Comparative analysis of ReliefF-SVM and CFS-SVM for microarray data classification

Mochamad Agusta Naofal Hakim, Adiwijaya, Widi Astuti
Department of Informatic Engineering, Telkom University, Indonesia

ABSTRACT

Cancer is one of the main causes of death in the world where the World Health Organization (WHO) recognized cancer as among the top causes of death in 2018. Thus, detecting cancer symptoms is paramount in order to cure and subsequently reduce the casualties due to cancer disease. Many studies have been developed data mining approaches to detect symptoms of cancer through a classifying human gene data expression. One popular approach is using microarray data based on DNA. However, DNA microarray data has many dimensions that can have a detrimental effect on the accuracy of classification. Therefore, before performing classification, a feature selection technique must be used to eliminate features that do not have important information to support the classification process. The feature selection techniques used were ReliefF and correlation-based feature selection (CFS) and a classification technique used in this study is support vector machine (SVM). Several testing schemes were applied in this analysis to compare the performance of ReliefF and CFS with SVM. It showed that the ReliefF outperformed compared with CFS as microarray data classification approach.

Keywords: Classification Correlation-based feature selection Feature selection Microarray data ReliefF Support vector machine

This is an open access article under the CC BY-SA license.

1. INTRODUCTION

According to the World Health Organization (WHO), cancer is one of the world's leading causes of death where it is ranked second of the top ten causes of death globally, with estimates reaching 9.6 million deaths in 2018 [1]. Many studies have developed ways to detect cancer causality, and one of the popular approaches is through genetic expression. The human gene data has been expressed in the form of a storage medium called DNA microarray data [2]. Usually, the classification process must first be used on the DNA microarray data to detect the cancer causality. However, DNA microarray data has many dimensions that can adversely affect classification performance [1, 2]. One possible approach to overcome this problem is by selecting features that are considered more relevant in producing optimal classification performance [3]. Therefore, it is necessary to apply a feature selection technique prior to conducting the classification process to achieve more optimal cancer detection. However, not all feature selection techniques and implementation schemes can optimize classification process; many experiments are needed to compare feature selection techniques and the application of the classification scheme to achieve the most optimal results. The two popular approaches are ReliefF and correlation-based feature selection for feature selection process.

Journal homepage: http://ijece.iaescore.com
This paper would like to use two approaches, i.e., ReliefF and Correlation-based feature selection (CFS) to filter feature selection. These two techniques were chosen because they can independently assess the quality of good features, and have a simpler and faster computation time compared to other techniques [4, 5]; therefore, the filter technique is found appropriate to use for microarray data with many dimensions. The ReliefF approach is based on ranking approach, while the CFS is based on a subset approach. ReliefF gives a ranking value to each feature against its class attributes; the features with the highest weight will positively impact the classification process. Meanwhile, the CFS helps assess whether a subset of features uses merit’s calculations based on the correlation between features and classes, as well as the correlation between features with other features; the greater the merit’s value of a subset, the better its impact on the classification process [6]. The support vector machine (SVM) classification technique was chosen because it can produce better accuracy with microarray data compared to several other classification techniques [7-9].

The dataset used in this paper was taken from the Kent Ridge bio-medical dataset repository (http://leo.ugr.es/elvira/DBCRepository/) comprising colon tumour, ovarian, breast cancer, lung cancer, prostate, the central nervous system, and MLL leukaemia data. By conducting this research, the best implementation and performance of ReliefF and CFS techniques for improving the classification performance of the support vector machine (SVM) classification technique on microarray data can be identified. It was shown that the ReliefF outperformed compared with CFS for microarray data classification technique.

The rest of the paper is organized as follows: Section 2, discusses research methodology on how the step of research. The results and discussion are presented in section 3 to support the analysis of the results. Section 4 concludes with a summary of the comparison between two techniques.

2. RESEARCH METHOD

In the research development phase, two feature selection techniques were compared, namely ReliefF and correlation-based feature selection (CFS). These techniques are simple and widely used. This study aims to ascertain the performance of both techniques in the classification process using based on support vector machine (SVM).

