SOFTWARE TOOL ARTICLE

# REVISED Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME [version 2; peer review: 3 approved]

Malik Yousef[1], Burcu Bakir-Gungor[2], Amhar Jabeer[2], Gokhan Goy[2], Rehman Qureshi[3], Louise C. Showe[3]

[1]Zefat Academic College, Zefat, Israel
[2]Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey
[3]The Wistar Institute, Philadelphia, PA, USA

v2 **First published:** 19 Oct 2020, **9**:1255
https://doi.org/10.12688/f1000research.26880.1
**Latest published:** 05 Jan 2021, **9**:1255
https://doi.org/10.12688/f1000research.26880.2

## Abstract

In our earlier study, we proposed a novel feature selection approach, Recursive Cluster Elimination with Support Vector Machines (SVM-RCE) and implemented this approach in Matlab. Interest in this approach has grown over time and several researchers have incorporated SVM-RCE into their studies, resulting in a substantial number of scientific publications. This increased interest encouraged us to reconsider how feature selection, particularly in biological datasets, can benefit from considering the relationships of those genes in the selection process, this led to our development of SVM-RCE-R. SVM-RCE-R, further enhances the capabilities of SVM-RCE by the addition of a novel user specified ranking function. This ranking function enables the user to stipulate the weights of the accuracy, sensitivity, specificity, f-measure, area under the curve and the precision in the ranking function This flexibility allows the user to select for greater sensitivity or greater specificity as needed for a specific project.
The usefulness of SVM-RCE-R is further supported by development of the maTE tool which uses a similar approach to identify microRNA (miRNA) targets. We have also now implemented the SVM-RCE-R algorithm in Knime in order to make it easier to applyThe use of SVM-RCE-R in Knime is simple and intuitive and allows researchers to immediately begin their analysis without having to consult an information technology specialist. The input for the Knime implemented tool is an EXCEL file (or text or CSV) with a simple structure and the output is also an EXCEL file. The Knime version also incorporates new features not available in SVM-RCE.
The results show that the inclusion of the ranking function has a significant impact on the performance of SVM-RCE-R. Some of the clusters that achieve high scores for a specified ranking can also have high scores in other metrics.

## Open Peer Review

**Reviewer Status** ✓ ✓ ✓

| | Invited Reviewers | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| **version 2**<br>(revision)<br>05 Jan 2021 | ✓<br>report | ✓<br>report | ✓<br>report |
| | ↑ | ↑ | ↑ |
| **version 1**<br>19 Oct 2020 | ?<br>report | ?<br>report | ?<br>report |

1. **Marco Chierici**, Fondazione Bruno Kessler, Trento, Italy
2. **Osman Ugur Sezerman**, Acıbadem University, Istanbul, Turkey
3. **Abhishek Kumar**, Institute of Bioinformatics, International Technology Park, Bangalore, India

Any reports and responses or comments on the article can be found at the end of the article.

## Keywords

clustering, machine learning, recursive, gene expression, ranking, grouping, KNIME

**Corresponding author:** Malik Yousef (malik.yousef@gmail.com)

**Author roles: Yousef M**: Methodology, Project Administration, Supervision, Writing – Original Draft Preparation; **Bakir-Gungor B**: Formal Analysis, Supervision, Writing – Review & Editing; **Jabeer A**: Methodology, Software, Writing – Review & Editing; **Goy G**: Validation; **Qureshi R**: Writing – Review & Editing; **C. Showe L**: Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This study was supported by the National Cancer Institute [U01 CA200495-04; to L.C Showe]. The work of B.B.G. has been supported by the Abdullah Gul University Support Foundation (AGUV). The work of M.Y. has been supported by the Zefat Academic College.
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**How to cite this article:** Yousef M, Bakir-Gungor B, Jabeer A *et al.* **Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME [version 2; peer review: 3 approved]** F1000Research 2021, **9**:1255
https://doi.org/10.12688/f1000research.26880.2

**First published:** 19 Oct 2020, **9**:1255 https://doi.org/10.12688/f1000research.26880.1

## REVISED Amendments from Version 1

Made changes to the abstract so that the ranking algorithm is brought to focus.

Updated introduction, it now mentions ranking algorithm.

Corrected typos in the pseudo-code and made it clear.

Change a sub-heading in the methods section to emphasize the ranking function.

Added information about the input tables used in the workflow in the data section.

Changed the legend title in Figure 1.

Results now show which datasets were used to represent figures.

Table 3 values have changed to reflect consistency in the manuscript.

Added a new subheading in the results section to emphasize the comparison of different algorithms (SVM-RCE and SVM-RCE-R).

Figure 5 has f measures added to it.

Typos and spellings have been corrected throughout the manuscript and method of mentioning weights (w1,w2...,w3) have been updated to maintain consistency.

Finally, updated DOI in the manuscript to reflect the input tables https://doi.org/10.5281/zenodo.4327346.

**Any further responses from the reviewers can be found at the end of the article**

## Introduction

The application of a variety of new technologies for measuring gene expression has generated publicly available datasets with very high feature dimensionalities (tens of thousands of genes)[1,2]. Because expression of certain groups of genes can be functionally related, they can be grouped according to a specific metric, which can be defined by the biological processes and interactions the group represents. Since most of the existing feature selection approaches have been borrowed from the field of computer science and statistics, they fail to consider the associations between gene expression features. We now propose to address that issue. In our initial study we suggested an algorithm called SVM-RCE[3], where genes were grouped using a k-means based clustering algorithm. Our following study, SVM-RNE[4] incorporated the possibility of grouping subsets of genes according to gene sub-networks. Our recent tool maTE[5] suggested an alternative grouping based on microRNA targets and replaced k-means clustering with ensemble clustering[6].

Sahu and Mishra[7] have stressed the weakness of Signal-to-Noise Ratio (SNR) and t-statistics, which are widely used for gene rankings in the analysis of gene expression data, as using SNR and t-statistics as filtering techniques will likely select redundant features. They instead suggest that the genes are first grouped into clusters based on the similarities of their expression values, followed by the application of different filtering techniques to rank the genes in each cluster. The assigned ranks were then used to select the most informative genes from each cluster resulting in improved classification. The problem of dealing with clusters of features or groups of correlated features, in remote sensing data sets, was also recently addressed by Harris and Niekerk[8]. They stress the importance of first clustering the features by affinity propagation, and then applying a ranking function to overcome the weakness of the traditional feature selection approaches, which are likely to result in the selection of sub-optimal features. Therefore, the ranking function we propose is founded upon a process of assigning weights to the various clusters based on their performance metrics. This allows the user to specify which metric they want to focus on depending on their needs. The implementation of the ranking function is explained in further detail in the ranking function section.

## Methods

### The SVM-RCE workflow

The SVM-RCE algorithm can be described by three main steps:

1. The Clustering step combines the genes, based on expression, into groups using a clustering algorithm such as K-means. The merit of this step is to put genes with similar expression patterns into one cluster in order to deal with them together. In general, we refer to this step as a grouping function.
2. The Rank step ranks each cluster using a function we have used in the SVM-RCE[3] using Rank(X(S), f, r) as the average accuracy of the linear SVM over the data X represented by the S genes computed as f-folds cross validation repeated r times. We set *f* to 3 and *r* to 5 as default values (See Pseudocode 1).
3. The RCE step removes the lower ranked clusters of genes and can be implemented to remove one cluster or a percentage of clusters as specified by the researcher, e.g. removing the lower 10% of the clusters.

We have applied the step of recursive cluster elimination based on the hypothesis that the clustering algorithm will generate new sets of clusters and that some of the genes will move between clusters and we have shown this to be the case.

