Proceedings of the 6th International Conference on Statistics: Theory and Applications (ICSTA'24)
Barcelona, Spain – August 19 - 21, 2024
Paper No. 144
DOI: 10.11159/icsta24.144

Enhancing Gene Expression Data Analysis through SVM-Based Recursive Cluster Elimination and Weighted Center Approaches

Malik Yousef[1,2], Nurten Bulut[3], Burcu Bakir Gungor[3], Bahjat F. Qaqish[4]
[1]Department of Information Systems, Zefat Academic College
Zefat, Israel
[2]Galilee Digital Health Research Center, Zefat Academic College
Zefat, Israel
malik.yousef@gmail.com
[3]Department of Electrical and Computer Engineering, Abdullah Gul University
Kayseri, Turkey
nurten.bulut@agu.edu.tr; burcu.gungor@agu.edu.tr
[4]My Department of Biostatistics University of North Carolina at Chapel Hill
NC, Chapel Hill, USA
bahjat_qaqish@unc.edu

Abstract **-** The complexity and high dimensionality of gene expression data pose significant challenges for effective feature selection and accurate classification in bioinformatics. This study introduces two novel algorithms, Support Vector Machine-Recursive Cluster Elimination (SVM-RCE) and its advanced version, SVM-RCE with Center Weights (SVM-RCE-CW), designed to optimize feature selection by leveraging clustering techniques and machine learning models. Both algorithms aim to reduce the feature space, thereby enhancing the interpretability and performance of classification models. We present a comprehensive comparison of these methods against traditional feature selection techniques, demonstrating their efficacy in achieving significant dimensionality reduction while maintaining or improving classification accuracy in several gene expression datasets.

Keywords**:** Recursive Cluster Elimination, Feature Selection, Clustering, Gene Expression Data Analysis.

1. Introduction

It The analysis of gene expression data is pivotal in understanding biological processes and disease mechanisms. However, the high dimensionality of such data often leads to overfitting and complicates model interpretation. Traditional feature selection methods struggle to handle the complex dependencies among features effectively. Informative Feature Clustering and Selection (IFCS), a method proposed for gene expression data analysis, involves two steps: computing feature weights to determine gene importance and selecting genes from different gene clusters using a stratified feature selection method. The superiority of IFCS over six popular feature selection methods was demonstrated in several gene expression datasets [1].

In [2], an approach to simultaneous feature selection and semi-supervised clustering was developed by formulating the problem as a multi objective optimization task and utilizing a multi objective optimization technique called AMOSA.

Another study proposed an extension to a pathway analysis method, significance analysis of microarray gene set reduction (SAMGSR), for feature selection in longitudinal gene expression data [3]. By applying the reduction step twice, they aimed to select relevant genes for longitudinal omics data. Their work bridges feature selection and gene set analysis, providing a pathway-based approach to feature selection.

Other researchers [4], addressed the challenge of selecting relevant genes from big gene expression data using a bi-dimensional principal gene feature selection method. Their method aimed to improve computing efficiency by extracting relevant and important genes from large-scale gene expression data.

Researchers developed a multi filter model of feature selection to enhance colon cancer classification [5]. Their two-stage approach involved gene selection before classification techniques were applied, resulting in improved success rates for identifying cancer cells. They utilized Information Gain and a Genetic Algorithm for feature selection and applied the minimum Redundancy Maximum Relevance (mRMR) technique for gene ranking.

Yousef et al. [6], [7] propose a method called Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data. The authors address the challenge of identifying informative genes and reducing the dimensionality of gene expression data to improve classification accuracy. The RCE method involves the following steps: feature clustering, feature selection and recursive elimination. In the clustering step, the gene expression data are initially clustered using a k-means clustering algorithm. This step aims to group genes with similar expression patterns together. In the feature selection step, a feature selection algorithm is applied within each cluster to identify the most informative genes. A support vector machine (SVM) classifier is used to evaluate the importance of each gene in the classification task. In recursive elimination, the least informative genes are eliminated from each cluster, and the process is repeated iteratively. This recursive elimination step helps refine the gene set and improve classification performance. The performance of RCE is evaluated in benchmark datasets by comparing its classification accuracy with other feature selection methods, such as t-tests and principal component analysis (PCA). They demonstrate that RCE outperforms these methods in terms of classification accuracy and number of selected features.

In [8], Yousef et al. introduced a novel method denoted SVM with Recursive Network Elimination (SVM-RNE). They showed that integrating network information within SVM-based recursive feature reduction yields notable improvements in both performance and biological interpretability.

