain $S_{j,1}, S_{j,2}, \ldots, S_{j,L}$ and $I(S_{j,1}; y_j), I(S_{j,2}; y_j), \ldots, I(S_{j,L}; y_j)$ by mRMR($X_j, y_j$)
end for
for t = 1 to n do                                  //(for each data point)
    for i $\in \cup_{j=1}^{N} S_j$ do              //(for each selected feature)
        $X^{new}$(t,i) = $X(t,i)f(i,t)$
    end for
end for
return $X^{new}$
End
```

Now, the optimal feature subset for the j^{th} cluster is defined as $\vee_{l=1}^{L} S_{j,l}$ Thus, the membership function for i^{th} feature of t^{th} data point in the j^{th} cluster is defined as;

$$f_j(i,t) = \vee_{l=1}^{L} f_{j,l}(i,t) \tag{9}$$

By rearranging the eq.(8) and eq.(9) the total membership function calculated as;

$$f_j(i,t) = \vee_{l=1}^{L} \sigma(\frac{\mathbb{1}(i \in S_{j,l}) I(S_{j,l}; y_j)\lambda}{d_{t,j}}) \tag{10}$$

As explained in subsection A we again write the eq. (10) in summation form as;

$$f_j(i,t) = \sigma(\sum_{l=1}^{L} \frac{\mathbb{1}(i \in S_{j,l}) I(S_{j,l}; y_j)\lambda}{d_{t,j}}) \tag{11}$$

We can now calculate $f(i,t)$ as done in eq. (5);

$$f(i,t) = \vee_{j=1}^{N} f_j(i,t) \tag{12}$$

By rearranging the eq.(11) and eq.(12) the total membership function is calculated as;

$$f(i,t) = \sigma(\sum_{j=1}^{N} \sum_{l=1}^{L} \frac{\mathbb{1}(i \in S_{j,l}) I(S_{j,l}; y_j)\lambda}{d_{t,j}}) \tag{13}$$

Using this membership function we scale the i^{th} feature of the t^{th} data point as;

$$X^{new}(t,i) = X(t,i)f(i,t) \tag{14}$$

Similar to the previous method, the features which do not appear in any of the $S_{j,l}$ are discarded and this scaled dataset i.e. transformed dataset act as a new dataset with reduced features. The complete approach is described briefly in Algorithm 2 and shown in Fig 1. Let $S_j = \cup_{l=1}^{L} S_{j,l}$ where, S_j is the set of all selected features for the j^{th} cluster.

IV. Experiment

A. Datasets

To validate the performance of the proposed approach the experimentation is done on the six different binary as well as multi-class datasets. We implement our method on multi-class classification problems with continuous valued features. The data sets used have very few samples (in 100s) and large number of features (in 1000s). The data sets used are the 'Arcene' [5], 'ALLAML' [41], 'Colon' [41], 'Prostate-GE' [41], 'TOX' [42] and 'GLIOMA' [41]. The complete details of the datasets are tabulated in the Table I.

B. Implementation Details

The optimal number of clusters were selected using silhouette score analysis where maximum number of clusters ranged till 22. We used a linear SVM classifier with squared hinge loss, L_2 penalty, with $C = 1.0$ and one-vs-rest strategy for multi-class classification, (from sklearn) for making predictions and 5-fold cross-validation accuracy has been reported. Implementations for *mRMR* and entropy estimation were obtained from skfeature library [43]. For calculating accuracy on our method, we have trained a different classifier per cluster since our method advocates solving for cluster-specific scenario. The value of λ is taken to be 10. A constant of 0.0001 was added to each distance to accommodate for the cases when the cluster centers are the data points, and all the distances were divided by the largest cluster center to data point distance to yield appropriately normalized distances.

To implement both proposed methods for the same number of 'equivalent'features (explained below), fixing the parameters of the second method automatically fixes the parameters

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on December 01,2020 at 00:58:40 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Brief description of datasets

Datasets	No. of samples	No. of features	No. of classes
Colon [41]	62	2000	2
Prostate [41]	102	5966	2
TOX [42]	171	5748	4
GLIOMA [41]	50	4434	4
ALLAML [41]	72	7129	2
Arcene [5]	200	10000	2

for the first method. Thus, the parameters are fixed for the second method which are ($t = 10$) feature subsets were selected per cluster with each feature subset being of size ten ($k = 10$). These parameters have been kept constant for all datasets. The actual implementation for selecting top t feature subsets was obtained by modifying the *mRMR* implementation of skfeature [43] library. We used a greedy forward search approach to calculate the top t feature subsets.