Based on Figure 1, the system includes the normalization of data into values between [0, 1] [10]. The data was divided into training data and testing data, respectively. Then, the feature selection process, namely ReliefF and correlation-based feature selection (CFS) be carried out on the training data, producing top-rank features that were then used as new training data to form the support vector machine (SVM) classifier model [11]. The final step involved testing the system by inputting the testing data to determine the accuracy of the system.

Figure 1. General system flowchart
2.1. The datasets

The dataset used in this paper is microarray data comprising seven DNA microarrays. The dataset was obtained from the Kent-Ridge biomedical data repository, as reported by Li [10]. Each microarray data was grouped into training data and testing data according to the ratio listed in Table 1. This division was done by directly dividing the data randomly, along with the results of the distribution of training and testing data, as shown in Table 1.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Training</th>
<th>Testing</th>
<th>Number of Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>24481</td>
<td>78</td>
<td>19</td>
</tr>
<tr>
<td>Colon tumor</td>
<td>2000</td>
<td>43</td>
<td>19</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>15154</td>
<td>177</td>
<td>76</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>12533</td>
<td>149</td>
<td>32</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>12600</td>
<td>102</td>
<td>34</td>
</tr>
<tr>
<td>MLL leukemia</td>
<td>12583</td>
<td>57</td>
<td>15</td>
</tr>
<tr>
<td>Central Nervous System</td>
<td>7129</td>
<td>42</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 1. Distribution of data into testing data and training data

Normalization is useful for converting attribute values into a range between 0 and 1 so that the data will be easier to process [12]. The normalization was done using (1) [10].

\[ X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \] (1)

Equation (1) was used to calculate the normalization value, where \( X' \) is the result of normalization, \( X \) is original value of the attribute, \( X_{\min} \) is the smallest value of all the data in one attribute, and \( X_{\max} \) is the largest value of all the data in one attribute.

2.2. The ReliefF technique

The initial process of selecting the ReliefF feature requires the separation of training data based on class and the determination of the number of possible instances based on its class [13, 14]. The training data was separated according to its class to facilitate the search for near-hit and near-miss for each randomly selected instance, where near-hit and near-miss search were calculated based on the level of similarity of data using the Euclidian distance formula of (2) [10], and considering that the training data was numerical and continuous.

\[ d(x, y) = \sqrt{\sum (x_i - y_i)^2} \] (2)

Equation (2) was used to calculate distance value, where \( x \) and \( y \) are instances, and \( i \) is the attribute value. The near-hit and near-miss values determined were then used to calculate the weight of all features based on the formula of the ReliefF technique given by (3).

\[ W[F_i] = W[F_i] - \frac{\text{diff}(F_i, x_l, NH)}{m} + \sum_{x \in \text{class}(x_l)} \frac{[P(C) - \text{diff}(F_i, x_l, NM(C))]}{m} \] (3)

where \( W \) is the weight of each feature, \( F_i \) is the feature that is being calculated, \( x \) is the selected random instance, \( NH \) is the near-hit instance value, \( NM \) is the value of near-miss instance, \( P(C) \) is the prior probability of class \( C \), and \( m \) is the number of repetitions of the random instance retrieval [15, 16].

The function \( \text{diff}(F_i, I_1, I_2) \) was used to calculate the value of \( F_i \) using the difference in value between \( I_1 \) and \( I_2 \) instances. This calculation is defined by (4).

\[ \text{diff}(F_i, I_1, I_2) = \frac{\text{value}(F_i, I_1) - \text{value}(F_i, I_2)}{\text{max}(F_i) - \text{min}(F_i)} \] (4)

where \( F_i \) is the feature that is being calculated, \( I_1 \) is the selected random instance, and \( I_2 \) is the near-hit instances or near-miss instances [17]. The calculation of possible occurrences of instances using prior probability can be done using (5).

\[ P(C) = \frac{P_C}{1 - P_C} \] (5)
where \( C \) is a class that will calculate the probability of occurrence which consists of \( P_c \) a class probability and \( P_{ck} \) an instance probability. The following section provides a flowchart describing the system and process of the ReliefF feature selection technique as shown in Figure 2.