**Pseudocode 1. The Ranking method R(), a main component of the SVM-RCE**

```
Ranking Algorithm - $R(X_s,M,f,r)$
  $X_s$: any subset of the input gene expression data X, the
  features are gene expression values
  M $\{m_1,m_2,\dots,m_p\}$ is a list of groups produced by k-means.
  $f$ is a scalar ($0 \le f \le 1$): split into train and test data
  r: repeated times (iteration)
  res={} for aggregating the scores for each $m_i$
Generate Rank for each $m_i$, $Rank(m_i)$:
  For each $m_i$ in M
  $sm_i=0$;
     Perform $r$ times (here r=5) steps 1–5:
     1.  Perform stratified random sampling to split $X_s$ into train
         $X_t$ and test $X_v$ data sets according to $f$ (here 80:20)
     2.  Remove all genes (features) from $X_t$ and $X_v$ which are
         not in the group $m_i$
     3.  Train classifier on $X_t$ using SVM
     4.  $t$ = Test classifier on $X_v$ –calculate performance of test data
     5.  $sm_i = sm_i + t$;
  $Score(m_i) = sm_i / r$ ; Aggregate performance
  $res = \bigcup_{i=1}^{p} Score(m_i)$
Output
Return res ( $res = \{Rank(m_1),Rank(m_2),\dots,Rank(m_p)\}$ )
```

## Incorporation of novel ranking function

The algorithm of Recursive Cluster Elimination[3] considers clusters of similar features/genes and applies a rank function to each group as described in Pseudocode 1. Since we are using the clustering algorithm k-means we refer to these groups as clusters, but it could be any other biological or more general function that groups the particular features, such as KEGG pathways or microRNA targets, as we have suggested in several other studies[4,5]. As illustrated in Pseudocode 1, in the original code of SVM-RCE we used the accuracy which was the performance as the determinant for ranking the clusters. The data for establishing that ranking was divided between training and testing. The data represented by each gene/feature is then assigned to a specific cluster and the rank function is then applied as the mean of *r* repeat times of the repeated training and the testing performance while recording different measurements of accuracy (sensitivity, specificity, etc.).

In this new version implemented in Knime[9] we have incorporated more user specific ranking function. The user provides the weights of the following ranking function that correspond to the mean of each measurement achieved by the *r* times of the internal:

$$R(w_1,w_2,w_3,w_4,w_5,w_6)= w_1 \times acc + w_2 \times sen + w_3 \times spe + w_4 \times fm + w_5 \times auc + w_6 \times prec$$

Where the *acc* is the accuracy, *sen* is the sensitivity, *spe* is the specificity, *fm* is the f-measurement, *auc* is the area under the curve and *prec* is precision.

The coefficient weights represent the importance of each measurement for searching those clusters of genes that contribute to the final performance requirements. For example, if the user is interested in achieving greater specificity than sensitivity, the user would choose weights of 0.7 for the parameter *spe* and 0.3 for *sen*, stating that he is searching for clusters of genes that contribute to high specificity. However, one can also choose all the weights to be zero, with the weight of accuracy is set as 1, the rank function will then only rely on the accuracy.

## Implementation in Knime

We have used the free and open-source platform Knime[10] for re-coding SVM-RCE (Figure 1–Figure 3) due to its simplicity and useful graphical presentations. Knime is a highly integrative tool that allows the user to include other programming languages such as R, Python and Java. In addition, one can also add external packages as such WEKA, H2O and so on. Figure 1 presents the workflow that includes SVM-RCE-R as a meta-node. The workflow can be executed on multiple input files. The node "List Files" will be indicated on the folder that has the input files. The workflow loops through those files and runs the SVM-RCE-R meta-node. The "Loop End" is also collecting specific results that can be subjected to further analysis.

The SVM-RCE-R meta-node consists of two components (two meta-nodes). The meta-node "Genes Filter t-test" (Figure 1b) is used to reduce the dimension of the features by applying the t-test to the training part of the data. Following that is the RCE component.

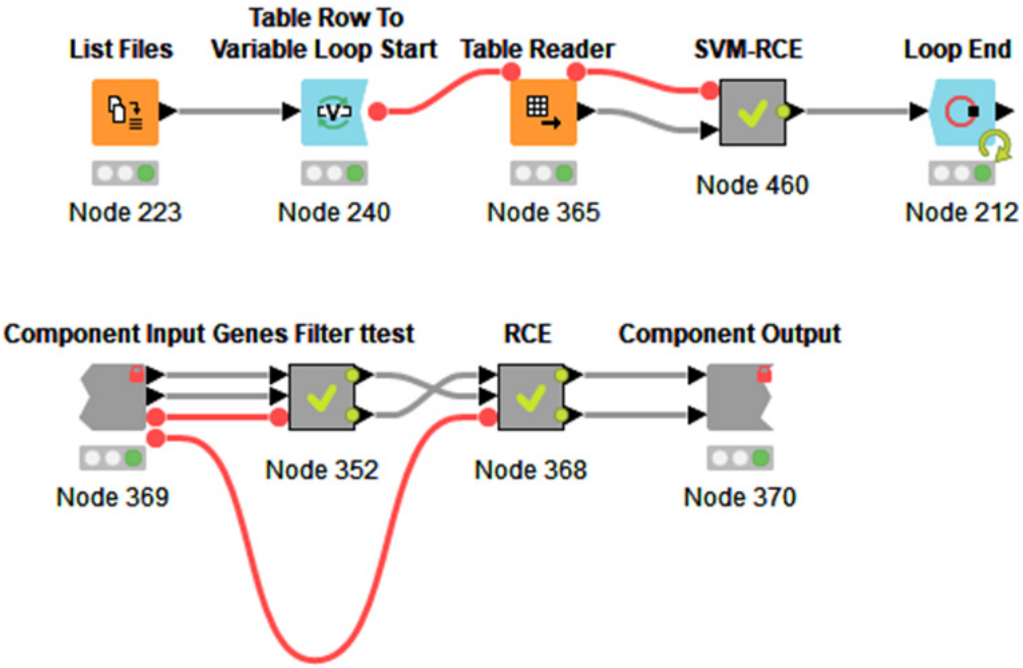

**Figure 1.** (**a**) The main Knime workflow for SVM-RCE-R that can be executed on multiple input files. (**b**) The internal meta-node SVM-RCE-R that consists of two components.

