Feature selection algorithm for high dimensional biomedical data classification based on redundant removal

Bingtao Zhang¹,², Peng Cao³, Yi Zhang¹, Chaochao Zhang¹, Zhe Li¹, Zhidiao Qu¹, Xiaopeng Wang², Tao Lei¹, Hanshu Cai¹*, Bin Hu¹*
¹ School of Information Science and Engineering, Lanzhou University, Lanzhou, China
² School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, China
³ School of Architecture and Urban Planning, Lanzhou Jiaotong University, Lanzhou, China
* Corresponding author

Address: School of Information Science and Engineering, Lanzhou University, Tianshui South Road, Lanzhou, Gansu Province, China
E-mail: zhangbt14@lzu.edu.cn, 174601298@qq.com, zhangyi2014@lzu.edu.cn, zhangchch2017@lzu.edu.cn, lizh2014@lzu.edu.cn, quzhid17@lzu.edu.cn, wangxiaopeng@mail.lzjtu.cn, leitao@sust.edu.cn, caihsh13@lzu.edu.cn, bh@lzu.edu.cn

High dimensional biomedical data contain thousands of features, and accurate identification of the main features in these data can be used to classification related data. However, it is usually a large number of irrelevant or redundant features seriously influence classification accuracy. To solve this problem, a new feature selection algorithm based on redundant removal is proposed in this study. Firstly, two redundant criteria are determined by vertical relevance and horizontal relevance. Secondly, an approximate redundancy feature framework based on mutual information (MI) is defined to remove redundant and irrelevant features. Finally, to evaluate the effectiveness of our proposed method, contrast experiments based on the classic feature selection algorithm are conducted using (K-nearest neighbour) KNN classifiers, and the results show that our algorithm can effectively improve the classification accuracy.

1. INTRODUCTION

High dimensional data analysis (Tamaresis et al., 2014) is a very hot research area, especially in cancer data (Lee et al., 2013 ), or mental illness data (Jiang et al., 2017; Li et al., 2017). Usually high dimensional data contain many weak relevance or irrelevance features. Hypothesis all the features are treated equally, computational complexity and accuracy of the prediction can be seriously affected. Therefore, feature selection is considered to be an essential procedure in high dimensional data processing.

Feature selection (Saefys et al., 2007) refers to selecting relevant features while to remove irrelevant and redundant features. As one of the important part of knowledge discovery technology, feature selection can effectively improve the computing speed of subsequent prediction algorithm, enhance the compactness of the prediction model, increase the generalization ability of the corresponding model. Additionally, the major purpose of high dimensional data feature selection is to overcome the curse of dimensionality (Li et al., 2016; Zhang et al., 2018).

In general, the process of feature selection mainly consists (Mafarja et al., 2018): search strategy and evaluation criterion. Evaluation criterion can be categorized into the wrapper method and the filter method. The wrapper method (Chrysostomou et al., 2017) to evaluate superiority and inferiority of the optimal feature subset under the premise of keeping the classification algorithm unchanged. And the corresponding classification accuracy is adopted as an index to select optimal feature subset. It is necessary to execute the feature selection process again when the classification algorithm is changed. Hence, the complexity is too high, especially for high dimension data. The filter method (Hancer et al., 2018; Lei et al., 2018), the search of feature space depends on the intrinsic correlation of the data itself rather than the
classification algorithm. The filter method is increasingly attractive because of its simplicity and fast speed. So filter method is more general applied than the wrapper method.

According to the above discussion, in this paper, a filter feature selection method is proposed. First of all, four kinds of boundary extremes are analyzed, and then two redundant criterions are proposed. Meanwhile, in order to quantify the redundancy criterion, the core module based on mutual information (MI) (Estevez et al., 2009) is proposed: the definition of approximate redundancy feature. Finally, the experiment is given.

The remainder of this article is organized as follows: Section 2 provides basic concepts related to this research. A feature selection algorithm based on redundancy removal is proposed in Section 3. In Section 4, we describe our experimental design, experimental results. Finally, in section 5 concludes the work of this study.

2. BASIC CONCEPTS

In order to facilitate follow-up research, some basic concepts (John et al., 1994) used in this study are listed as follows.

