cluster are more similar even though their geometric distance is large.

This process of combining multiple clusters of a set of objects without accessing the original features results in a new feature space based on the cluster labels in each iteration. This consensus of clustering is called Ensemble Clustering (EC). Combination clustering is a more challenging task than the combination of supervised classifications. Topchy et al. [6] and Strehl and Ghosh [7] addressed this issue by formulating consensus functions that avoid an explicit solution to the corresponding problem.

A clustering based learning method was proposed by Derbeko et al. [8]. In this study, several clustering algorithms are run to generate several (unsupervised) models and the labelled data is than applied to entire clusters (making the assumption that all points in the particular cluster have the same label). In this way, the algorithm develops a number of hypotheses. The one that minimizes the PAC-Bayesian boundary is the one chosen to be used as the classifier under the assumption that at least one of the clustering runs will produce a good classifier and that the algorithm will find it.

Abd-Allah and Shimshoni [9], used the ensemble clustering methods within the k-nearest neighbor classifier by developing a distance function based on ensemble clustering. Then, based on this classifier they developed a selective sampling algorithm that selected the most informative samples from an unlabelled dataset to be labelled by a teacher in order to improve the training data [10].

The first study using RCE–Recursive cluster elimination proposed by Yousef et al. [5,6] and also suggests using gene networks for clustering the genes with RCE. Both studies have reported a high performance compared with other similar methods.

A recent study by Yousef et al. [11] has used EC classification on plant microRNA data and, compared it to SVM and one-class classifiers. They showed that EC-K Nearest Neighbors (EC-KNN) outperforms all other methods indicating that the EC makes a significant contribution to the resulting higher accuracies.

The most recent study in this field Du et al. [12], proposed a hybrid feature selection method based on Multiple Kernel Learning (MKL) and was favourably compared with several different approaches. We have used the same data (they kindly provided) to test our new approach and have compared our results to their published MKL results and other reported results.

Results

Data used for assessment of classification accuracy

We have used the same data sets from the experiment by Du et al. [12] that are available from the link: http://csbl.bmb.uga.edu/ICSB/SMKL-FS/index.html. This data consists of three types of gene expression data that were obtained by Du et al. [12] from Gene Expression Omnibus (GEO) by Barrett et al. [13] and the Cancer Genome Atlas (TCGA) by Weinstein et al. [14]. The paired samples in the expression datasets only considered which were from tumor and which from adjacent non-tumor tissues to form the two-classes for the classification algorithm. The following steps are the summary of the pre-processing procedure applied on the data set by Du et al. [12]:

1. Missing value stage: For each missing value that is less than 20% of the sample, these values are estimated as the local least squares imputation (LLSimpute) method. Then, the different probes for the same mRNA (or miRNA) are merged by the maximum expression value of these probes for each sample.

2. Normalization stage: The median absolute deviation (MAD) method is applied to normalize expression between samples.

More information on the pre-processing steps applied on the datasets has shown in Du et al. experiment [12]. Table 1 list the eight cancer types of mRNA microarray datasets including the number of samples and GEO id.

The second data set analyzed includes mRNA and miRNA sequencing results from the following 8 datasets:

KIDNEY (88), BRCA (71), LUNG (47), HNSC (37), LIHC (46), PRAD (43), STAD (29) and THCA (56) (Values in parenthesis are the equal number of samples for both classes). These data are also obtained from Du et al. experiment [12].

Comparing results using SVM-RCE-EC and different classification approaches

The SVM-RCE-EC approach was compared to 8 previously reported methods [12] including: SVM-RFE [15], SVM-RCE [4], mRMR [16], IMRelief [17], SlimPLS [18] and SMKL-FS [12] using 10-fold Cross-Validation (CV) on a variety of datasets. The performance is calculated as the mean effectiveness measurement [12].