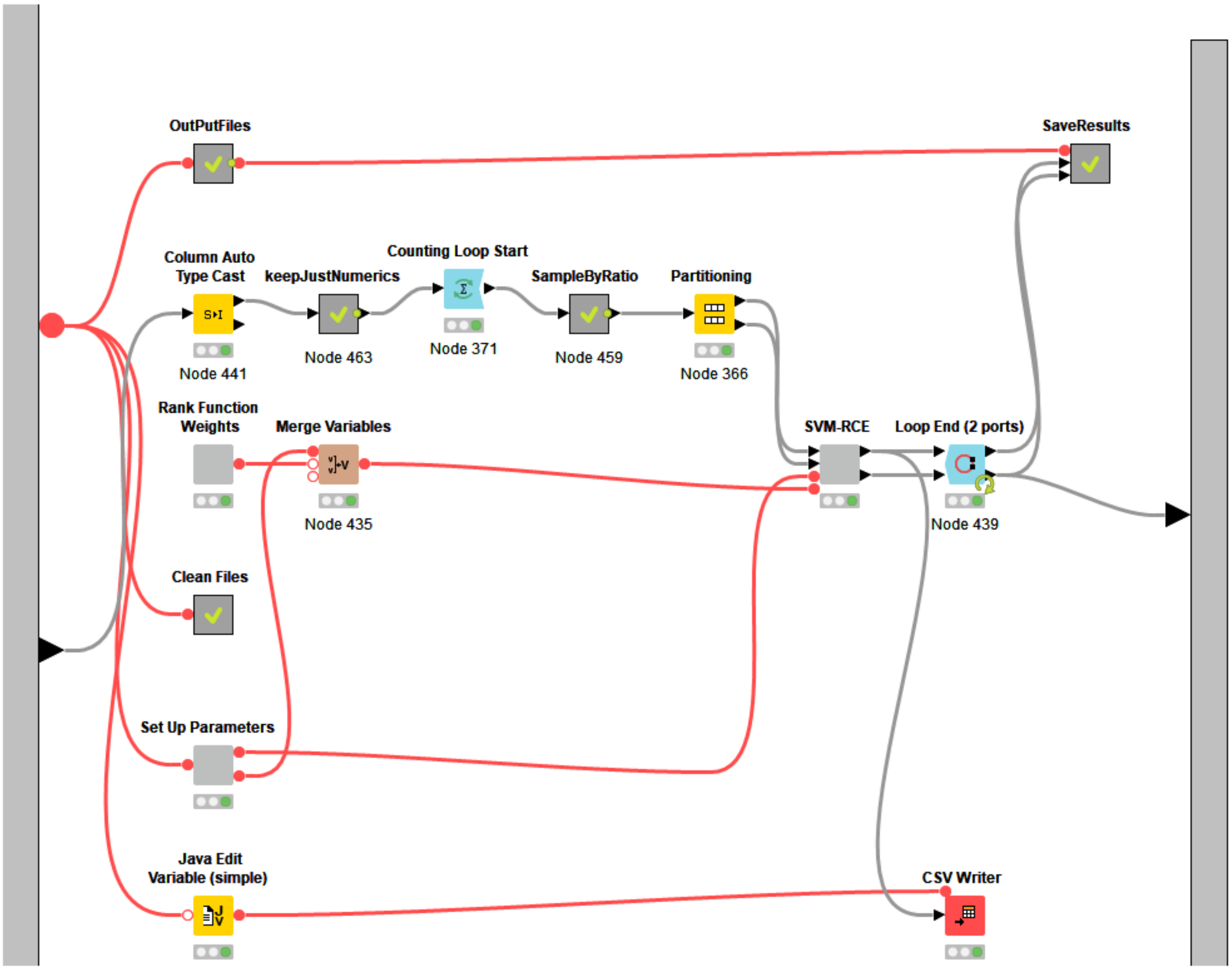

**Figure 2. The interface of the SVM-RCE-R.**

The interface of the SVM-RCE-R is presented in Figure 2. This part of the tool is used to set different parameters. The user can specify the number of iterations for Monte Carlo cross-validation (MCCV) by configuring the node “Counting Loop Start”. MCCV is the process of randomly selecting (without replacement) some fraction of the data to form the training set, and then assigning the rest to the test set. The node “Partitioning” is used to specify the ratio of the training/testing splitting.

The most important component “Rank Function Weights” is related to the rank function R(), where the user specifies the values of the weights w1, w2, .., w6. We show in the results section that these values have an impact on the performance of the SVM-RCE-R.

Figure 3, meanwhile, shows nodes present in the meta-node SVM-RCE. It is designed so that it follows the pseudocode, thereby making it user-friendly.

## Operation

The workflow was developed in KNIME which is compatible with Mac, Linux and Windows OS. We would recommend using a quad core CPU with at least 8 GB of RAM to run the workflow. Moreover, users will need to install Python 3 and R environments, Anaconda is recommended for the installation of Python 3 meanwhile R > 1.5 should be installed with Rserve package which can be found at https://cran.r-project.org/web/packages/Rserve/index.html.

## Gene expression data

12 human gene expression datasets were downloaded from the Gene Expression Omnibus at NCBI[11]. For all datasets, disease (positive) and control (negative) data were available (Table 1). All of the datasets are gene expression data with different number of samples and were used as is in our workflow. The columns of the datasets indicate the sample identification code and the rows contain the names of the genes. Moreover, the sample input datasets can be found in the underlying data

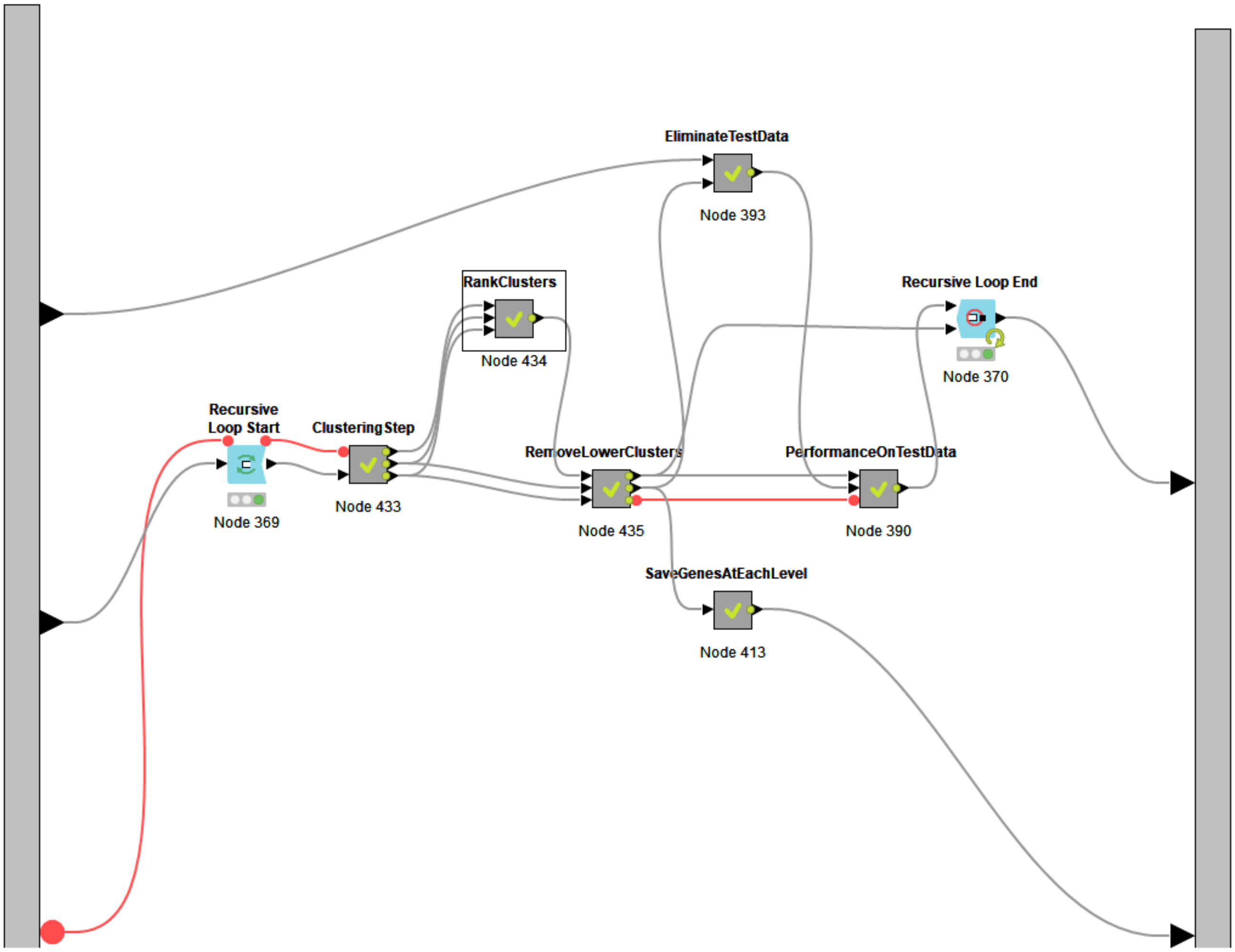

**Figure 3. The nodes present in the meta-node SVM-RCE-R.**

repository. Those 12 datasets served to test the SVM-RCE-R tool and to compare its performance with two other approaches; the filter and embedded approaches[12,13]. The first approach performs feature selection using information gain (SVM-IG) on the training part while the second approach is compared with SVM with recursive feature elimination (SVM-RFE)[14]. We have also implemented a workflow for SVM-RFE that is based on the Scikit-learn package[15] in Knime.