(i) Strong relevance: Fi is strongly relevant feature iff there exists \(P(F_i | A_j) > 0\) such that
\[
P(C | F_i, A_j) 
eq P(C | A_j)
\] (1)

(ii) Weak relevance: Fi is weakly relevant feature iff it is not strongly relevant (i.e. \(P(C | F_i, A_j) = P(C | A_j)\)), there exists \(A' \subseteq A\) and \(P(F_i, A') > 0\) such that
\[
P(C | F_i, A') 
eq P(C | A')
\] (2)

(iii) Irrelevance: Fi is irrelevent feature iff it are not strongly relevant and weakly relevant, there all \(A' \subseteq A\) and \(P(F_i, A') > 0\) such that
\[
P(C | F_i, A') = P(C | A')
\] (3)

where \(P\) is a probability measure. \(F\) is feature set, \(F = \{F_1, F_2, ..., F_n\}\). \(F_i: F_i = \{f_1, f_2, ..., f_n\}\). \(A_i: A_i = F - \{F_i\}\). \(C\) is class attribute. \(C = \{C_1, C_2, ..., C_n\}\).

Strong relevance shows that the feature is very important for classification accuracy, so it can’t be arbitrarily removed. Weak relevance indicates that this feature can sometimes contribute to improve prediction accuracy. Irrelevance indicates that this feature is useless on the improvement of classification accuracy, so it can be directly deleted.

3 METHOD

3.1 Redundancy criterion

A redundancy criterion based on the correlation is proposed to lay the foundation for further feature selection. Based on three basic concepts in the section 2, the redundancy of feature \(F_i\) is analyzed under four extreme values of \(R_{i,c}\) (the relevance between any feature \(F_i\) and class attribute \(C\)) and \(R_{ij}\) (the relevance between any pair of feature \(F_i\) and \(F_j\), \(i \neq j\)). And four extreme values are shown in table 1.

<table>
<thead>
<tr>
<th>(R_{i,c})</th>
<th>(R_{ij})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>large</td>
</tr>
<tr>
<td>2</td>
<td>large</td>
</tr>
<tr>
<td>3</td>
<td>small</td>
</tr>
<tr>
<td>4</td>
<td>small</td>
</tr>
</tbody>
</table>

From table 1, it is easy to draw the following conclusions:

Conclusion 1: \(R_{i,c}\) is large, which means that \(F_i\) contains more information about \(C\). \(R_{i,c}\) is large, which means that the correlation between \(F_i\) and \(F_j\) is strong. If \(R_{i,c} = 1\), then \(F_i\) and \(F_j\) is complete correlation, hence \(F_i\) is redundant. If \(R_{i,c} \neq 1\), it is difficult to determine the feature \(F_i\) whether or not is redundant.

Conclusion 2: \(R_{ij}\) is small, which means that the correlation between \(F_i\) and \(F_j\) is weak. Hence \(F_i\) can’t replace \(F_j\). In other words, no matter the size of the \(R_{i,c}\), the feature \(F_i\) is not redundant.

Conclusion 3: \(R_{i,c}\) is small, which means that \(F_i\) contains less information about \(C\). \(R_{ij}\) is large, which means that the correlation between \(F_i\) and \(F_j\) is strong. In this case, the feature \(F_i\) is redundant with higher probability. With the increase of \(R_{ij}\), this probability is also increasing.

Conclusion 4: \(R_{ij}\) is small, which means that the correlation between \(F_i\) and \(F_j\) is weak. This conclusion is consistent with the conclusions 2, no matter the size of the \(R_{i,c}\), the feature \(F_i\) is not redundant.

Based on the above four conclusions, two redundant criterions can be obtained:

Criteria 1: when \(R_{ij}\) is large, whether \(F_i\) is redundant is uncertain.

Criteria 2: when \(R_{ij}\) is small, no matter the size of the \(R_{i,c}\), the feature \(F_i\) is not redundant.

3.2 Approximate redundancy feature

Assuming that the \(R_{i,c}\) of the feature \(F_i\) is very close to \(R_{max}\) (the maximum value of \(R_{i,c}\)), it indicates that
\( F_i \) contains a lot of information about class attribute \( C \). In this condition, only if the value of \( R_{ij} \) is large enough, \( F_i \) can be considered as an approximate redundancy feature. Otherwise, it can’t be considered as redundancy feature. The reason is that \( F_i \) plays an important role in improving the accuracy of classification, and can’t be easily removed as redundancy. By contrast, Assuming that the \( R_{ic} \) of feature \( F_i \) is not very close to \( R_{max} \), it indicates that \( F_i \) contains relatively less information about class attribute \( C \). In this condition, As long as the value of \( R_{ij} \) is relatively large, \( F_i \) is considered as an approximate redundancy feature. The reason is that \( F_i \) is not plays a main role in improving the accuracy of classification. In addition to the above conditions, \( F_i \) is removed as an approximate irrelevance feature when the difference between \( R_{ic} \) and \( R_{max} \) is quite large. Based on the above analysis and discussion, the approximate redundant feature is formally described in definition 1.