In SVM-RCE-OPT [9], the aim is to ascertain the optimal weights for the scoring function initially proposed in the SVM-RCE-R study [7], employing advanced optimization methods.Optimizing the scoring function's weights can significantly enhance the overall performance of SVM-RCE in a majority of scenarios. In certain instances, substantial improvements were observed in accuracy and AUC metrics.

This study introduces a novel algorithm, the Support Vector Machine-Recursive Cluster Elimination with Center Weights (SVM-RCE-CW). SVM-RCE-CW integrates support vector machine (SVM) classification, recursive feature elimination and clustering techniques, with a focus on using feature cluster centers for assessing cluster importance.

2. Methods

2.1. The original SVM-RCE algorithm

The Support Vector Machine-Recursive Cluster Elimination (SVM-RCE) algorithm is a method for feature selection and dimensionality reduction, particularly in gene expression data analysis. Its objective is to enhance the performance of classification models by iteratively removing less informative features (genes) based on their clustering and predictive power. The process involves several stages:

Clustering: Initially, the algorithm applies a clustering technique, such as K-means, to the features, grouping genes into clusters based on their expression patterns. The motivation that genes within a cluster exhibit similar expression profiles and potentially contribute similarly to the biological condition being studied.

Data Subset Creation: For each cluster identified in the first step, a data subset consisting of data on all samples but only the features in the cluster, is extracted from the training data.

Evaluation and Scoring: The importance of each cluster in the prediction task is its assigned score. In the original version of SVM-RCE, the score was the predictive accuracy of a model that uses genes in the cluster as predictors. The training is repeated for each cluster. However, in the updated version named SVM-RCE-CW (with "CW" denoting "Center Weights"), predictive accuracy is assessed in a linear model with cluster centers (cluster means) as predictors. The estimated coefficients serve as scores for the respective clusters. The training is done only once. This reflects a move towards understanding the importance of each cluster adjusted for the other clusters.

Elimination: Clusters deemed least important based on their scores are eliminated, and their constituent genes are removed from the dataset. This results in a reduced feature set for the next iteration.

Iteration: The entire process is repeated with the reduced set of features, progressively refining the feature space by eliminating the least informative clusters in each round.

Final Model Training: Feature elimination concludes when either a predetermined number of features is reached or when all but the most critical cluster have been removed. Then, a final model is trained on the reduced dataset. This model is expected to be more efficient, potentially more accurate, and easier to interpret than one trained on the full set of features. The pseudo code of SVM-RCE is given in Table 1.

Table 1: The pseudo code of SVM-RCE

```
Procedure SVM-RCE (D, $initial_k$, $d_{percent}$):
  Input:
    D: Gene expression dataset with samples as rows and genes as columns.
    $initial_k$: Initial number of clusters to form.
  number_of_clusters_schedule: A vector containing a decreasing sequence of integers that controls the number of clusters retained in each round.

  Output: Final reduced data set and the performance of the SVM on it

  Step 1: Initialization
  - Split the samples (rows of D) randomly into $D_{train}$ (90%) and $D_{test}$ (10%).
  - k = $initial_k$

  Step 2: Clustering and Sub-dataset Creation
  While k > 1 begin:
    - Apply K-means to cluster the genes (columns) in $D_{train}$ into k clusters.
    - For i=1 to k, form data subset $S_i$ consisting of all the rows of $D_{train}$ but only the columns that correspond to cluster i.

  Step 3: Scoring
      - Perform internal cross-validation (CV) by randomly splitting the rows of $S_i$ into training (90%) and testing (10%) parts.
      - Repeat the CV process f times (f=5 is recommended).
      - Compute the mean accuracy across the f iterations.
      - The mean accuracy serves as the score assigned to cluster i.
      - End for.

  Step 4: Cluster Elimination
    - Define n_retain  to be the number of clusters to be retained according to the number of clusters schedule.
      Compute n_remove = k - n_retain.
    - Remove the lowest scoring  n_remove clusters from $D_{train}$ and $D_{test}$
      forming $D_{train}$* and $D_{test}$*. Removing a cluster means removing all its genes.

  Step 5: Model Training and Testing on Reduced Dataset
    - Train an SVM model on $D_{train}$*.
    - Test the trained SVM on $D_{test}$*.
    - Evaluate the performance of the SVM (e.g., accuracy, precision, recall).

  Step 6: Preparation for Next Iteration
    - Update $D_{train}$ and $D_{test}$ to $D_{train}$* and $D_{test}$ *.
    - Reduce k according to the elimination scheme (e.g., k = k - num_to_remove).
 End while  (k = 0)
  Step 7: Return the final reduced data set and the performance of the SVM on it.
```

2.2. SVM-RCE-CW: SVM-RCE with Center Weights

The Support Vector Machine-Recursive Cluster Elimination with Center Weights (SVM-RCE-CW) is a feature selection algorithm that builds upon the SVM-RCE algorithm by introducing a new approach to feature cluster scoring based on a prediction model that encompasses all clusters and in which each cluster is represented by its mean.

