This thesis proposed a ranking and clustering-based feature selection framework for high-dimensional data classification. The proposed method conducted the relevancy analysis using ranking approach in order to identify the irrelevant features followed by the relevancy analysis using clustering approach to identify the redundant feature and remove them. As the result of removing the irrelevant and redundant features from the high-dimensional space, the classification accuracy was improved for high-dimensional data.

Initially, the performance of the feature subset-based and feature ranking-based feature selection methods was analyzed. Then, the performance of different clustering techniques was validated for observing their suitability for redundancy analysis in high-dimensional space. Then, the filter-based feature selection method was proposed to reduce the redundancy and to improve the classification accuracy. Further, the ranking and clustering-based feature selection framework was proposed. The proposed framework was tested on various real world high-dimensional datasets and its performance was compared with various state-of-the-art features selection methods in terms of classification accuracy, runtime, and redundancy rate. The rest of this chapter summarizes the results of the experiments conducted to validate the performance of the proposed work.
Initially, the performance of the feature subset-based and the feature ranking-based features selection methods were evaluated in Section 5.1. In this investigation, various feature subset-based methods and feature ranking-based methods were employed and the performance of these methods was tested on a wide range of datasets. Three different types of classifiers were adopted to observe their performance in terms of classification accuracy. The summary of the experimental results are tabulated in terms of average classification accuracy and number of selected features in Table 6.1.

Table 6.1  Summary of the performance of the feature subset-based and feature ranking-based methods in terms of average classification accuracy and number of selected features

<table>
<thead>
<tr>
<th>Feature selection methods</th>
<th>Average classification accuracy in %</th>
<th>No. of selected features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NB</td>
<td>J48</td>
</tr>
<tr>
<td>Feature subset-based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRFS</td>
<td>78.55</td>
<td>77.27</td>
</tr>
<tr>
<td>COFS</td>
<td>78.03</td>
<td>80.94</td>
</tr>
<tr>
<td>Feature ranking-based</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CQFS</td>
<td>77.76</td>
<td>80.28</td>
</tr>
<tr>
<td>IGFS</td>
<td>77.92</td>
<td>80.21</td>
</tr>
</tbody>
</table>

From Table 6.1, it is evident that there is no significant difference in the performance of the feature subset-based and feature ranking-based feature selection methods in terms of number of features selected and the classification accuracy produced with various classifiers. However, as observed from the literatures, the subset-based feature selection methods have high computational and space complexity. Hence, the ranking-based approaches are suitable for feature selection, especially in the high-dimensional space.
Then a pragmatic investigation on clustering methods was conducted as discussed in Section 5.2 in order to identify a suitable clustering method for redundancy analysis for the proposed framework. Therefore, three different clustering methods namely, K-means clustering (KC), expectation maximization clustering (EC), and hierarchical clustering (HC) were adopted. The performances of these clustering methods were evaluated in terms of average intra-cluster redundancy rate and runtime. The summary of results of this investigation is tabulated in Table 6.2.

The better clustering method produces the higher average intra-cluster redundancy rate. This is due to the fact that the higher average intra-cluster redundancy rate depicts that the corresponding cluster contains similar features. From Table 6.2, it is observed that the KC clustering method performs better in the redundancy analysis for the high-dimensional space in terms of redundancy rate and the runtime since it produces higher redundancy rate and reduces the runtime for the high-dimensional data. Further, EC cluster has more computational complexity than KC and its overall performance in terms of intra-cluster redundancy rate is better than HC. However, HC takes more time to form clusters due to the inherent computational complexity and it exhibits poor performance in terms of overall intra-cluster redundancy rate compared to EC and KC. Further, HC induces buffer overflow when the number of features is more (high-dimensional data) due to high space complexity. Therefore, HC is not a suitable choice for redundancy analysis in high-dimensional space. So, KC clustering technique seems to perform better even in high-dimensional space compared to EC and HC clustering techniques.
Table 6.2  Summary of average runtime and mean value of average intra-cluster redundancy rate of KC, EC, and HC clustering methods on all datasets except ORL10P and PIX10P (since HC induces buffer overflow for the datasets ORL10P and PIX10P) with respect to number of clusters

<table>
<thead>
<tr>
<th>No. of Clusters</th>
<th>Runtime</th>
<th>Redundancy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KC</td>
<td>EC</td>
</tr>
<tr>
<td>2</td>
<td>0.594</td>
<td>3.257</td>
</tr>
<tr>
<td>3</td>
<td>0.593</td>
<td>7.509</td>
</tr>
<tr>
<td>4</td>
<td>1.233</td>
<td>4.942</td>
</tr>
<tr>
<td>5</td>
<td>1.139</td>
<td>11.013</td>
</tr>
<tr>
<td>6</td>
<td>2.225</td>
<td>11.433</td>
</tr>
<tr>
<td>7</td>
<td>1.544</td>
<td>10.462</td>
</tr>
<tr>
<td>8</td>
<td>3.791</td>
<td>22.649</td>
</tr>
<tr>
<td>9</td>
<td>3.535</td>
<td>12.791</td>
</tr>
<tr>
<td>10</td>
<td>3.179</td>
<td>12.273</td>
</tr>
</tbody>
</table>