**Definition 1** (approximate redundancy feature): There is any pair of correlation feature \( F_i \) and \( F_j \), and \( R_{ic} \geq R_{ij} \).

(i) \( F_i \) is an approximate redundancy feature iff there exists \( R_{max} - R_{ij} \leq \delta, 0 \leq \delta \leq 0.2 \), such that

\[
R_{i,j} \geq R_{max} \tag{4}
\]

(ii) \( F_i \) is an approximate redundancy feature iff there exists \( R_{max} - R_{ij} > \delta \) & \( R_{max} - R_{ij} \leq \alpha \),

\[
0 \leq \delta \leq 0.2, 0.5 \leq \alpha \leq 0.7 \text{ such that } \\
R_{i,j} \geq (R_{ij} + R_{ic}) / 2 \tag{5}
\]

Where \( \bar{R} \) is a mean value of \( R_{ic} \), that is \( \bar{R} = \frac{1}{n} \sum_{i=1}^{n} R_{ij} \).

In addition, definition 1 shows that \( F_i \) can be approximated as an alternative for \( F_j \).

### 3.3 Correlation calculation

A nonlinear method based on MI is applied in this paper, and the reason is that the high dimensional data usually exist in the form of nonlinear in the real world. The correlation between any pair of variables \( (X, Y) \) can be calculated in the following formulas (6) or (7).

\[
IG(X, Y) = H(X) + H(Y) - H(X, Y) \tag{6}
\]

where \( H(X) \), and \( H(X, Y) \) can be calculated on the basis of formulas (7) and (8).

\[
H(X) = -\sum_{x} p(x) \log_2 p(x) \tag{7}
\]

\[
H(X, Y) = -\sum_{x} p(x) \sum_{y} p(y \mid x) \log_2 p(y \mid x) \tag{8}
\]

To prevent the scale of data is not unified and to reduce the effect of extreme value, each \( IG(X; Y) \) is normalized to the range \([0, 1]\) using formula (9).

\[
R = \frac{2 \times IG(X; Y)}{H(X) + H(Y)} \tag{9}
\]

### 3.4 Performance evaluation

In this paper, classification accuracy and the number of selected features are two indicators used to design the performance evaluation function (Hu et al., 2016; Chuang et al., 2008), which is shown in formula (10).

\[
\text{performance} = w_1 \times \text{Acc} + w_2 \times (1 - \frac{n}{N}) \tag{10}
\]

\( w_1 \) and \( w_2 \) are predefined weight coefficients, which are used to adjust the importance of two indicators in the performance evaluation function. In this study, the values of \( w_1 \) and \( w_2 \) are set to 0.999 and 0.001 respectively. \( \text{Acc} \) is classification accuracy as defined in formula (11). \( n \) is the number of selected features and \( N \) is the total number of features.

\[
\text{Acc} = \frac{C_{num}}{C_{num} + I_{num}} \times 100\% \tag{11}
\]

\( C_{num} \) and \( I_{num} \) are the number of correct and incorrect classification features respectively.

### 4 EXPERIMENTS

#### 4.1 Data description

Five well-known biomedical datasets (Table 2) were used to evaluate the performance of our proposed algorithm. These dataset includes three aspects of disease (cancer) diagnosis, such as gene expression, sera mass-spectrometric etc. The data dimension range was from 2000 to 10000. The first two datasets were taken from the Kent Ridge Biomedical dataset (Li & Liu, 2004), and the last one datasets were taken from the UCI dataset (Asuncion and Newman, 2007).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Attributes</th>
<th>Instances</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ColonTumor</td>
<td>2000</td>
<td>62</td>
<td>2</td>
</tr>
<tr>
<td>DLBCL-Stanford</td>
<td>4026</td>
<td>47</td>
<td>2</td>
</tr>
<tr>
<td>Arcene</td>
<td>100000</td>
<td>200</td>
<td>2</td>
</tr>
</tbody>
</table>

#### 4.2 Experimental procedure

We designed and conducted the following experiments: three kinds of high dimensional biomedical data were compared and analyzed by our proposed algorithm, Relief (a filter method based on the nearest neighbor distance) (Kononenko, 2004), maximum relevance and minimum redundancy (mRmR, a filter method based on MI) (Peng et al., 2005), under the same conditions, respectively. In this experiment, the same conditions refers to: Random forest (RF, numTrees=10) (Zhang et al., 2018) were adopted
as classifier to evaluate classification accuracy respectively.