![Figure 2. ReliefF feature selection flowchart](image)

2.2.1. The ReliefF algorithm

The algorithm of ReliefF technique is adopted from [18], it is based on several steps as shown:

The Input: for each training instance a vector of attribute values and the class value

Output: The vector \( w \) of estimations of the qualities of attributes.

1. set weight of features \( W[A] := 0.0 \)
2. for \( i := 1 \) to \( m \) do
3. select an instance \( r_i \) randomly;
4. find \( k \)-nearest hits \( h_j \);
5. for each \( C \neq \text{class}(r_i) \) do
   6. from class \( C \) find \( k \) nearest misses \( m_j(C) \);
   7. for \( A := 1 \) to \( a \) do
5. Calculate:

\[
W[A] = W[A] - \sum_{j=1}^{k} \frac{d_i f f (A, r_i, h_j)}{(m, k)} + \frac{\sum_{C \neq \text{class}(r_i)} P_c}{1 - \text{class}(r_i)} \frac{\sum_{i=1}^{k} d_i f f (A, r_i, h_j)}{(m, k)}
\]

9. end.

2.3. The CFS technique

The CFS runs by calculating components based on a heuristics value called the \( \text{Merit}_s \) value, which represents the quality of each feature combination or feature subset. \( \text{Merit}_s \) is calculated using (6):

\[
\text{Merit}_s = \frac{k \bar{r}_{ff}}{k+k(k-1)\bar{r}_{ff}}
\]

where \( \text{Merit}_s \) is subset feature value, \( k \) is number of features, \( \bar{r}_{ff} \) is average value of class minus the feature correlation, and \( \bar{r}_{ff} \) is average value of feature minus the feature intercorrelation [6].
The components of $Merit_s$ include the correlation value of a feature with other features and the correlation between features and classes owned. In this case, a $Merit_s$ value search was done using the forward selection algorithm, starting with a blank feature subset until the best combination according to the threshold was found [19, 20]. Then, the $Merit_s$ value of each subset will be calculated, and subset with the biggest $Merit_s$ value will be selected. Figure 3 shows the CFS flowchart.

![CFS Flowchart](image)

**Figure 3. CFS flowchart**

Using ReliefF and CFS, the features of the microarray data were obtained. These features were then used as parameters for developing the classification model. The development of the SVM classification model includes developing the best hyperplane model to separate the data according to each class, based on supporting points or support vectors that are at the class separation limit [21, 22]. The selection of points as support vectors was influenced by the shape and character or condition of the features of the data. By getting the best features, the margin on the support vector can be maximized [23-25]. Equation (7) presents the function that must be optimized from the hyperplane margin where $w$ is the unit vector found in the hyperplane.

$$
\min_w c(w) = \frac{1}{2} ||w||^2
$$

However, before using the data to develop the model, the dataset was first divided into testing data and training data. Because the testing data consists of 2 and 3 classes, the type of kernel to be used for SVM was first tested against all data. The kernel that showed good compatibility and good accuracy for all data was then used as one of the parameters for SVM.

### 2.3.1. The CSF algorithm

The algorithm of CSF technique using forward selection begins with the zero value of subset then add a feature one by one and calculate the $Merit_s$ value of each features combination greedily [19]. It is based on several steps as shown:

The Input: For each training instance a vector of attribute values and the class value.
Output: maximum subset $Merit_s$ value.

1. set subset $S := [\ ]$, threshold := determined by writer
2. while threshold != 0
3. for each attribute
4. $r_c :=$ average attribut relation with class
5. $r_f :=$ average attribut relation with another attribut
6. $k :=$ count number of subset
7. $Merit_s = \frac{r_c}{(r_c + r_f)}$
8. if maximum $Merit_s$ value has finded then
9. reinitialize threshold number to determined value
10. else threshold minus 1
11. end.

*Comparative analysis of ReliefF-SVM and CFS-SVM for microarray... (Mochamad Agusta Naofal Hakim)*
3. RESULTS AND DISCUSSION

In results and discussions section, will discuss about results of research. The result are divided into three scenarios and presented in a figure and table.