In SVM-RCE-CW, the dataset is initially subjected to clustering, typically using K-means, to group features (e.g., genes) based on their expression patterns across samples. For each cluster, the algorithm calculates its center, which is a representative point that captures the average behavior of features within the cluster. A sub-dataset is then formed around these centers, highlighting the core characteristics of each cluster.

The core innovation of SVM-RCE-CW is the scoring mechanism for each cluster. Instead of relying on cross-validation accuracy from a classifier(Table 1 ,Step 3: Evaluation and Scoring) , SVM-RCE-CW trains a linear model on the sub-datasets centered around cluster centers. The weights assigned by this linear model to the features are interpreted as a measure of the cluster's importance. Clusters with lower weight magnitudes are considered less critical for the predictive modeling task and are eliminated in a recursive fashion.

This elimination process continues until a predetermined criterion is met, such as reaching a specific number of clusters.

SVM-RCE-CW mainly updates Step 3 of the original SVM-RCE algorithm, where the scoring mechanism for each cluster is based on the weights of a linear model trained on the cluster centers. The updated Step 3 for the SVM-RCE-CW algorithm, reflecting the new scoring mechanism is described in Table 2.

Table 2: The scoring step in SVM-RCE-CW

Step 3: Scoring (Updated)
(1) For each $cluster_i$ (i=1 to k):
    - Compute the center of $cluster_i$, the mean vector of all (genes) in the cluster.
(2) Create a data subset $D_{centers}$ that includes all the centers with the original class labels
(3) Fit a linear model (e.g., linear SVM or linear regression) on $D_{centers}$.
(4) The score assigned to each cluster is the absolute value of the estimated weight or coefficient of its center.