## Results

We have tested SVM-RCE-R on the aforementioned datasets and used the performance results to verify our new ranking function. For the comparison of the three approaches, we have considered five datasets (GDS1962, GDS3646, GDS3874, GDS3900, GDS5499) as listed in Table 2. We have applied SVM-RCE-R, obtaining the performance over 100 iterations. At each iteration we have split the data into 90% for training and 10% for testing. The average of all different performance measurements is then aggregated. For additional comparison we refer to the first study published about SVM-RCE-R[3].

The results indicate that SVM-RCE-R outperforms or is equivalent to the other approaches in all the datasets except in determining the specificity for GDS3646 with a case to control ratio of 5 to 1[16] and GDS3874.

We have also considered different values of the rank function *R(w1,w2,w3,w4,w5,w6)* by specifying different values of the measurements weights, w1,..,w6 and have generated six rank functions as listed in Table 3. For each rank function we have applied the SVM-RCE-R obtaining the performance over 100 iterations. At each iteration we have split the data into 90% for training and 10% for testing. The average

**Table 1. Description of the 12 data sets used in our study.** The data sets are obtained from GEO. Each entry has the GEO code the name of the data, the number of samples and the classes of the data.

| GEO Accession | Title | Sample count | Classes |
|---|---|---|---|
| GDS1962 | Glioma-derived stem cell factor effect on angiogenesis in the brain | 180<br>pos=157<br>neg=23 | non-tumor=23 (neg)<br>astrocytomas=26 (pos)<br>glioblastomas=131 (pos) |
| GDS2519 | Early-stage Parkinson's disease: whole blood | 105<br>pos=50<br>neg=55 | healthy control=22 (neg)<br>neurodegenerative disease control=33 (neg)<br>Parkinson disease=50 (pos) |
| GDS3268 | Colon epithelial biopsies of ulcerative colitis patients | 202<br>pos=73<br>neg=129 | normal=73<br>ulcerative colitis=129 |
| GDS2547 | Metastatic prostate cancer (HG-U95C) | 164<br>pos=75<br>neg=89 | normal=75<br>tumor=89 |
| GDS5499 | Pulmonary hypertensions: PBMCs | 140<br>pos=99<br>neg=41 | control=41 (neg)<br>idiopathic pulmonary arterial hypertension=30 (pos)<br>scleroderma-associated pulm. arterial hypert=42 (pos)<br>systemic sclerosis (SSc) without pulm. hypert=19 (pos)<br>SSc, interstitial lung disease & pulm. hypert=8 (pos) |
| GDS3646 | Celiac disease: primary leukocytes | 132<br>pos=110<br>neg=22 | healthy control=22<br>celiac disease=110 |
| GDS3874 | Diabetic children: peripheral blood mononuclear cells (U133A) | 117<br>pos=93<br>neg=24 | healthy=24<br>type 1,2 diabetes=93 |
| GDS3837 | Non-small cell lung carcinoma in female nonsmokers | 120<br>pos=60<br>neg=60 | lung cancer=60<br>control=60 |
| GDS5037 | Severe asthma: bronchial epithelial cell | 108<br>pos=88<br>neg=20 | mild asthma=50<br>control=20<br>severe asthma=38 |
| GDS4516_ 4718 | Colorectal cancer: laser microdissected tumor tissues<br>Colorectal cancer: homogenized tumor tissues | 148<br>pos=104<br>neg=44 | laser microdissected tumor tissues=104<br>homogenized tumor tissues=44 |
| GDS3900 | Fear conditioning effect on the hybrid mouse diversity panel: hippocampus and striatum | 198<br>pos=100<br>neg=98 | hippocampus=100 (pos)<br>striatum=98 (neg) |
| GDS3929 | Tobacco smoke effect on maternal and fetal cells | 183<br>pos=128<br>neg=55 | non-smoker=128 (pos)<br>smoker=55 (neg) |

**Table 2.** Comparison results between SVM-RCE-R, SVM-RFE and SVM-IG including the standard deviation.

| | Information | | SVM-RCE-R | | | | SVM-RFE | | | | SVM-IG | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | #samples | #genes | Sensitivity | Specificity | Accuracy | #Genes | Sensitivity | Specificity | Accuracy | #Genes | Sensitivity | Specificity | Accuracy | #Genes |
| **GDS1962** | 180 | 54613 | 0.94 ± 0.09 | 1.00 ± 0.00 | 0.96 ± 0.07 | 97.00 ± 39.11 | 0.93 ± 0.13 | 0.97 ± 0.06 | 0.94 ± 0.08 | 100 | 0.90 ± 0.12 | 0.97 ± 0.11 | 0.93 ± 0.08 | 100 |
| **GDS3646** | 132 | 22185 | 0.80 ± 0.19 | 0.50 ± 0.33 | 0.71 ± 0.19 | 108.70 ± 45.91 | 0.74 ± 0.32 | 0.64 ± 0.21 | 0.67 ± 0.17 | 100 | 0.63 ± 0.24 | 0.67 ± 0.25 | 0.65 ± 0.18 | 100 |
| **GDS3874** | 117 | 22283 | 0.82 ± 0.18 | 0.53 ± 0.17 | 0.71 ± 0.13 | 112.80 ± 38.37 | 0.84 ± 0.16 | 0.65 ± 0.24 | 0.70 ± 0.16 | 100 | 0.79 ± 0.46 | 0.79 ± 0.10 | 0.79 ± 0.24 | 100 |
| **GDS3900** | 198 | 25697 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 141.20 ± 67.81 | 1.00 ± 0.00 | 1.00 ± 0.00 | 1.00 ± 0.00 | 100 | 0.99 ± 0.02 | 1.00 ± 0.01 | 0.99 ± 0.01 | 100 |
| **GDS5499** | 140 | 49577 | 0.93 ± 0.08 | 0.90 ± 0.13 | 0.92 ± 0.07 | 145.00 ± 44.54 | 0.93 ± 0.07 | 0.90 ± 0.10 | 0.88 ± 0.08 | 100 | 0.90 ± 0.11 | 0.83 ± 0.14 | 0.87 ± 0.08 | 100 |

of all different performance measurements is then aggregated. All of the datasets that are used for the comparison between the performance of six different functions are listed in Table 3 and the results are shown in Figure 4.

Figure 4 shows that there is deviation of the performance measurements for each R. However, we observed that the deviation is clear if we consider each data set individually which will be discussed in further detail in the next section.

## SVM-RCE vs SVM-RCE-R

In order to examine the effect of the Rank function, we plotted the results obtained on the cluster level 2 as appears in Figure 5 (See *Underlying data* for all the results for the 12 datasets[16]) for each data set.