3.1. Scenario 1

In this testing process, a comparison was made between the kernels to be used in the SVM classifier according to the data distribution and to provide the most optimal accuracy results. The kernels chosen included the RBF, Polynomial, and Linear kernels with parameters (d) degrees with a value of 3 and (C) with a value of 1, while the testing data consisted of 7 cancer data. Table 2 shows the results of the 3 types of kernel experiments for SVM using microarray data.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Colon tumor</th>
<th>Lung cancer</th>
<th>Prostate</th>
<th>MLL leukemia</th>
<th>Breast cancer</th>
<th>Ovarian</th>
<th>Central nervous system</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBF</td>
<td>78.95</td>
<td>50</td>
<td>50</td>
<td>80</td>
<td>36.84</td>
<td>88.16</td>
<td>72.22</td>
<td>65.17</td>
</tr>
<tr>
<td>Polynomial</td>
<td>68.42</td>
<td>50</td>
<td>73.53</td>
<td>73.33</td>
<td>36.84</td>
<td>68.42</td>
<td>61.11</td>
<td>61.67</td>
</tr>
<tr>
<td>Linear</td>
<td>73.68</td>
<td>84.36</td>
<td>94.12</td>
<td>86.67</td>
<td>84.21</td>
<td>100</td>
<td>61.11</td>
<td>83.45</td>
</tr>
</tbody>
</table>

In testing the system with the first testing scenario, the accuracy, as shown in Table 2, indicates that the most suitable SVM kernel for all testing data was the linear kernel, which returned the highest average accuracy of 83.45%, while the RBF kernel produced an average accuracy of 65.17% and the Polynomial kernel reached 61.67% accuracy only. Although the linear kernel produced good average accuracy, in tests conducted on some data such as the central nervous system and colon tumor data, it still produced a lower accuracy compared to the RBF kernel. From the results, it can be concluded that most of the distribution of data used in this test followed a linear or linear separable data distribution, except for some data such as the central nervous system and colon tumor data.

3.2. Scenario 2

The second testing scenario produced values in the form of an optimal number of features, the highest accuracy, and the average accuracy for each feature subset selection, as outlined in Table 3. A threshold parameter with a value of 300 was inserted in this paper. A threshold is used as a delimiter of the number of features that will be tested as a feature subset. Five features were taken in a subset until the threshold was determined and the number was searched. The feature subset should provide the most optimal accuracy. Besides the parameter in the form of a threshold, the number of Nearest Instance to be used was also tested in this scenario, where the nearest instance value was chosen randomly from 1, 2, or 3. The CFS threshold parameter values of 3 and 5 were inserted in this paper. The CFS threshold parameter was used as a limit to allow the forward selection algorithm to widen its search if the new Merit results were lesser than the results of previous Merit. Otherwise, another subset will be given the chance to add members in the form of new features.

<table>
<thead>
<tr>
<th>Data</th>
<th>SVM+ReliefF</th>
<th>SVM+CFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon tumor</td>
<td>100</td>
<td>89.47</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>96.88</td>
<td>84.38</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>94.12</td>
<td>82.35</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>100</td>
<td>98.68</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>84.21</td>
<td>63.16</td>
</tr>
<tr>
<td>MLL leukemia</td>
<td>100</td>
<td>86.67</td>
</tr>
<tr>
<td>Central nervous system</td>
<td>88.89</td>
<td>83.33</td>
</tr>
</tbody>
</table>

Based on Table 3, when testing the SVM classification technique using the ReliefF selection feature, the colon tumor data was classified with an accuracy of 100% with 15 features and 3 nearest instances. The lung cancer data was classified with 96.88% accuracy with 10 features and 3 nearest instances. The ovarian cancer data was classified with 100% accuracy in all number of instances, with the least features being 130 features. The prostate cancer data was classified with a maximum accuracy of 94.12% in all instances and 170 as the minimum number of features. Breast cancer data was classified with 84.21% accuracy with at least...
65 features and a total of 3 nearest instances. The MLL leukemia data with 3 classes was classified with 100% accuracy at all nearest instances, with 5 features minimum at the nearest instances of 1 and 2. The central nervous system data was classified with the highest accuracy of 88.89% at the nearest instances of 2 and 3 and a minimum number of features of 125 at instance 2.