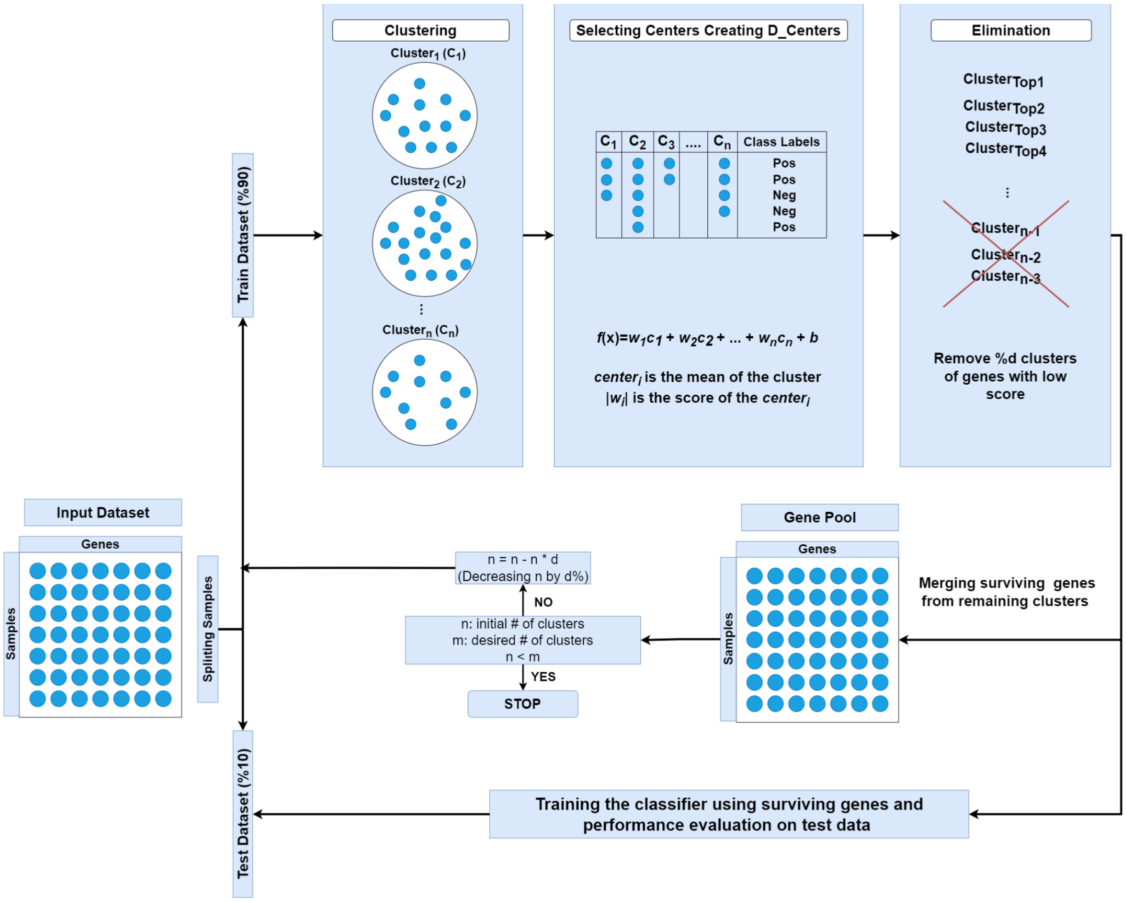
Clustering
$\text{Cluster}_1\ (C_1)$
$\text{Cluster}_2\ (C_2)$
$\text{Cluster}_n\ (C_n)$
Selecting Centers Creating D_Centers
C_1 C_2 C_3 .... C_n Class Labels
Pos
Pos
Neg
Neg
Pos
$f(x)=w_1c_1 + w_2c_2 + \dots + w_nc_n + b$
$center_i$ is the mean of the cluster
$|w_i|$ is the score of the $center_i$
Elimination
$\text{Cluster}_{\text{Top1}}$
$\text{Cluster}_{\text{Top2}}$
$\text{Cluster}_{\text{Top3}}$
$\text{Cluster}_{\text{Top4}}$
Cluster_{n-1}
Cluster_{n-2}
Cluster_{n-3}
Remove %d clusters of genes with low score
Train Dataset (%90)
Input Dataset
Genes
Samples
Splitting Samples
n = n - n * d (Decreasing n by d%)
NO
n: initial # of clusters
m: desired # of clusters
n < m
YES
STOP
Gene Pool
Genes
Samples
Merging surviving genes from remaining clusters
Training the classifier using surviving genes and performance evaluation on test data
Test Dataset (%10)

Fig. 1: The flowchart of the SVM-RCE-CW algorithm.

3. Results and Discussion

The present study utilized 17 gene expression datasets (GDS2547, GDS3268, GDS3646, GDS3837, GDS3874, GDS5037, GDS5499, GDS2519, GDS2609, GDS3875, GDS3929, GDS4228, GDS4824, GDS5093, GSE15573, GSE157103) downloaded from GEO [10] to test the new approach and compare it with the original algorithm. The performance was obtained as the average over 10 Monte Carlo Cross Validation (MCCV)[11] iterations. Samples were allocated to training and testing in a 9:1 ratio.

The evaluation criteria are accuracy, sensitivity, specificity, area under the curve (AUC) and the number of surviving genes in the last two stages. Each metric is averaged over the 17 datasets. The algorithm's successive rounds involve the removal of clusters with the lowest scores.

Both algorithms start with an initial value of k = 100, which is then sequentially decreased through the series: 90, 80, 70, 60, 50, 40, 30, 20, 10, 5, 2, 1.

Table 3: Evaluation of SVM-RCE and SVM-RCE-WC algorithms according to the number of surviving genes for 1 and 2 clusters based on the average of 17 datasets.

```
Step 3: Scoring (Updated)
(1) For each $cluster_i$ (i=1 to k):
        - Compute the center of $cluster_i$, the mean vector of all (genes) in the cluster.
(2) Create a data subset $D_{centers}$ that includes all the centers with the original class labels
(3) Fit a   linear model (e.g., linear SVM or linear regression) on $D_{centers}$.
(4) The score assigned to each cluster is the absolute value of the estimated weight or coefficient of its center.
```

Table 3 presents the results obtained by SVM-RCE and SVM-RCE-WC algorithms from the last two rounds, with 2 clusters and 1 cluster, averaged over the 17 datasets. The two algorithms have comparable performance metrics. The slight advantage in most metrics for 2 clusters over 1 cluster are expected as more predictors get used in the case of two clusters.