C. 'Equivalent' number of features for baseline

For measuring accuracy with respect to the *mRMR* method we need to define the final number of 'equivalent' features selected by our algorithm. Here, we have used two methods to define this 'equivalent' number of features. In the first methods, total number of features of the rescaled feature vector is taken to be the number of 'equivalent' features for each epoch in the cross-validation. An intuitively better approach for selecting the number of 'equivalent' features for *mRMR* is to take the normalized sum of the fuzzy memberships of all the features averaged over the training points and the number of clusters. The expression for number of 'equivalent' features is given by F as;

$$F = \frac{\sum_{t=1}^{n} \sum_{i \in (\cup_{j=1}^{N} S_j)} f(i,t)}{(max_{i,t} f(i,t)) \times N \times n} \tag{15}$$

Such an expression for the number of 'equivalent' features selected by our algorithm is justified because assignment of different memberships to different features means they do not participate equally in the process of classification. Also, membership function quantitatively measures the participation of each feature for each data point. Since different memberships are assigned to a single feature for different clusters as well as different data points so we need to divide the total sum by the product of the total number of data points (n) and the total number of clusters (N). Using this number of 'equivalent' features, we can make a fair comparison of the performances of our methods over *mRMR*.

We try to maintain the same average number of features for both proposed methods (one-feature subset per cluster and t-feature subsets per cluster). The final number of features in the rescaled feature vector obtained by the first method is approximately the product of number of clusters and number of features in each subset ($N \times k$). This number is equated to the total number of features in the scaled feature vector obtained from the second method to calculate the number of features per subset to be supplied to first method in each epoch of the cross-validation.

V. Results and Discussion

The results obtained show that our method performs as well if not better than the original feature selection filter (here is *mRMR*) over the whole dataset. The fact that different features were selected for different clusters justifies the motivation for the problem. Thus, different clusters have a different set of optimal features.

A. Implication of 'equivalent features'

As predicted before, the results when considering the 'equivalent' number of features as total number of features of final rescaled feature vector are not very conclusive, with substantial increase in accuracy in three datasets(GLIOMA, ALLAML, Arcene), slight increase in accuracy in one dataset (Prostate), substantial decrease in accuracy in one dataset (TOX) and miniscule decrease in accuracy in the remaining dataset (Colon) as shown in Table II. This decrease in accuracy is not because of failure of our algorithm but due to the improper choice of number of features for the baseline *mRMR* over the whole dataset.

However, when the number of 'equivalent' features are defined according to fuzzy weights, our method outperforms *mRMR* in all the datasets with a substantial increase in accuracy in three datasets (Arcene, ALLAML, GLIOMA, as shown in Table III. In Table III, the average number of features is calculated by averaging over the number of features of the rescaled output while the number of 'equivalent' features is calculated as described before using fuzzy weights. It turns out that the number of 'equivalent' features is almost equal to the average number of features divided by the number of clusters, which is significantly less than the average number of features. The difference in results highlight the role of choosing 'equivalent' number of features and further demonstrate that the first method does not provide a good baseline for comparison.

The *t*-best feature selection weighs each subset by the subset's mutual information with the response. Thus it could even account for features which are not relevant individually but belong to a relevant feature subset which wasn't accommodated by the single feature subset per cluster method. Due to this, we expected slight increase in accuracy of the second method but it is not seen in the results except for 'Arcene' dataset in Table II and 'Colon' dataset in Table III. This difference between our methods can be better reflected in datasets where the above mentioned kind of features are in abundance. However, in the 'Arcene' dataset in Table III, the accuracy of the second method comes out to be slightly lower than the first method which is not expected but can be justified in terms of many more irrelevant features being selected in the second method over the first. This can be improved by tuning k and t for each dataset separately.

Our method uses a filter algorithm at its core and would thus always perform close to how that algorithm performs. Thus, our algorithm is likely to perform poorly if the filter algorithm used by it is poor. The improvement which our algorithm provides over the original filter is to making it cluster-sensitive.

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on December 01,2020 at 00:58:40 UTC from IEEE Xplore. Restrictions apply.