Figure 1 compares the reported performance of the 8 other methods on the 8 gene expression datasets from Du et al. [12]. SVM-RCE-EC outperforms the best results obtained by SMKL-FS as reported by the study by about 7% [12] and outperforms the poorest results obtained by OSFS by about 28%. It also outperforms SVM-RCE by 13%. It

Cancer Data Set Type	#Tumor	# Non-tumor
LiverGSE5364, GSE22058, GSE14520, GSE12941	132	132
PancreaticGSE15471, GSE16515, GSE22780	63	63
Lung GSE5364, GSE19804, GSE22058, GSE10072, GSE7670, GSE2514	249	249
Colon GSE5364, GSE8671, GSE25070, GSE21510, GSE23878, GSE18105	70	70
Gastric GSE13911, GSE13195, GSE5081, GSE19826	93	93
Breast GSE5364, GSE15852, GSE10810, GSE16873, GSE5764, GSE14548	113	113
Thyroid GSE5364, GSE3678	23	23
Prostate GSE6919, GSE6956, GSE17951	88	88

Table 1: mRNA microarray datasets from GEO genomics data repository with the number of samples and GEO ID for each data set. The number of tumor samples and non-tumor are present and they are equal.

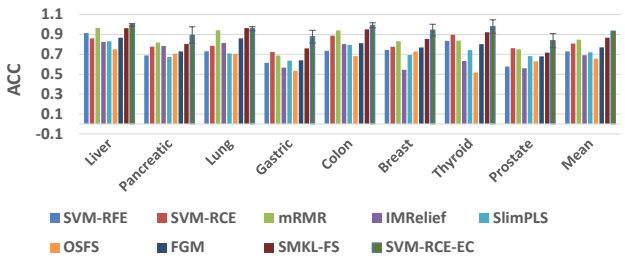

Figure 1: The average of the last 10 levels (mean effectiveness) applied to mRNA microarray data from Table 1. The first 8 columns of results are from Ding and Pang [16]. SVM-RCE-EC includes a standard deviation bar to indicate the deviation of the results per iteration.

should be noted that SVM-RCE-EC achieved high performance with the difficult Prostate data set, where most of the other methods failed and it outperformed SMK-FS on this data set by 12%. SVM-RCE-EC performance with the Gastric dataset reached 88% while most of the other methods performed poorly.

Figure 2 presents the results of all the methods applied on mRNA sequencing data. The SVM-RCE-EC performance compared to SMKL-FS is higher on average by about 2% and compared to SVM-RCE by about 3%. While these differences are small, even small improvements in accuracy can have meaningful clinical implications. SVM-RCE-EC performance is higher by 12% compared to OSF and SlimPLS, the lowest performing methods. Overall, most of the methods perform well with the data used for Figure 2 indicating that most of those data are not difficult to separate.

Figure 3 presents the results from the analyses of microRNA gene expression. In this dataset, we find similar performance on average between SVM-RCE and SMKL-FS. It should be noted that SVM-RCE-EC performs better than all other methods on the STAD and LIHC data, SMKL-FS slightly outperformed SVM-RCE-EC on the first four cancer types listed. On the STAD data, SMKL-FS achieved an accuracy of 0.880, while SVM-RCE-EC achieved an accuracy of 0.93-an improvement of 5%.

We also show that the accuracy of both SVM-RCE and SVM-RCE-EC improves as the number of cluster/genes is decreased (Figure 4).

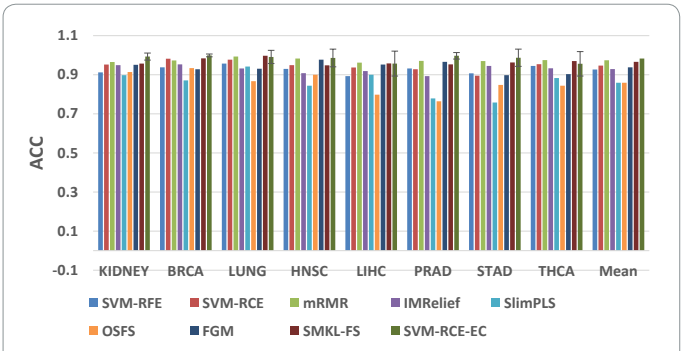

Figure 2: The average of the last 10 levels (mean effectiveness) applied on mRNA sequencing. The first 8 columns results are copied from Ding and Pang's experiment [16]. SVM-RCE-EC includes a standard deviation bar to indicate the deviation of the results per iteration.