4. Conclusion

This study presented and compared two algorithms: Support Vector Machine-Recursive Cluster Elimination (SVM-RCE) and the novel SVM-RCE with Center Weights (SVM-RCE-CW), designed to enhance feature selection in gene expression data analysis. In the original SVM-RCE, scoring each cluster individually through internal cross-validation focuses on the predictive power of each cluster in isolation. This approach evaluates the effectiveness of a cluster based on its standalone predictive accuracy, assessed by repeatedly training and testing a model exclusively on data from the cluster. While this method can be computationally intensive due to multiple iterations of model training and validation for each cluster, it provides a straightforward measure of each cluster's contribution to the overall prediction accuracy.

Conversely, the updated method, SVM-RCE-WC, adopts an encompassing approach by training a single linear model on the centers of all clusters. This strategy not only reduces computational demands by avoiding the extensive cross-validation process but also adjusts each cluster for the other clusters. The scoring of clusters based on the weights or coefficients assigned to their centers in the linear model reflects not just the contribution of each cluster in isolation, but also how clusters relate to and interact with each other in the context of the model. This interaction can reveal synergies or redundancies among clusters, offering insights into the structure of the data and the underlying biological or logical processes.

References

[1] Y. Yang, P. Yin, Z. Luo, W. Gu, R. Chen, and Q. Wu, “Informative Feature Clustering and Selection for Gene Expression Data,” IEEE Access, vol. 7, pp. 169174–169184, 2019, doi: 10.1109/ACCESS.2019.2952548.

[2] S. Acharya, S. Saha, and Y. Thadisina, “Multiobjective Simulated Annealing-Based Clustering of Tissue Samples for Cancer Diagnosis,” IEEE J. Biomed. Health Inform., vol. 20, no. 2, pp. 691–698, Mar. 2016, doi: 10.1109/JBHI.2015.2404971.

[3] S. Tian, C. Wang, and H. H. Chang, “To select relevant features for longitudinal gene expression data by extending a pathway analysis method,” F1000Research, vol. 7, p. 1166, Jul. 2018, doi: 10.12688/f1000research.15357.1.

[4] X. Hou, J. Hou, and G. Huang, “Bi-dimensional principal gene feature selection from big gene expression data,” PLOS ONE, vol. 17, no. 12, p. e0278583, Dec. 2022, doi: 10.1371/journal.pone.0278583.

[5] M. Al-Rajab, J. Lu, and Q. Xu, “A framework model using multifilter feature selection to enhance colon cancer classification,” PLOS ONE, vol. 16, no. 4, p. e0249094, Apr. 2021, doi: 10.1371/journal.pone.0249094.

[6] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, “Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data,” BMC Bioinformatics, vol. 8, 2007, doi: 10.1186/1471-2105-8-144.

[7] M. Yousef, B. Bakir-Gungor, A. Jabeer, G. Goy, R. Qureshi, and L. C. Showe, “Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME,” F1000Research, vol. 9, p. 1255, Oct. 2020, doi: 10.12688/f1000research.26880.1.

[8] M. Yousef, M. Ketany, L. Manevitz, L. C. Showe, and M. K. Showe, “Classification and biomarker identification using gene network modules and support vector machines.,” BMC Bioinformatics, vol. 10, p. 337, 2009, doi: 10.1186/1471-2105-10-337.

[9] M. Yousef, A. Jabeer, and B. Bakir-Gungor, “SVM-RCE-R-OPT: Optimization of Scoring Function for SVM-RCE-R,” in Database and Expert Systems Applications - DEXA 2021 Workshops, vol. 1479, G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky, and S. Khan, Eds., in Communications in Computer and Information Science, vol. 1479. , Cham: Springer International Publishing, 2021, pp. 215–224. doi: 10.1007/978-3-030-87101-7_21.

[10] T. Barrett, S. Wilhite, P. Ledoux, C. Evangelista, I. Kim, M. Tomashevskyet, K. Marshall, K. Phillippy, P. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. Robertson, N. Serova, S. Davis and A. Soboleva, "ncbi geo: archive for functional genomics data sets—update", Nucleic Acids Research, vol. 41, no. D1, p. D991-D995, 2012. https://doi.org/10.1093/nar/gks1193

[11] Q.-S. Xu and Y.-Z. Liang, “Monte Carlo cross validation,” Chemom. Intell. Lab. Syst., vol. 56, no. 1, pp. 1–11, Apr. 2001, doi: 10.1016/S0169-7439(00)00122-2.