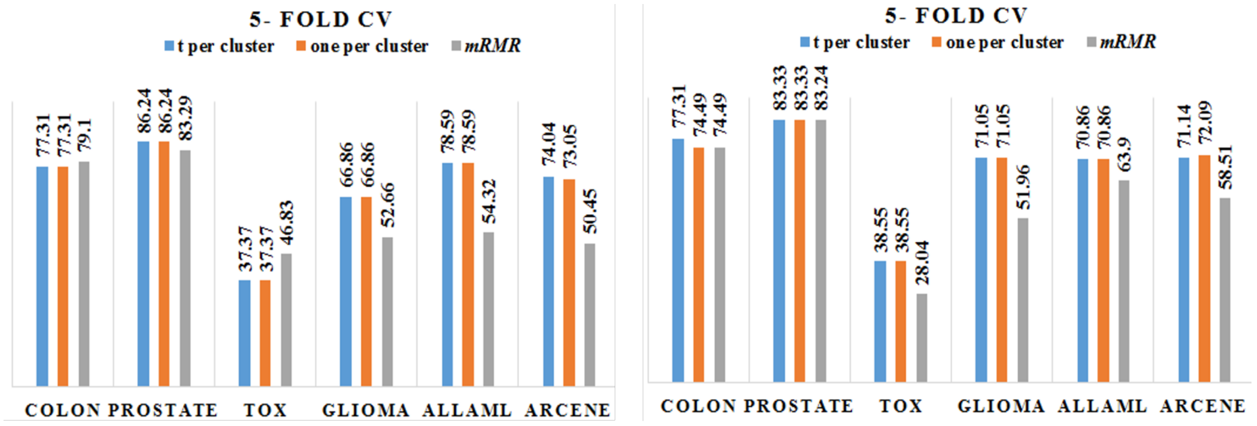

Fig. 2: Bar chart of test accuracy using 5-fold cross-validation with respect to : (a) scaled feature vector after feature selection (b) sum of normalized fuzzy weights of the scaled feature vector after feature selection

TABLE II: Acc. (in %) with respect to number of features in the scaled feature vector after feature selection

Datasets	Clusters	Avg. number of features	*t*-feature subsets per cluster	*one*-feature subset per cluster	*mRMR*
Colon [41]	2	27.6	77.31	77.31	**79.10**
Prostate [41]	2	35.4	**86.24**	**86.24**	83.29
TOX [42]	2	35	37.37	37.37	**46.83**
GLIOMA [41]	2	23.4	**66.86**	**66.86**	52.66
ALLAML [41]	2	26	**78.59**	**78.59**	54.32
Arcene [5]	3	46.8	**74.04**	73.05	50.45

TABLE III: Acc. (in %) with respect to sum of normalized fuzzy weights of the scaled feature vector after feature selection

Datasets	Clusters	Avg. number of features	Number of Equivalent features	*t*-feature subsets per cluster	*one*-feature subset per cluster	*mRMR*
Colon [41]	2	23.8	12.2	**77.31**	74.49	74.49
Prostate [41]	2	36	18	**83.33**	**83.33**	83.24
TOX [42]	2	36	18.2	**38.55**	**38.55**	28.04
GLIOMA [41]	2	24.4	12.4	**71.05**	**71.05**	51.96
ALLAML [41]	2	24.8	12.8	**70.86**	**70.86**	63.90
Arcene [5]	3	48	16.2	71.14	**72.09**	58.51

Even then, our algorithm will not always perform better than the original filter because filters are used for selecting a few features from a high-dimensional input data which has very low number of samples. Almost all filter selection algorithms tend to perform better with increase in number of samples.

In our algorithm, we have decreased the number of samples by using the filter on each cluster of the data, so ideally the filter should perform worse for each cluster, but the increase in accuracy in our result is justified by the subsequent fuzzy combination of features. In some data sets like 'Prostate', the fuzzy combination and reduction in input sample size for filters compensate each other and the filter and our method both give similar accuracy. The most ideal cluster size for our algorithm would be in ranges where the performance of the filter doesn't increase significantly when going from the cluster's data points to the all the data points. The data sets where we have a large increase in our accuracy over the filter indicate the context-sensitivity of the data set. In some crude sense, our method multiplies the less relevant features by a number less than 1, so it reduces the variance of irrelevant features, which is one of the goals of feature selection.

VI. Conclusion and Future Work

The existence of different feature sets for different contexts (clusters) demonstrates the need for handling context-sensitivity. Our algorithm can handle this context sensitivity of feature selection and ensure the optimal combination of features are involved for each data point according to its location in the dataset with respect to the cluster centers. The choice of number of features for comparison was also important. Our definition of 'equivalent' number of features was able to represent the improvement of our algorithm over the baseline much better than the naive selection of total number of features. Decrease in number of samples should decrease quality of features selected by any feature selection algorithm, which is the case when mRMR is applied individually for each cluster. However, the increase in accuracy due to fuzzy combination of the optimal features is still able to overcome this possible decrease. Also, the second method proposed by us makes even more intuitive sense since it operates at feature subset level completely and should be compatible with subset-based feature selection methods. The first method is poor for subset-based feature selection, but it can be very easily used in scenarios where individual features are independently assigned scores. The flexibility of our methods allow extension and possible improvement in performance of most filter feature selection algorithms.