**Table 3. Different values of weights generating different rank functions.** w1,weight of accuracy; w2, weight of sensitivity; w3, weight of specificity; w4, weight of f-measurement; w5, weight of area under the curve.

| | |
|---|---|
| R1 | w1=0.2, w3=0.3, w2= 0.4, w5=0.1 |
| R2 | w1=1.0, the rest are zero |
| R3 | w5= 1.0, the rest are zero |
| R4 | w4=1.0, the rest are zero |
| R5 | w3=0.2, w2=0.8 |
| R6 | w3=0.8, w2=0.2 |

For example, the accuracy obtained with R5 is significantly greater than R4 by about 12%, while reaching 4%–6% more than the other ranks. Interestingly we are getting a 4% improvement over the standard rank we have been using with the old version of SVM-RCE, which was R2.

GDS2547 data reached an accuracy of ~79% applying R6 and 63% with R3, a difference of 16%, which is about 9% over the standard rank using the previous version SVM-RCE. However, for GDS5037 the max performance obtained with the standard rank R2 reached a difference of 16% over the minimum values reached by R5.

We have calculated the overall difference between the max value of each rank and the R2 that was used in the old version to get 5%.

This indicates that one can dramatically improve the performance of SVM-RCE-R by searching for the optimal values of the weights of the rank function.

We also conducted an additional experiment in order to examine the effect of gradually changing the values of sensitivity and specificity weights in the rank function. We ran two experiments on GDS3646 and GDS1962 data considering the values of (1,0) (0,1) (first argument is sensitivity weight while second one is specificity weight) increasing by 0.1 to reach (0,1) for the weights of sensitivity and specificity, respectively. The results are represented in Figure 6 for cluster level 2.

Figure 6 shows that the two graphs are behaving differently over the set of weights, showing that the results depend on the

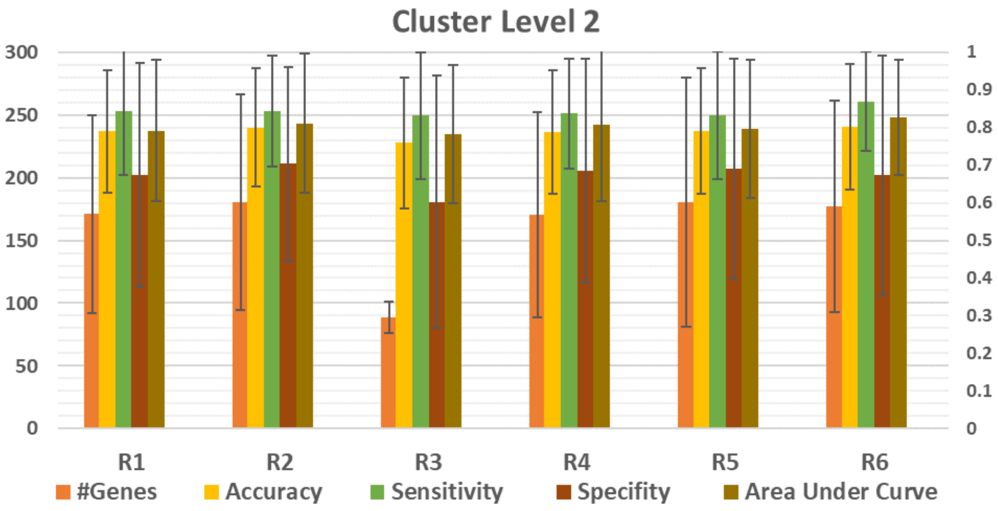

**Figure 4. Comparison between the performance of six (R1,..,R6) different functions, listed in Table 3.** The average of 100 iterations if computed for different performance measurements for each R1,...,R6 over the 12 datasets. The results of the level of cluster 2 is presented. #Genes is the average number of genes in level 2. The average accuracy, sensitivity, specificity and area under the curve (AUC) is presented for R1,..R6.

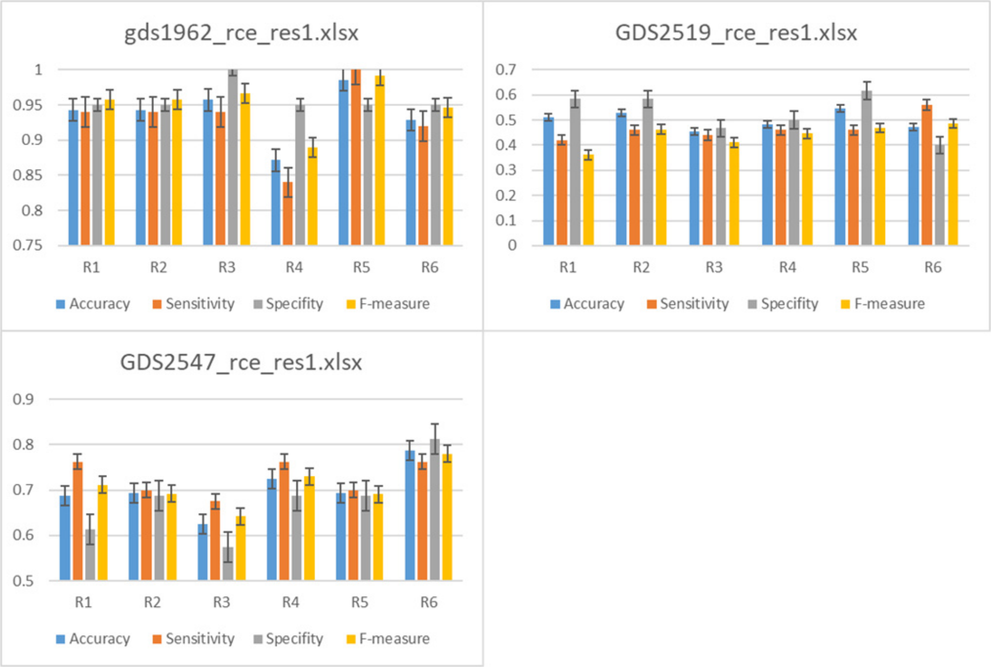

**Figure 5. The performance of SVM-RCE-R over different data considering different Ranks functions(R1..R6) where R2 is SVM-RCE.** *The results show the increase/decrease of those rank functions on the accuracy, sensitivity and specificity.*

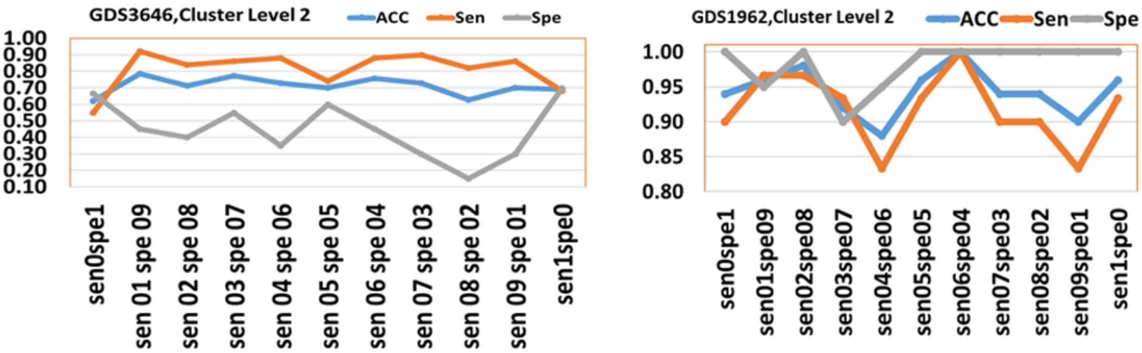

**Figure 6. The performance of values of weights for Sen and Spe over (0,0) gradually increased by 0.1 to (1,1).** The axes labels are the values, for example sen01spe09 is associated for weight of 0.1 of sensitivity and 0.9 for specificity. The accuracy (ACC), sensitivity (Sen) and specificity (Spe) are plotted.

specific data. Interestingly we see that for GDS1962 data, the optimal performance for all measurements is with weight 0.6 and 0.4 for sensitivity and specificity, respectively. Although the maximum accuracy is achieved over (0.1,0.9) weights pair, for GDS3646 data, the specificity at this point is very low and not usable for prediction, while (0.5,0.5) seems to provide reasonable performance for both sensitivity and specificity. Additionally, we have computed the number of common genes by considering the top 50 significant genes for each pair (sen01sep09 vs sen02spe08, …) having on average 11 genes. That is another indication that the rank function also has a significant impact on the list of the significant genes.