From the testing scenarios conducted, it can be said that using the ReliefF feature selection on all the data successfully provided more optimal results, compared with scenarios without using ReliefF. This is evidenced by the increased accuracy with feature reduction but for some data in particular such as colon tumor, lung cancer, MLL leukemia, and the central nervous system data, which saw increased classification accuracy in all test scenarios.

The number of features was not the only parameter that affected classification accuracy. If the nearest instance parameter for some data were to be observed such as colon tumor, prostate cancer, ovarian cancer, lung cancer, MLL leukemia, and the central nervous system data, an increase in the number of the nearest instances resulted in an accuracy that was not lower than the accuracy with the smaller nearest instance, although the breast cancer data at the second nearest instance had a lower accuracy than the nearest instance of 1. However, in the third instance, the accuracy again increased and was greater than the second instance. This is likely because there is a nearest instance that calculates the weight of the feature that made it less optimal, so the top features that had weight were the most likely to contain noise.

3.3. Scenario 3

For the third testing scenario shown in Table 4, an increase in accuracy was observed in almost all microarray data after the selection feature technique was applied. The colon tumor data was classified with 100% and 89.47% accuracy using the ReliefF technique and the CFS technique, respectively. The Central Nervous System data was initially classified with an accuracy of 61.11%, which increased to 88.89% using ReliefF and 83.33% using CFS. Besides that, the classification accuracy of lung cancer, prostate cancer, and MLL leukemia data also increased using the ReliefF technique, but the CFS technique produced the same accuracy for some data-the same as that of testing without feature selection i.e. 84.37% for lung cancer data, and 86.67% for MLL leukemia data. Decreased accuracy was observed in some datasets such as from 84.21% to 63.16% for breast cancer data; from 100% to 98.68% for ovarian cancer data; and from 94.12% to 82.36% for prostate cancer data.

<table>
<thead>
<tr>
<th>Data</th>
<th>SVM</th>
<th>SVM+ReliefF</th>
<th>SVM+CFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon tumor</td>
<td>73.68</td>
<td>100</td>
<td>89.47</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>84.38</td>
<td>96.88</td>
<td>84.38</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>94.12</td>
<td>94.12</td>
<td>82.35</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>100</td>
<td>100</td>
<td>98.68</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>84.21</td>
<td>84.21</td>
<td>63.16</td>
</tr>
<tr>
<td>MLL leukemia</td>
<td>86.67</td>
<td>100</td>
<td>86.67</td>
</tr>
<tr>
<td>Central nervous system</td>
<td>61.11</td>
<td>88.89</td>
<td>83.33</td>
</tr>
</tbody>
</table>

From Table 4 an Figure 4, the SVM classification technique using CFS for colon tumor data produced an accuracy of 89.47% with 16 features and a threshold of 3. For the lung cancer data, classification accuracy reached 84.375% with 45 features at a threshold value of 3; ovarian cancer data achieved 98.68% accuracy for all numbers of instances, with the least number of features (30): prostate cancer data yielded a maximum accuracy of up to 82.35% at a threshold of 5 with 46 features; breast cancer data obtained 63.16% accuracy with 51 features at a threshold value of 5; MLL leukemia data with 3 classes obtained an accuracy of 86.67% at a threshold of 5 with 19 features; and the central nervous system data achieved an accuracy of 72.22% with 12 features, for all thresholds.

From the testing scenarios conducted, using CFS feature selection for some data such as MLL leukemia and lung cancer produced more optimal results compared to only SVM. Meanwhile, the colon tumor and central nervous System data also produced more optimal results and experienced increased accuracy compared to the SVM scenarios only, as shown by the fewer number of features used, but accuracy was still maintained and improved. However, in some data such as breast cancer, prostate cancer, and ovarian cancer data, in the testing scenarios, the accuracy of each experiment decreased. This is likely because the *Merit* search process used a Greedy forward selection technique that was not optimal, as can be proven from the results of the number of features obtained by the ReliefF technique for the three data above, which was a larger number than the other data, while the maximum feature subset found by CFS was lower in comparison.