A filter-based feature selection method was proposed in Section 5.3. The prime objective of this work is to remove the irrelevant and redundant features in order to improve the classification accuracy of the classifier. Therefore, the proposed method was developed by employing mechanisms for both relevancy and redundancy analysis. The information gain measure and the KC clustering were used for relevancy and redundancy analysis, respectively. The performance of the proposed method was tested on various real-world datasets. Three classifiers namely NB, J48, and IB1 were used to analyze the performance of the proposed and other feature selection methods in terms of classification accuracy. The redundancy rate was also calculated from the selected features to observe the performance of the proposed method in terms of redundancy reduction. The summary of the...
classification accuracy and the redundancy rate of the proposed and other methods compared are tabulated in Table 6.3.

Table 6.3 Summary of classification accuracy and redundancy rate of the feature selection methods on datasets

<table>
<thead>
<tr>
<th>Feature selection method</th>
<th>Average classification accuracy in %</th>
<th>Average redundancy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NB</td>
<td>J48</td>
</tr>
<tr>
<td>Proposed</td>
<td>68.90</td>
<td>73.82</td>
</tr>
<tr>
<td>IGFS</td>
<td>65.53</td>
<td>69.86</td>
</tr>
<tr>
<td>CQFS</td>
<td>65.02</td>
<td>69.70</td>
</tr>
<tr>
<td>GRFS</td>
<td>65.24</td>
<td>68.25</td>
</tr>
<tr>
<td>SUFS</td>
<td>65.46</td>
<td>70.19</td>
</tr>
</tbody>
</table>

From the Table 6.3 it is discovered that the proposed method produces the better classification accuracy compared to the other feature selection methods. Further, it reduces the redundancy significantly compared to other feature selection methods. It is also apparent that the ranking methods mainly concentrate on the relevancy analysis therefore they fail to reduce the redundancy and do not produce better classification accuracy. Nevertheless, the symmetric uncertainty-based feature selection (SUFS) produces the better classification accuracy compared to the information gain-based feature selection (IGFS).

Considering the above facts, the proposed ranking and clustering based feature selection framework was developed using IRFE algorithm with both relevancy and redundancy analysis mechanisms as discussed in Section 5.4. The symmetric uncertainty measure and the KC clustering were used for relevancy and redundancy analysis, respectively. The performance of the proposed framework IRFE was tested on various real world high-dimensional
datasets and compared with various state-of-the-art feature selection methods such as OneR, ReliefF, JMI, MRMR, MIFS, CMIM, DISR, CIFE, CondRed, and MCFS. Three classifiers namely NB, J48, and kNN were used to evaluate the performance of the proposed framework IRFE and other state-of-the-art feature selection methods in terms of classification accuracy. The performance of the proposed framework in reducing the redundancy and improving the classification accuracy with respect to the number of feature clusters formed were also be analyzed. The summary of the classification accuracy and runtime of the proposed feature selection framework and other state-of-the-art feature selection methods are tabulated in Table 6.4.

Table 6.4  Comparison of the proposed feature selection framework and other state-of-the-art feature selection methods in terms of average classification accuracy and runtime

<table>
<thead>
<tr>
<th>Feature selection method</th>
<th>Average classification accuracy</th>
<th>Average runtime (Seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NB</td>
<td>J48</td>
</tr>
<tr>
<td>OneR</td>
<td>0.7578</td>
<td>0.7309</td>
</tr>
<tr>
<td>ReliefF</td>
<td>0.7499</td>
<td>0.7234</td>
</tr>
<tr>
<td>JMI</td>
<td>0.7691</td>
<td>0.7082</td>
</tr>
<tr>
<td>MRMR</td>
<td>0.7756</td>
<td>0.7300</td>
</tr>
<tr>
<td>MIFS</td>
<td>0.6660</td>
<td>0.6724</td>
</tr>
<tr>
<td>CMIM</td>
<td>0.6996</td>
<td>0.6855</td>
</tr>
<tr>
<td>DISR</td>
<td>0.7723</td>
<td>0.7172</td>
</tr>
<tr>
<td>CIFE</td>
<td>0.6324</td>
<td>0.6464</td>
</tr>
<tr>
<td>CondRed</td>
<td>0.5961</td>
<td>0.5770</td>
</tr>
<tr>
<td>MCFS</td>
<td>0.7428</td>
<td>0.6921</td>
</tr>
<tr>
<td>IRFE</td>
<td>0.8198</td>
<td>0.7797</td>
</tr>
</tbody>
</table>
From Table 6.4, it is obvious that the proposed feature selection framework produces higher classification accuracy compared to all other state-of-the-art feature selection methods and takes lesser computation time than all the methods compared except MCFS. However, MCFS does not produce better classification accuracy.

6.1 FUTURE SCOPE

In general, the feature selection is employed for machine learning applications for removing the irrelevant and redundant features from the dataset. The proposed feature selection framework can be extended with different mechanisms for relevancy or redundancy analysis. This proposed framework can be employed for the classification applications especially where the high-dimensional space are available.
REFERENCES