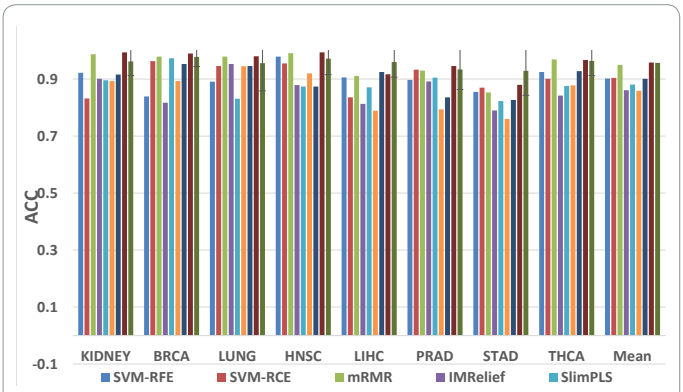

Figure 3: The average of the last 10 levels (mean effectiveness) applied on miRNA sequencing. The first 8 columns results are copied from Ding and Pang's experiment [16]. SVM-RCE-EC includes a standard deviation bar to indicate the deviation of the results per iteration.

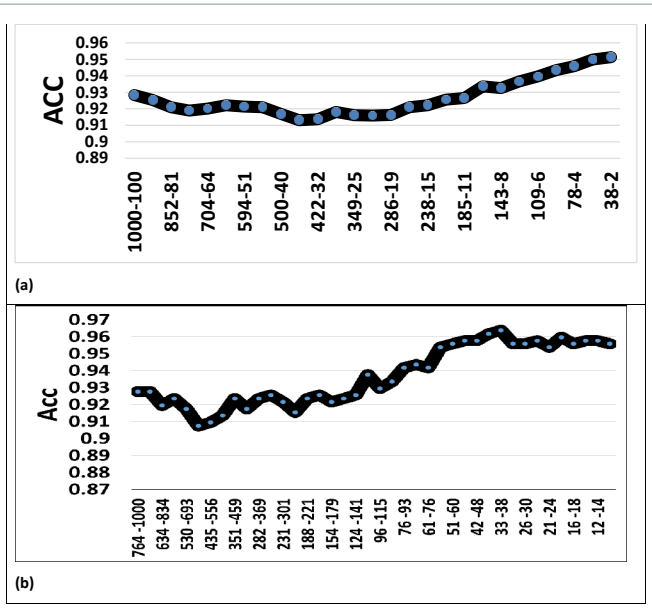

Figure 4: (a) The performance of SVM-RCE on the lung dataset over different level of clusters/genes. (b) The performance of the SVM-RCE-EC on the lung dataset over different level of clusters/genes. The x-axis is the cluster/genes levels while the y-axis s is the accuracy (ACC) values.

This is an indication that the selection of the optimal number of gene clusters needed for the best possible solution is progressing as less informative clusters are removed. It should be noted that while the decrease in the number of clusters in SVM-RCE is fixed as 10%, the clusters that are removed at each level in SVM-RCE-EC is not fixed but is determined by the ensemble procedure.

Looking inside the gene clusters

Figure 5 is a heatmap showing the expression patterns of the significant genes identified by SVM-RCE-EC and SVM-RCE. There are 27 genes in common between the top 50 significant genes selected by both SVM-RCE and SVM-RCE-EC. This is perhaps not expected given the similar performances in approximately ½ of the data sets analyzed. There are however, some differences in the 2 gene sets that are worth noting. The most significantly changed genes in both data sets are the genes under expressed in liver cancers. While 11 out of 50 gene are under expressed among the 50 SVM-RCE (-5.7 to -23. 6 fold) genes only 5 of them are included in the top 50 SVM-RCE-EC genes (-6.6 to -23.2 fold). All 5 genes products are found in the extra-cellular space with one of the 5, MTH1 also being nuclear. By contrast the most highly up regulated of the SVM-RCE gene is SQLE (3.9 fold in cancer) which is also in the SVM-RCE-EC list while RRM2 (6.5 fold) is only in the SVM-RCE-EC list. In addition, 23 of the 50 SVM-RCE-EC have nuclear localizations, with overlapping regulatory functions associated with transcription and maintenance of chromatin structure, compared to just 12 for the SVM-RCE list. The SVM-RCE-EC genes overall have more robust differences in expression between cancers and controls. Shown in the additional file top 50 GenesLiver.xls.