10 fold cross validation was adopted to evaluate the classification accuracy. Each data set was stratified into 10 folds, of which 9 folds were used as a training sample and the remaining 1 fold constitute a testing sample. The above experiments were implemented in Matlab 2017a.

### 4.3 Results

For the three datasets in Section 4.1, we have conducted the experiments described in Section 4.2. Three main statistical indicators were compared and analyzed in Table 3 which are: (1) Mean (%): the mean of performance, (2) Std: the standard deviation, and (3) MeanFN: the mean number of selected feature.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Algorithm</th>
<th>Mean (%)</th>
<th>Std</th>
<th>MeanFN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ColonTumor</td>
<td>Full Set</td>
<td>78.21</td>
<td>3.51</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>Relief</td>
<td>85.49</td>
<td>3.16</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>mRmR</td>
<td>86.61</td>
<td>3.85</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Our proposed algorithm</td>
<td>92.01</td>
<td>1.97</td>
<td>34</td>
</tr>
<tr>
<td>DLBCL-Stanford</td>
<td>Full Set</td>
<td>76.60</td>
<td>2.87</td>
<td>4026</td>
</tr>
<tr>
<td></td>
<td>Relief</td>
<td>75.48</td>
<td>3.34</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>mRmR</td>
<td>76.58</td>
<td>6.42</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Our proposed algorithm</td>
<td>82.99</td>
<td>2.33</td>
<td>29</td>
</tr>
<tr>
<td>Arcene</td>
<td>Full Set</td>
<td>73.14</td>
<td>4.87</td>
<td>10000</td>
</tr>
<tr>
<td></td>
<td>Relief</td>
<td>81.01</td>
<td>2.06</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>mRmR</td>
<td>78.80</td>
<td>4.82</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>Our proposed algorithm</td>
<td>85.67</td>
<td>2.11</td>
<td>51</td>
</tr>
</tbody>
</table>

*Boldface shows the best experimental result.*

From Table 3, we can observe the following aspects: (1) our algorithm obtained the best Mean among all the three feature selection algorithms. The best Mean(s) are 92.01%, 82.99%, and 85.67%, respectively. In addition, we notice that the maximum Mean improvement of our proposed algorithm was 13.80% compared with the full set. (2) The Std among all the three feature selection algorithms for two out of three experimental results obtained by our proposed algorithm is smaller than other two algorithms. (3) The three feature selection algorithms can effectively reduce the feature dimension, and the dimensionality reduction of our proposed algorithm was the most obvious. In addition, Figure 1 was obtained by statistical analysis of table 3. Figure 1 shows that one average attribute values (avg(Mean)). From comparison results, we can observe that our proposed algorithm were superior to the other two algorithms.

### 5 CONCLUSIONS

In this study, the relationship between two kinds of correlation (the correlation between features and classes, and the correlation between features and features) is established to eliminate redundant features. Due to the determination of completely redundant features is difficult to realize, therefore we analyze four kinds of boundary conditions between $R_{ij}$ and $R_{ik}$ and then a redundancy feature criteria is proposed. On this basis, the approximate redundancy features are defined in this study. Finally, we have proposed a new feature selection algorithm based on redundancy removal for high dimensional data classification.

**ACKNOWLEDGMENT**

This work was supported by the National Basic Research Program of China (973 Program) [2014CB744600]; the National Natural Science Foundation of China [61632014, 61210010, 61461025, 61871259, 61811530325]; Program of Beijing Municipal Science & Technology Commission [Z171100000117005]; the Program of International S&T Cooperation of MOST [2013DFA11140]; the Yong Scholar Fund of Lanzhou Jiaotong University [2016004] and the Teaching and Reform Project of Lanzhou Jiaotong University [JGY201841].

![Figure 1: The average attributes value (avg(Mean))]
REFERENCES