However our method can still be further improved by taking into account redundancy after the selection of cluster-specific optimal subsets is required. This notion of fuzzy combination of context-sensitive entities can be extended to not just the features but the classifier as well. The current implementation focuses on a hard approach, in which a different classifier is trained for each cluster but their fuzzy combination is not taken. Changing the core filter algorithm can change the performance of the method. The better the core algorithm, the better should be our method. Essentially, our method requires

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on December 01,2020 at 00:58:40 UTC from IEEE Xplore. Restrictions apply.

a feature selection algorithm which can output the optimal feature set along with the scores for either each feature or the complete feature set.

Another extension for our algorithm could be using fuzzy combination of optimal feature subsets selected by different feature selection algorithms. Extension of our method beyond filter feature selection to wrapper and embedded feature selection methods will also prove to be challenging. Our method shows better performance on standard datasets while selecting a small number of 'equivalent' features than the actual filter algorithm at its core. Thus context-based feature selection and subsequent fuzzy combination offer a promising method for context-based treatment of the feature selection problem.

REFERENCES

[1] A. C. Harrington, C. Rosenow, and J. Retief, "Monitoring gene expression using DNA microarrays," Current opinion in Microbiology, vol. 3, no. 3, pp. 285-291, June 2000.

[2] W. P. Kuo, E. Y. Kim, J. Trimarchi, , T. K. Jenssen, S. A. Vinterbo, and L. Ohno-Machado, "A primer on gene expression and microarrays for machine learning researchers," Journal of Biomedical Informatics, vol. 37, no. 4, pp. 293-303, August 2004.

[3] Z. M. Hira, and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," Advances in Bioinformatics, May 2015.

[4] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano et al., "Machine learning in bioinformatics," Briefings in Bioinformatics, pp. 86-112, March 2006.

[5] I. Guyon, and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning research, vol. 3, pp. 1157-1182, March 2003.

[6] A. L. Blum, and P. Langley, "Selection of relevant features and examples in machine learning," Artificial intelligence, vol. 97, no. 1, pp. 245-271, December 1997.

[7] R. Kohavi, and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence vol. 97, no. 1-2, pp. 273-324, December 1997.

[8] P. E. Meyer, C. Schretter and G. Bontempi, "Information-Theoretic feature selection in Microarray data using Variable Complementarity," IEEE Journal of Selected Topics in Signal Processing, vol. 2 no. 3, June 2008.

[9] C. Ding, and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," Journal of Bioinformatics and Computational Biology, vol. 3, no. 2, pp. 185-205, April 2005.

[10] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, August 2005.

[11] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Transactions on Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.

[12] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," Journal of Machine Learning Research, vol. 5, pp. 1531-1555, November 2004.

[13] P. Jafari and F. Azuaje, "An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors," BMC Medical Informatics and Decision Making, vol. 6, no. 1, pp. 27, June 2006.

[14] M. A. Hall, "Correlation-based feature selection for machine learning," University of Waikato Hamilton, April 1999.

[15] J. E. R. Hruschka, E. R. Hruschka, and N. F. F. Ebecken, "Feature selection by Bayesian networks," Conference of the Canadian Society for Computational Studies of Intelligence, pp. 370-379, May 2004.

[16] A. Rau, F. Jaffrezic, J. L. Foulley, and R. W. Doerge, "An empirical Bayesian method for estimating biological networks from temporal microarray data," Statistical Applications in Genetics and Molecular Biology, vol. 9, no. 1, January 2010.

[17] P. Yang, B. B. Zhou, Z. Zhang, and A. Y. Zomaya, "A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data," BMC Bioinformatics, vol. 11, no. 1, January 2010.

[18] T. Jirapech Umpai and S. Aitken, "Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes," BMC Bioinformatics, vol. 6, no. 1, p. 148, December 2005.

[19] C. H. Ooi and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," Bioinformatics, vol. 19, no. 1, pp. 3744, January 2003.

[20] T. C. Lin, R. S. Liu, Y. T. Chao and S. Y. Chen, "Classifying subtypes of acute lymphoblastic leukemia using silhouette statistics and genetic algorithms," Gene, vol. 518, no. 1, pp. 159-163, April 2013.