Evaluation

The process for evaluating the over-all performance of SVM-RCE-EC is described in the Materials and Methods and illustrated in Figures 6-8. Briefly, we used 10-fold cross validation (9 fold for training and 1 fold for testing). After each round of feature selection or cluster reduction, the accuracy was calculated on the independent hold-out test

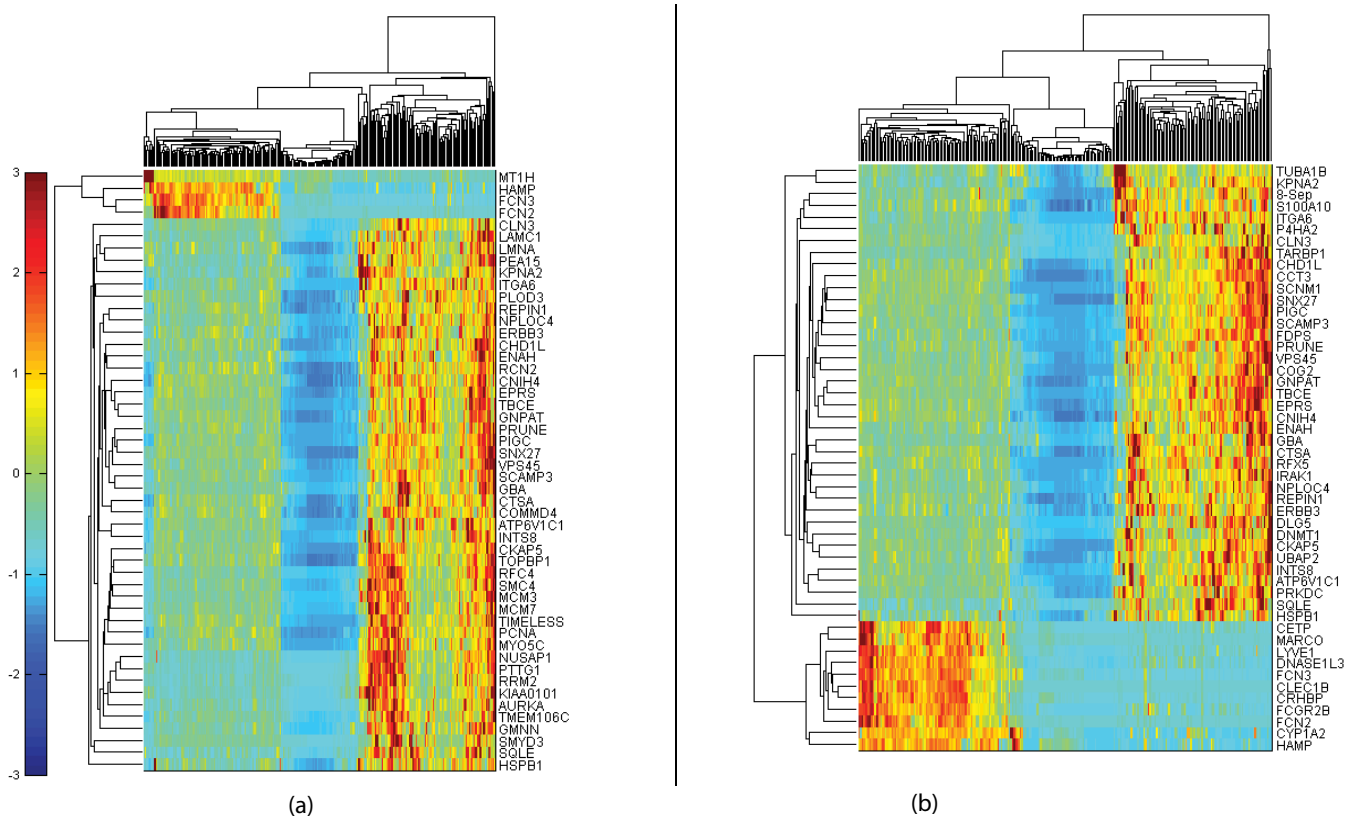

Figure 5: The Heatmap (Hierarchal clustering) of top 50 genes discovered by SVM-RCE-EC (panel a) and SVM-RCE (panel b) on the liver data set. The vertical lines indicate the clustering of the samples while the horizontal lines indicate the clustering of the genes.

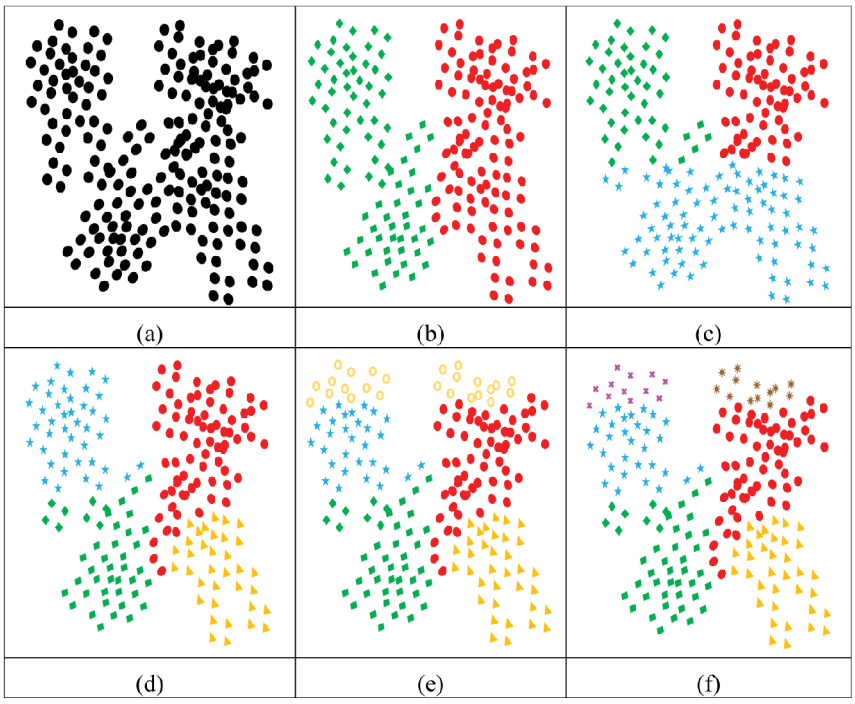

Figure 6: Example of Euclidean distance and Clustering based distance methods to differentiate the points into clusters. (a) The original data, (b) k-means with k=2, (c) k-means with k=3, (d) k-means with k=4, (e) k-means with k=5, (f) the equivalence classes.

Algorithm SVM-RCE-EC (input data *D*)
X= the training dataset
s = genes list (all the genes) or top *n_g* genes by t-test
n = infinity number
m = final number of clusters
d = the reduction parameter
While ($n >= m$) do

1. n= Cluster the given genes S into clusters $S_1, S_2, \ldots, S_n$ using Ensemble Clustering (**EC Clustering step**)
2. For each cluster $i=1..n$ calculate its $Score(X(s_i), f, r)$ **(SVM scoring step)**
3. Remove the *d%* clusters with lowest score (**RCE step**)
4. Merge surviving genes again into one pool *S*

Figure 7: The pseudo code of the SVM-RCE-EC Algorithm.