## Discussion

As gene expression data sets become more complex, new computational tools that deal with features in a non-traditional way are needed to address this complexity. Our approach does not simply tackle the problem of inherent redundant or correlated features, it also suggests that defining the grouping metrics is equally important when searching that specific feature space that each researcher would like to focus on. Different biological systems/problems can require an output with a greater emphasis on either specificity, sensitivity or overall accuracy. Although specifying a certain metric, for instance, specificity, has higher priority during clustering, there can be cases where the clusters have high values for other metrics, which can be inferred from our results. Therefore, finding the optimal ranking will be one of the topics that we will further focus on. We now provide the capability to decide whether the specific problem being addressed will benefit more from reducing false positives or false negatives.

This new version of RCE now provides the user with the ability to control the analyses and to also design the ranking function that will allow exploration of the data in a way that addresses the specific goals of the analysis. Additionally, since it is easy to change the learning algorithm from SVM or to combine SVM with other machine learning algorithms, it further expands the utility of RCE-R. These additional components will be added to the next version of RCE as well as additional features for optimization procedures. Currently, our program estimates each cluster separately; a future version will combine different numbers of clusters using a search algorithm in order to identify the optimal combination that will return the highest accuracy.

## Data availability

### Source data

Human gene expression datasets from Gene Expression Omnibus, Accession numbers: GDS1962, GDS2519, GDS3268, GDS2547, GDS5499, GDS3646, GDS3874, GDS3837, GDS5037, GDS4516_GDS4718, GDS3900, GDS3929

### Underlying data

Zenodo: Ajabeer/SVM-RCE-R-results-Omnibus-dataset: Supplementary Data for SVM-RCE-R. https://doi.org/10.5281/zenodo.4327346[16].

This project contains the following underlying data:

- all_res1_clusters.xlsx files (contains the summary of all res_1 files for all 12 datasets for R1-R6)
- logResults.csv files (contains the scoring values and class labels for each run of the SVM-RCE loop for each of the 12 datasets, R1-R6)
- rankedGenes.xlsx files (contains the names of the genes that ranked according to the rank function with their levels, rank function values and scores for each of the 12 datasets, R1-R6)
- res1.xlsx files (contains the mean values of genes and the scoring metrics values calculated: Accuracy, Sensitivity, Specificity, F-measure, AUC, Cohens Kappa, for each cluster level for each of the 12 datasets, R1-R6)
- res2.xlsx files (contains the number of genes for each level, scoring metrics values calculated: Accuracy, Sensitivity, Specificity, F-measure, AUC, Cohens Kappa, for each cluster for each iteration for each of the 12 datasets, R1-R6)
- table files (contains the datasets from GEO expression datasets used as the input file for the workflow)

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

## Software availability

The SVM-RCE-R Knime workflow, step-by-step tutorial and a detailed documentation are available on the following web site: https://malikyousef.com/svm-rce-in-knime/

Source code available from: https://github.com/malikyousef/SVM-RCE-R-KNIME

Archived source code at time of publication: https://zenodo.org/record/4066639#.X3sQVlLis2w[9]

License: GNU General Public License v3.0

Detailed terms and conditions of KNIME can be found at https://www.knime.com/downloads/full-license.

## References

1. Clough E, Barrett T: **The Gene Expression Omnibus Database.** *Methods Mol Biol.* 2016; **1418**: 93–110.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Brazma A, Sarkans U, Shojatalab M, *et al.*: **ArrayExpress - A public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res.* 2003; **33**(Database issue): D553–5.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Yousef M, Jung S, Showe LC, *et al.*: **Recursive Cluster Elimination (RCE) for classification and feature selection from gene expression data.** *BMC Bioinformatics.* 2007; **8**: 144.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Yousef M, Ketany M, Manevitz L, *et al.*: **Classification and biomarker identification using gene network modules and support vector machines.** *BMC Bioinformatics.* 2009; **10**: 337.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Yousef M, Abdallah L, Allmer J: **maTE: discovering expressed interactions between microRNAs and their targets.** *Bioinformatics.* 2019; **35**(20): 4020–4028.
PubMed Abstract | Publisher Full Text

6. AbdAllah L, Khalifa W, Showe LC, *et al.*: **Selection of Significant Clusters of Genes based on Ensemble Clustering and Recursive Cluster Elimination (RCE).** *J Proteomics Bioinform.* 2017; **10**(8): 186–192.
Publisher Full Text

7. Sahu B, Mishra D: **A novel approach for selecting informative genes from gene expression data using Signal-to-Noise Ratio and t-statistics.** In, *2011 2nd International Conference on Computer and Communication Technology, ICCCT-2011.* 2011.
Publisher Full Text

8. Harris D, VanNiekerk A: **Feature clustering and ranking for selecting stable features from high dimensional remotely sensed data.** *Int J Remote Sens.*

2018; **39**(23): 8934–8949.
Publisher Full Text

9. malikyousef: **malikyousef/SVM-RCE-R-KNIME: SVM-RCE-R (Version v1.0).** Zenodo. 2020.
http://www.doi.org/10.5281/zenodo.4327346

10. Berthold MR, Cebron N, Dill F, *et al.*: **KNIME: The Konstanz Information Miner.** In *SIGKDD, Explorations.* 2008; 319–326.
Publisher Full Text

11. Barrett T, Wilhite SE, Ledoux P, *et al.*: **NCBI GEO: Archive for functional genomics data sets - Update.** *Nucleic Acids Res.* 2013; **41**(Database issue): D991–5.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Pan W: **A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.** *Bioinformatics.* 2002; **18**(4): 546–554.
PubMed Abstract | Publisher Full Text

13. Lazar C, Taminau J, Meganck S, *et al.*: **A survey on filter techniques for feature selection in gene expression microarray analysis.** *IEEE/ACM Trans Comput Biol Bioinform.* 2012; **9**(4): 1106–19.
PubMed Abstract | Publisher Full Text

14. Guyon I, Weston J, Barnhill S, *et al.*: **Gene Selection for Cancer Classification using Support Vector Machines.** *Machine Learning.* 2012; **46**: 389–422.
Publisher Full Text

15. Pedregosa F, Varoquaux G, Gramfort A, *et al.*: **Scikit-learn: Machine learning in Python.** *J Mach Learn Res.* 2011; **12**(85): 2825−2830.
Reference Source

16. Ajabeer: **Ajabeer/SVM-RCE-R-results-Omnibus-dataset: Supplementary Data for SVM-RCE-R (Version v1.0.0).** Zenodo. 2020.
http://www.doi.org/10.5281/zenodo.4327346

# Open Peer Review

## Current Peer Review Status: ✓ ✓ ✓

**Version 2**

Reviewer Report 22 January 2021

https://doi.org/10.5256/f1000research.31433.r76608

✓ **Marco Chierici**
Fondazione Bruno Kessler, Trento, Italy

I am satisfied with the revised version.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Bioinformatics, artificial intelligence, computational biology

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 18 January 2021

https://doi.org/10.5256/f1000research.31433.r76606

✓ **Abhishek Kumar**
Institute of Bioinformatics, International Technology Park, Bangalore, Karnataka, India

Accepted.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Multiomics, genomics and ML

**I confirm that I have read this submission and believe that I have an appropriate level of**

**expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 11 January 2021

https://doi.org/10.5256/f1000research.31433.r76607

✓

**Osman Ugur Sezerman**
Department of Biostatistics and Bioinformatics, Life Science Institute, Acıbadem University, Istanbul, Turkey

I am satisfied with the changes.