*Comparative analysis of ReliefF-SVM and CFS-SVM for microarray... (Mochamad Agusta Naofal Hakim)*
As for the effect of threshold on the accuracy produced, almost all data with more thresholds had an accuracy that was equal or better than smaller threshold values. Hence, it can be concluded that a threshold value of 5 still allowed for increased accuracy even with an additional number of features. However, the lung cancer data threshold of 3 had better accuracy than a threshold of 5, likely due to the presence of noise when the threshold was reduced.

Besides that, the effect of accuracy generated from the ReliefF-SVM testing scheme can be said to be stable against the ratio of various data-solving ratios. This is evident for the data with a ratio of 70% such as the ovarian cancer, colon tumor, and the central nervous system training data, which produced increased accuracy. For lung cancer and MLL leukemia data, with more training data according to a ratio of 80% to 82%, an increase in accuracy could also be achieved. In addition, although there was no increase in accuracy for the prostate cancer and breast cancer data using a training data ratio of 75%-80%, the number of features were still substantially reduced and the same accuracy with testing was observed without using the feature selection technique. This result was also observed in other testing data using the SVM-CFS scheme, with data based on a ratio of 70% in the colon tumor and the central nervous system training data. Both of which was able to provide increased accuracy. Meanwhile, for MLL leukemia data and Lung cancer data with a ratio of 80%-82%, although there was no increase in accuracy, the number of features managed to be reduced substantially. However, for the prostate cancer, ovarian cancer, and breast cancer data with different data ratios, all experienced a decrease in accuracy, but the ratio was probably not the main cause of decreased accuracy; and other factors such as thresholds in the scheme could also cause this result.

4. CONCLUSION AND FUTURE WORK

Based on the results of the testing scenario conducted in this study, it can be concluded that the ReliefF and correlation feature selection (CFS) techniques on microarray data classification using support vector machine (SMV) can generally provide more optimal results compared to the classification process without using feature selection. However, some tests under the CSF-SVM scenarios produced results with decreased accuracy compared to SVM without feature selection.

In SVM testing, without using feature selection, by testing several types of kernels to determine optimal classification performance, the best accuracy was achieved using the Linear kernel, which returned an average accuracy of 83.45%. In contrast, the RBF kernel produced an average accuracy of 65.17% and the Polynomial kernel produced 61.67% accuracy.

The final accuracy obtained from the three testing scenarios for all data under the SVM scenario without using the feature selection technique was an average accuracy of 83.45%. Meanwhile, ReliefF-SVM produced an average accuracy of 94.87%, and CFS-SVM produced 84% accuracy. From these results, it can be concluded that the testing scheme involving ReliefF as the feature selection technique with SVM had the best classification accuracy.

Some suggestions related to this research include improving the CFS-SVM testing scheme using algorithms for searching such as Forward Selection and with a CFS threshold with 5 as the maximum. This would still return a high probability of increased accuracy, so it is better for future studies to use a greater...
value for data with large dimensions such as microarray data. Then, to provide more valid test results for all data used, cross-validation evaluation could be done to provide more valid evaluation results because this technique treats all data including testing data and training data.

REFERENCES


BIOGRAPHIES OF AUTHORS

Mochamad Agusta Naofal Hakim was born in Nganjuk, Indonesia, in 2015 received the Bachelor degree of Informatics Engineering from Telkom University Bandung, Indonesia. His research interest includes on Data Mining, Machine Learning, and Microarray Data.

Adiwijaya is a professor of mathematics at School of Computing, Telkom University. He is interested in the research area of graph theory and its applications, data science, and information science. He joined Telkom University since 2000 and has become professor since 2016.

Widi Astuti obtained bachelor degree and master degree of Informatics at Telkom University. Her research interests include data mining, big data, machine learning and artificial intelligence.