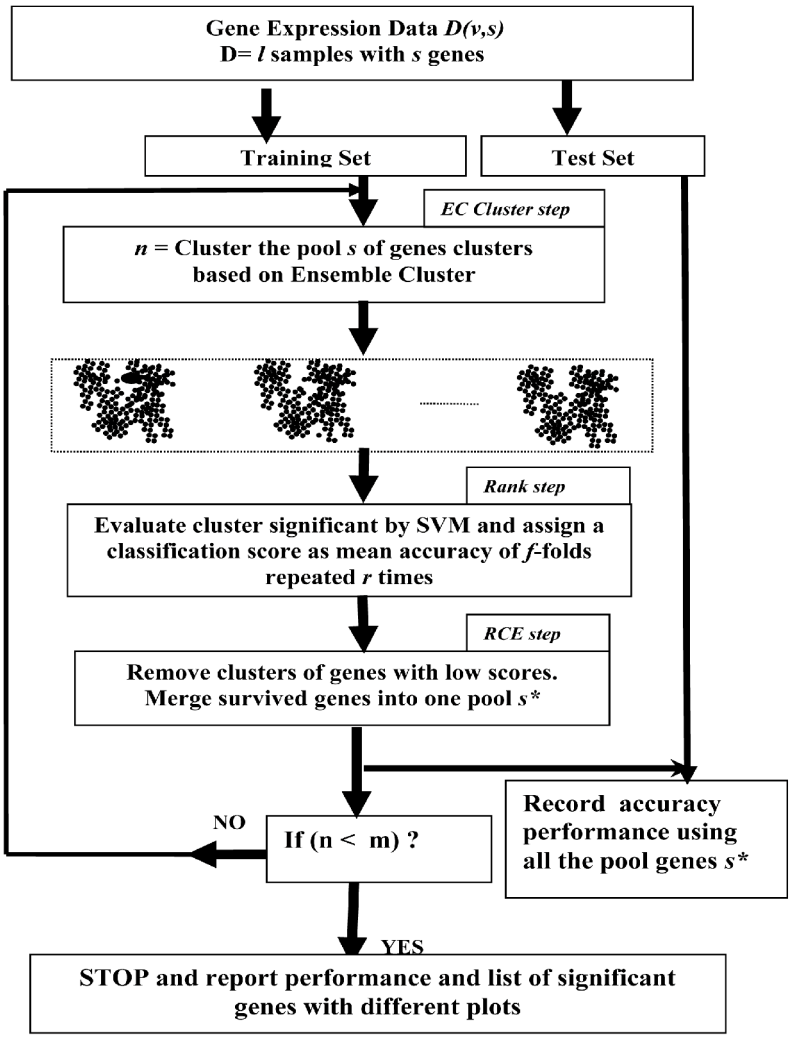

Figure 8: The description of the SVM-RCE-EC algorithm. A flowchart of the SVM-RCE-EC algorithm consists of main three steps: the EC-Cluster step for clustering the genes based on ensemble clustering, the Rank step for assessment of significant clusters and the RCE step to remove clusters with low rank.

set (Figure 8). The classifier performance was calculated with sensitivity (SE) and specificity (SP) and accuracy (ACC) statistics as follows:

TP=Number of true positives examples.

FP=Number of false positives examples.

TN=Number of to true negatives examples.

FN=Number of false negative examples.

SE=TP/(TP+FN), SP=TN/(TN+FP), ACC=(TP+TN)/(TP+TN+FP+FN)

SVM-RCE-EC produces a series of feature subsets and the algorithm evaluates each subset the SE, SP and ACC. To have a general evaluation of performance we consider the mean effectiveness measurement which is the average of last k subset accuracies [12]. We have chosen k to be 10 folds suggested in Du et al. experiment [12].

Conclusion

This study presents an improved method of SVM-RCE called SVM-RCE-EC for classification of gene expression datasets by selecting significant clusters of genes based on the ensemble clustering approach. SVM-RCE-EC demonstrated improved classification accuracy compared to other methods tested (SVM-RFE [15], SVM-RCE [4], mRMR [16], IMRelief [17], SlimPLS [18] and SMKL-FS [12]) reported by the study of Du et al. [12] and this was particularly true for data sets where the 2 classes were difficult to separate and where other methods either failed or had low performance.