[21] H. Glass and L. Cooper, "Sequential search: a method for solving constrained optimization problems," Journal of the ACM, vol. 12, no. 1, pp. 7182, January 1965.

[22] D. Kleftogiannis, K. Theofilatos, S. Likothanassis, and S. Mavroudi, "YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features," IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), vol. 12, no. 5, pp. 1183-1192, September 2015 .

[23] C. Furlanello, M. Serafini, S. Merler and G. Jurman, "Semi-supervised learning for molecular profiling" IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 2, no. 2, pp. 110-118, April 2005.

[24] R. K.Sevakula, V. Singh, N. K. Verma, C. Kumar and Y. Cui, "Transfer Learning for Molecular Cancer Classification using Deep Neural Networks", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018. (Accepted)

[25] K. B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for gene selection in cancer classification with expression data," IEEE transactions on Nanobioscience, vol. 4, no. 3, pp. 228-234 September 2005.

[26] S. Rajurkar, V. Singh, N. K. Verma and Y. Cui, "Deep stacked autoencoder with deep fuzzy network for transcriptome based tumor type classification", BMC Bioinformatics, vol. 18, 2017

[27] L. K. Luo, D. F. Huang, L. J. Ye, Q. F. Zhou, G. F. Shao and H. Peng, "Improving the computational efficiency of recursive cluster elimination for gene selection" IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 1, pp. 122-129, January 2011.

[28] U. Maulik, A. Mukhopadhyay, and D. Chakraborty, "Gene-expression-based cancer subtypes prediction through feature selection and transductive SVM," IEEE transactions on Biomedical Engineering, vol. 60, no. 4, pp. 1111-1117, April 2013.

[29] L. Huan and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," Proceedings of Seventh International Conference on Tools with Artificial Intelligence, pp. 388-391, November 1995.

[30] W. Sewall, "The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating," Evolution, vol. 19, no. 3, pp. 395-420, September 1965.

[31] U . Alon, N . Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays," Cell Biology, vol. 96, no. 12 pp. 6745- 6750, Jun 1999.

[32] A.Blum and P. Langley, "Selection of relevant features and examples in machine learning" Artificial Intelligence, vol. 97, no. 1-2 pp. 245- 271, December 1997.

[33] P. Melin and O. Castillo, "Type-1 Fuzzy Logic. In: Hybrid Intelligent Systems for Pattern Recognition Using Soft Computing. Studies in Fuzziness and Soft Computing," Springer, Berlin, Heidelberg, vol. 172, 2005.

[34] L. A. Zadeh, "Fuzzy sets," Information and control, vol. 8, no. 3, pp. 338-353, June 1965.

[35] V. Singh, R. Dev, N. K Dhar, P. Agrawal and N. K. Verma, "Adaptive Type-2 Fuzzy Approach for Filtering Salt and Pepper Noise in Grayscale Images", IEEE Transactions on Fuzzy Systems, 2018. (Accepted for Publication)

[36] N. K. Verma, and M. Hanmandlu, "From a gaussian mixture model to non-additive fuzzy systems," IEEE Transactions on Fuzzy Systems, vol. 15, no. 5, pp. 809-827, October 2007.

[37] R. K. Sevakula and N. K. Verma, "Compounding General Purpose Membership Functions for Fuzzy Support Vector Machine Under Noisy Environment," IEEE Transactions on Fuzzy Systems, 25(6), pp.1446-1459, 2017.

[38] D. Arthur, and S. Vassilvitskii, "k-means++: The advantages of careful seeding," In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, pp. 1027-1035. January 2007.

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on December 01,2020 at 00:58:40 UTC from IEEE Xplore. Restrictions apply.

[39] P. J Rousseeuw, “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis,” Journal of Computational and Applied Mathematics, pp. 53-65, November 1987.

[40] T. J Ross. Fuzzy logic with engineering applications.John Wiley and Sons,2009.

[41] http://featureselection.asu.edu/datasets.php

[42] E. Y. Kwon, S. K. Shin, Y. Y. Cho, U. J. Jung et al,“Time-course microarrays reveal early activation of the immune transcriptome and adipokine dysregulation leads to fibrosis in visceral adipose depots during diet-induced obesity,” BMC Genomics, vol. 13, no. 1, September 2012.

[43] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, and H. Liu, “Feature selection: A data perspective,” ACM Computing Surveys (CSUR), vol. 50, no. 6, p. 94, December 2017.

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on December 01,2020 at 00:58:40 UTC from IEEE Xplore. Restrictions apply.