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Bioinformatics

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

## Version 1

Reviewer Report 14 December 2020

https://doi.org/10.5256/f1000research.29685.r73698

?

**Abhishek Kumar**
Institute of Bioinformatics, International Technology Park, Bangalore, Karnataka, India

Yousef *et al.* have presented SVM-RCE-R as the improved implementation of the original algorithm SVM-RCE in Knime with additional features. The most important new feature is the ranking function R(). The authors demonstrate R() is important in improving the performance over the original study SVM-RCE. The Knime implementation demonstrates the novelty of the integration of biological information into the machine learning feature selection component. The approach is performing grouping by clustering and ranking by internal cross-validation. The approach is interesting and would also contribute to the new trend of integrative biological knowledge into the process of feature sections.

Major issues:

I have two major concerns:

1. The authors must define the "user-specific ranking function" more clearly at the start of the manuscript.

2. Usage of collected datasets are not clear and authors must provide better examples of results derived from datasets.

Minor issues:

The authors must improve carefully figure and table legends plus typos and errors:

1. Table 3 title: 'ACC=accuracy', 'Spe=specificity', and so on.

2. Figure 4: "The average of 100 iterations if computed” should be is 'computed'.

3. Results section: "For the comparison of the three approaches, five datasets are considered for," -> the last 'for' should be removed.

4. Figure 3: It should be 'SVM-RCE-R' and not "SVM-RCE".

5. Figure 1: It should be 'SVM-RCE-R' and not "SVM-RCE".

**Is the rationale for developing the new software tool clearly explained?**

Yes

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Multiomics, genomics and ML

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 15 Dec 2020

**Malik Yousef**, Zefat Academic College, Zefat, Israel

Dear Dr. Abhishek Kumar,

Thank you very much for your feedback suggestions, which significantly improve the manuscript and enrich the content. We would like to address your comments:

- ○ *The authors must define the "user-specific ranking function" more clearly at the start of the manuscript.*

We have now made changes based on the earlier reviewers as well to clarify and focus the manuscript more on the user-specific ranking function. We have made the relevant changes in the introduction, abstract as well as defining a more focused heading on the novel ranking feature.

- ○ *Usage of collected datasets are not clear and authors must provide better examples of results derived from datasets.*

We noticed that we were not clear in describing which datasets were used for which figures, therefore now we have mentioned in detail which datasets are used in their relevant tables and figures.

- ○ *Figure 4: "The average of 100 iterations if computed" should be is 'computed'.*

We wrote it down as "if" since the user has the option of computing the algorithm to his/her needs and it can differ to their choices.

We have also gone through all the typos and spelling mistakes that you have so kindly provided us, thank you very much. Please let us know if you further feedback.

Yours truly,
The authors.

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 02 December 2020

https://doi.org/10.5256/f1000research.29685.r73308

?

**Osman Ugur Sezerman**
Department of Biostatistics and Bioinformatics, Life Science Institute, Acıbadem University, Istanbul, Turkey

In this manuscript, the authors describe the implementation of feature selection method, SVM-RCE-R, in KNIME and demonstrate the usefulness of this tool. As they clearly describe, this is an improvement of the previously developed SVM-RCE. The most important novel feature in SVM-RCE-R is the user-specific ranking function, allowing the researcher to select for different performance metrics. As KNIME provides an easy-to-use interface, and the tool requires simple input formats, it will most likely be a valuable tool for biomedical researchers with many different backgrounds.

**Major issues:**

- The introduction of the user-specific ranking function could be emphasized more clearly, perhaps in the Introduction section.

- The analyzed datasets presented in the results section should be more clearly defined for each result, the number of datasets are confusing.

- The abstract states that “The input for the Knime tool is an EXCEL file (or text or CSV) with a simple structure...”, this structure should be described in the main text.

**Minor issues:**

- In Table 3, the metric names should be replaced by the weights of these metrics.

- Minor typos:

  - Page 3, last paragraph: it should be “user-specific”, “ranking functions” should be singular (i.e., “ranking function”).
  - Page 4, in the user-specific ranking function, “pres” should be “prec”.
  - Page 5, “Reserve” should be “Rserve”.
  - Please use "meta-node" throughout the manuscript.
  - “are under curve” should be “area under the curve”.

**Is the rationale for developing the new software tool clearly explained?**

Yes

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Partly

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Bioinformatics

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 13 Dec 2020

**Malik Yousef**, Zefat Academic College, Zefat, Israel

Dear Prof. Ugur Sezerman,
We deeply appreciate your taking the time to provide your valuable feedback to us. We have revised the manuscript based on that feedback and would also like to address the comments:

- *The introduction of the user-specific ranking function could be emphasized more clearly, perhaps in the Introduction section.*

We have now included a description of the novelty of our ranking function in the abstract to better focus on this topic. Moreover, we have also updated the introduction section with your recommendations. Finally, we have revised the methods section so that it clearly states and describes the user specific ranking function.

- *The analyzed datasets presented in the results section should be more clearly defined for each result, the number of datasets are confusing.*

We have updated the results section and now clearly mention which datasets were used in each graph or table. We have also included how many datasets were used for comparison results and have clearly stated their names.

- *The abstract states that "The input for the Knime tool is an EXCEL file (or text or CSV) with a simple structure...", this structure should be described in the main text.*

In the data section, we now include a description of the input data, and we have also updated the underlying data for the input data files.

As for the minor comments, all the relevant mistakes pointed out have been corrected and updated. We once again thank you for your feedback and contribution and we welcome any further feedback.

Kind regards,
The authors

***Competing Interests:*** No competing interests were disclosed.

Reviewer Report 02 December 2020

https://doi.org/10.5256/f1000research.29685.r73564

?

**Marco Chierici**

Fondazione Bruno Kessler, Trento, Italy

The authors describe the Knime implementation of SVM-RCE-R, an algorithm for feature ranking and selection that is described as an improved version of the previously developed SVM-RCE. The motivation behind this new implementation is clearly stated and mainly resides in Knime's user-friendliness and ease of use, besides Knime being an Open Source project. Because of this, the software has the potential to be a valuable tool for researchers in the computational biology community.

My first concern is about potential confusion: it is not clear whether the manuscript focuses on the Knime implementation of a previously devised algorithm (SVM-RCE-R) or whether it introduces both the algorithm and its implementation. I think this aspect should be better stated in both the Abstract and the Introduction.

My other concerns follow.

**Conveyance of novelty:**

- The novelty of SVM-RCE-R vs. SVM-RCE, which, to my understanding, is the introduction of the ranking functions, should be pointed out more effectively. In particular, the Methods section opens with the description of the SVM-RCE algorithm: this comes unexpectedly to the reader since the manuscript should focus on describing SVM-RCE-R. The novelty element sneaks into the “Weighted rank function” subsection, while it should be better highlighted. This could be achieved by merely restructuring the Methods section or by adjusting the subsection titles.

**Data and analysis details:**

- It has to be clearly stated which datasets are used for each experiment discussed in the Results section: as far as I have understood, the experiments concerning the ranking functions (e.g., Figure 4) consider all 12 datasets, while those comparing the three approaches (e.g., Table 3) consider five datasets. Still, Figure 5 presents results for three datasets while the main text states “for each data set”. The beginning sentence of Results can be misleading, as it may seem that all results are obtained on five datasets.