SVM-RCE-EC is a search method that queries a space that consists of gene clusters rather than individual genes in order to identify those clusters of potentially interacting genes that contribute most to the differences in phenotypes to return improved classification performance in distinguishing 2 different sample classes. We find that the ensemble approach, which measures the frequency of which genes group together in order to identify the most significant clusters, provided more robust, informative clusters compared to several other methods used alone or even by combining methods.

Embedding the ensemble clustering approach within the original SVM-RCE has allowed us to solve the problem of how to determine the appropriate numbers of clusters that will be retained/eliminated with each iteration, rather than leaving the decision of how many clusters are retained/eliminated to the user. The cluster size is determined arbitrarily at the onset of the analysis by the investigator and, as the algorithm proceeds, the least informative clusters are progressively removed. Moreover, SVM-RCE-EC does not produce redundant clusters, as is sometimes the case for SVM-RCE. In summary SVM-RCE-EC using the ensemble cluster approach appears to be a promising method particularly when dealing with data from very similar sample classes that are difficult to separate. In the current version of SVM-RCE-EC, the hamming distance was used to group the genes together, although it is possible to apply different measurements and explore their influence on the classification performance.

Methods

SVM-RCE-EC uses an ensemble clustering method, to identify robust gene clusters, and Support Vector Machines (SVMs) [19], to score (rank) those gene clusters for accuracy of classification. SVM-RCE-EC is using ensemble clustering to group genes into clusters. After scoring of the clusters by SVM, the clusters with lower scores removed. The remaining considerable are then moved to the next ranking step.

Ensemble clustering

The previous method uses the k-means clustering algorithm for the initial gene grouping. We found that genes belonging to the same cluster usually share some common traits even though their geometric distance might be large. Since the clustering results of the k-means clustering algorithm are affected by the initial centroids that were selected randomly, we explored the effects of running the k-means several times with different k values. This problem of combining multiple clustering of a set of objects without accessing the original features is called cluster ensemble. Abd-Allah and Shimshoni [10] developed a new method that combines several clustering results in a matrix, called clustering matrix, and define a distance function between the objects based on this matrix and then tested it using the nearest neighbor classifier. Other empirical ensemble clustering methods were also described in [6,7,20]. The following example illustrates this situation (Figure 6). Considering the dataset in Figure 6a the number of clusters is unknown, the clusters are not elliptic and the number of points within each cluster is unbalanced. Therefore, running k-means with fixed k is not a good idea and in high probability, we will get poor clusters. Therefore, if we run the k-means clustering algorithm different values of the parameter k we might cover all the cases. In the following example we decided to run k=2, 3, 4, 5 as described in Figures 6b-6e. Figure 6f describes the shared points that belong together in all the iterations.

The SVM-RCE-EC methods

The algorithm of SVM-RCE-EC is described at Figures 7 and 8 where D is the data with *s* the genes. The training set is *X* that consists of *l* samples. The Score *(X(s), f, r)* function is the average accuracy for *f*-folds cross validation of the linear SVM [19], repeated *r* times (default values to 3 and r to 5). We apply Score *(X(s), f, r)* on each clusters of genes as the Score $S_1, S_2, \ldots, S_n$ as the Score ($X(s_i)$, *f, r*).

The central algorithm of SVM-RCE-EC is based on the SVM-RCE algorithm [4]. However, SVM-RCE-EC differs from SVM-RCE in two main aspects. The first one is the method for grouping the genes and second is the number of clusters in each iteration. In SVM-RCE, one can control the number of clusters at each iteration. In SVM-RCE-EC the number of clusters is determine by the method itself and that number does not necessarily diminish at different (iterations) levels as demonstrated in Figure 7, step (1) where n is determined by the method whereas in the SVM-RCE n is defined by the user.

We have considered the k-means [21] clustering method as the clustering step with SVM-RCE and SVM-RCE-EC.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of supporting data

Results_SVN-RCE-EC vs others.xls: The results in numeric numbers of Figures 1-3.

Finalaccurecy_micro_miRNA_Sequencing.xls: The results for SVM-RCE-EC applied on micro miRNA sequencing datasets.

Finalaccurecy_micro_mRNA_microarray.xls: The results for SVM-RCE-EC applied on micro mRNA microarray datasets.