- Were the GEO datasets used “as is” or was some preprocessing applied? More details should be included regarding the input format: for example, is it a gene expression table? What should be on the rows and the columns? Should there be row and column names?

**Presentation of results:**

- A comparison with SVM-RCE is only briefly mentioned, while the discussion would benefit from a more extended comparison (as done with SVM-RFE and SVM-IG), also to show SVM-RCE-R improvements over the previous version.

- The authors state that “SVM-RFE has a slightly better accuracy although significantly lower specificity than SVM-RCE-R”, but according to Table 3 it appears that average SVM-RFE accuracy is always lower than, or equal to, average SVM-RCE-R accuracy, and the same for SVM-RFE specificity (except for one dataset): please elaborate. Moreover, how was the significance assessed?

- For the ranking function, the authors chose accuracy, sensitivity, specificity, F1 score, AUC, and precision. This is quite a comprehensive set of metrics, which the authors could further enrich with the Matthews Correlation Coefficient (MCC), a balanced measure of accuracy and precision that can still be effectively used when sample classes are highly imbalanced.

- The notation used in Table 3 should be improved: it is misleading to indicate the weights using the metric names to which they refer. For example, Acc=0.2 should be replaced by $w_1$ =0.2, and so on. Please refer to the rank function notation introduced in “Weighted rank function” and rename the weights accordingly.

- Figure 4: since the goal is to compare different ranking functions, and the ranking function includes F1 among its terms, I suggest adding the average F1 to the figure. The same holds for Figure 5.

**Misc and minors:**

In general, please consider having the paper proof checked by a native English speaker to improve overall readability and address typos.

- Please consider improving the quality of Figures, especially 1-3.

- Please indicate the measure of dispersion represented by the error bars in the Figures.

- I am afraid the following sentence is not correct: “As illustrated in Pseudocode 1, in the original code of SVM-RCE we used the accuracy as the determinant for ranking the clusters.” There is no mention of accuracy in Pseudocode 1, only “performance”. Moreover, in Pseudocode 1 please consider improving “t = Test classifier on Xv –calculate performance” as it is not clear, and please change “aggregation the scores” to “aggregating the scores”.

- In Methods (p. 3): “test set performance” would be more accurate than “training-testing performance”.

- In Methods (p. 4): probably “prec” should be used instead of “pres”.

- In “Implementation in Knime”: the sentence “by configuring the node Counting Loop Start” more likely belongs to the end of the preceding paragraph.

- Figure 1 caption: please change “RCE based SVM” to “SVM-RCE” for consistency; please address the typo “compenents”.

- In “Operation” (p. 5): “Reserve” should be changed to “Rserve”.

- In Results (p. 5): please check “five datasets are considered for”.

- In Results (p. 6): "... six rank functions as listed in Table 2"; it should be Table 3.
- In Results (p. 9): please check "with 0.6 and weight 0.4".
- Please use either "meta node" or "meta-node" throughout the manuscript; the same holds for "SVM IG" vs SVM-IG.
- Please use "area under the curve" instead of "area under curve" (also in Figures).
- Please change w_1, w_2, ... and w1,..,w6 to $w_1$, $w_2$, ... (Results); please mind proper spacing and ellipsis in "w1,..w6".

**Is the rationale for developing the new software tool clearly explained?**
Yes

**Is the description of the software tool technically sound?**
Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**
Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**
Partly

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**
Yes

***Competing Interests:*** No competing interests were disclosed.

***Reviewer Expertise:*** Bioinformatics, artificial intelligence, computational biology

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 13 Dec 2020

**Malik Yousef**, Zefat Academic College, Zefat, Israel

Dear Marco Chierici,
Thank you for your constructive feedback. We found your insights helpful for improving the clarity of the manuscript. Below, we have addressed each of the items raised:

- *My first concern is about potential confusion: it is not clear whether the manuscript focuses on the Knime implementation of a previously devised algorithm (SVM-RCE-R) or* *whether it introduces both the algorithm and its implementation. I think this aspect should be better stated in both the Abstract and the Introduction.*

We have updated the abstract as well as the introduction to describe the novelty of the user specific ranking function as well the simplicity of KNIME implementation.

- *The novelty of SVM-RCE-R vs. SVM-RCE, which, to my understanding, is the introduction of the ranking functions, should be pointed out more effectively. In particular, the Methods section opens with the description of the SVM-RCE algorithm: this comes unexpectedly to the reader since the manuscript should focus on describing SVM-RCE-R. The novelty element sneaks into the "Weighted rank function" subsection, while it should be better highlighted. This could be achieved by merely restructuring the Methods section or by adjusting the subsection titles.*

In response to this suggestion, we have now changed the headings in the methods section. We now we have a specific subsection to bring the reader's attention to the user specific ranking function.

- *It has to be clearly stated which datasets are used for each experiment discussed in the Results section: as far as I have understood, the experiments concerning the ranking functions (e.g., Figure 4) consider all 12 datasets, while those comparing the three approaches (e.g., Table 3) consider five datasets. Still, Figure 5 presents results for three datasets while the main text states "for each data set". The beginning sentence of Results can be misleading, as it may seem that all results are obtained on five datasets.*

We have now made it clear in the results section about which datasets were used for which figures. Moreover, it is clearly stated results that we have used all the datasets for our results and specific datasets were used for comparison.

- *Were the GEO datasets used "as is" or was some preprocessing applied? More details should be included regarding the input format: for example, is it a gene expression table? What should be on the rows and the columns? Should there be row and column names?*

Based on your comment we have now included a description of the dataset as well. In addition, we have included the input data in our underlying data so that the results can easily be replicated.

- *A comparison with SVM-RCE is only briefly mentioned, while the discussion would benefit from a more extended comparison (as done with SVM-RFE and SVM-IG), also to show SVM-RCE-R improvements over the previous version.*

We agree, the section comparing SVM-RCE and SVM-RCE-R is not clearly mentioned. We have now included a sub section to indicate to the readers the comparison results.

- *The authors state that "SVM-RFE has a slightly better accuracy although significantly lower specificity than SVM-RCE-R", but according to Table 3 it appears that average SVM-RFE accuracy is always lower than, or equal to, average SVM-RCE-R accuracy, and the same for SVM-RFE specificity (except for one dataset): please elaborate. Moreover, how was the significance assessed?*

We have corrected the mistake which in the results for Table 3 and we pointed out that only SVM-RCE-R performs weaker in sensitivity for two of the datasets otherwise it outperforms or is on par with other approaches.

- *For the ranking function, the authors chose accuracy, sensitivity, specificity, F1 score, AUC, and precision. This is quite a comprehensive set of metrics, which the authors could further enrich with the Matthews Correlation Coefficient (MCC), a balanced measure of accuracy* *and precision that can still be effectively used when sample classes are highly imbalanced.*

We are planning to use the MCC in our further research where we search for the optimal combination of ranks. As you have mentioned it is a very good metric for highly imbalanced datasets. Thank you very much for the feedback about the metric.

- *The notation used in Table 3 should be improved: it is misleading to indicate the weights using the metric names to which they refer. For example, Acc=0.2 should be replaced by w1=0.2, and so on. Please refer to the rank function notation introduced in "Weighted rank function" and rename the weights accordingly.*

We have now updated the results in the table so that it is consistent with what is stated in the description of the ranking function.

- *Figure 4: since the goal is to compare different ranking functions, and the ranking function includes F1 among its terms, I suggest adding the average F1 to the figure. The same holds for Figure 5.*

We were focusing on the more widely used metrics to show the performance of our models, but as you have mentioned also looked into F-measure, so we now have included it in our results. All other typos and errors that you have so helpfully provided have been corrected in the revised manuscript.

Kind regards,
The authors

***Competing Interests:*** No competing interests were disclosed.