Feature Selection for Small Sample Sets with High Dimensional Data Using Heuristic Hybrid Approach

M. Biglari*, F. Mirzaei, H. Hassanpour

Computer Engineering and IT Department, Shahrood University of Technology, Shahrood, Iran

PAPER INFO

Abstract

Feature selection can significantly be decisive when analyzing high dimensional data, especially with a small number of samples. Feature extraction methods do not have decent performance in these conditions. With small sample sets and high dimensional data, exploring a large search space and learning from insufficient samples becomes extremely hard. As a result, neural networks and clustering algorithms perform poorly on this kind of data. In this paper, a novel hybrid feature selection technique is proposed, which can reduce drastically the number of features with an acceptable loss of prediction accuracy. The proposed approach operates in multiple stages, starting by removing irrelevant features with a low discrimination power, and then eliminating the ones with low variation range. Afterward, among each set of features with high cross-correlation, a single feature that is strongly correlated with the output is kept. Finally, a Genetic Algorithm with a customized cost function is provided to select a small subset of the remainder of features. To show the effectiveness of the proposed approach, we investigated two challenging case studies with sample set sizes of about 100 and the number of features larger than 1000. The experimental results look promising as they showed a percentage decrease of more than 99% in the number of features, with a prediction accuracy of more than 92%.

doi: 10.5829/ije.2020.33.02b.05

1. INTRODUCTION

One of the challenges in data mining is high dimensional data analysis [1–7]. Having a small sample set adds to the difficulty of the problem. Feature selection can be an effective solution to this problem by removing noisy, irrelevant, and redundant features from a large number of features. Moreover, it is evident that with a smaller number of features, it is easier to avoid overfitting and get a more accurate classifier [1]. However, selecting an appropriate feature selection technique if existed, is not a straight forward task.

When there is a large sample set such that the number of data is larger than the number of features, applying neural networks or multivariable regression analysis can lead to favorable results. The problem begins when there are a small number of data, each of which has a large number of features. Even in some environment, calculating and generating features are financially or timely expensive [8, 9]. Therefore, dimensionality reduction has significant importance. Dimensionality reduction can be managed by two approaches: I) feature selection, and II) feature extraction, which are completely different [2, 10]. Feature selection approaches try to select a subset of relevant or effective features from the original features set. On the other hand, feature extraction approaches, project the original feature space to another feature space with lower dimensionality or a space with better discrimination ability. The new features are usually a linear/nonlinear combination of the
original features. As a result, the analysis of these features will be harder than the original features, because their relevance to the problem statement is not directly assessable [2].

Feature selection methods can be classified into three categories, with respect to utilized learning models [1]: I) Filter-based methods [11–13], II) Wrapper methods [14–20], and III) Embedded Methods [6, 21, 22].

Filter-based methods select features based on statistical measurements that are independent of the learning algorithm and need less computational time. Some examples of these measurement criteria are as follows: Pearson’s correlation [23], information gain [17], Mutual Information (MI) [24, 25], Chi-square test [17], Fisher score, and variance threshold [1].

Wrapper methods wrap around a classifier to utilize it as a cost function to select the best possible subset of features. They use a kind of learning algorithm for testing the quality of the filtered features. As a consequence, their performance is affected by the classifier's accuracy. Furthermore, wrapper methods are more accurate but more computationally expensive than filter-based methods. Recursive feature elimination [19], and evolutionary algorithms are some well-known examples of wrapper methods [1].

Embedded methods employ hybrid learning and ensemble learning algorithms. These methods usually have better accuracy than the previous two categories, since they use a collective decision. Boosting and bagging [26] are examples of embedded methods.

The proposed method is considered as an embedded method since it makes use of some filter-based methods together with genetic algorithm (GA). In this paper, a novel hybrid feature selection approach is proposed. The suggested approach can be applied to small sample sets with high dimensional data where traditional methods are not applicable. Our hybrid approach is made of four stages. Firstly, features with low discrimination ability will be eliminated. Secondly, features with a small variation range will be omitted. Thirdly, among the features with high cross-correlation, all of them except one will be removed. Finally, a customized GA with a novel cost function will be applied to the remaining features, and an acceptable minimum number of features will be selected. Two case studies with a small number of samples and a high number of features are investigated to demonstrate the performance of the proposed method. For comparison purposes, a feed-forward neural network is considered with the initial features set and the reduced features subset. The experimental results indicate the superiority of the proposed method.

The organization of the paper is as follows: the proposed approach is described in section 3. The experimental results are provided in section 4. Finally, the paper is concluded in section 5.

2. THE PROPOSED APPROACH

The proposed approach contains four stages each of which tries to purify/reduce the original features set of length $N$. The goal is to select at most $1 \leq K \leq N$ features. Figure 1 displays the flowchart of the proposed approach.

The four stages of the proposed approach are discussed in the following. Furthermore, a technique is employed to determine a reasonable minimum number of features. The method is discussed in section 2.5.

2.1. Stage 1: Discrimination Ability The features with low discrimination ability are disregarded. By our definition, a feature with a lot of flat areas in its plot across different samples has low discrimination power. Figure 2 depicts a sample plot of four different features and output values over samples. It is evident that the output value is growing across samples. A discriminative feature should change across samples too.

In order to detect the flat areas in a feature plot, the histogram of the feature is calculated. A feature that has non-zero bins count smaller than a threshold $\beta$, is marked as a non-discriminative feature and will be removed. $\beta = floor(\frac{2}{\text{bins count}})$ is an appropriate empirical threshold for the minimum number of non-zero bins. Feature 4 in Figure 2 is an example of a non-discriminative feature that has many flat areas that are detectable in the histogram represented in Figure 3.

![Figure 1. Flowchart of the proposed approach](image-url)
correlation, but the one we were interested in was a linear correlation. If there would be some highly correlated features, there is no need to include them all in the final features subset. So we can select one of them and eliminate the others. In order to filter out this kind of features, the cross-correlation between feature $x_i, i \in \{1, \ldots, N\}$ and other features $x_j, j \in \{1, \ldots, N\} - \{i\}$ is measured, and if the result was greater than a threshold $\tau$, the feature $x_i$ or $x_j$ will be eliminated. $N$ represents the number of features. The employed experimental threshold value was $\tau = 0.99$. Between $x_i$ and $x_j$, the one with higher cross-correlation to the output, will be kept. Algorithm 1 demonstrates the pseudo-code for the proposed method. $\text{xcorr}(x_i, x_j)$ calculates the cross-correlation between vector $x_i$ and $x_j$.

Algorithm 1. The proposed algorithm to remove features with high cross-correlation

$X = \text{feature set for all samples}$

$O = \text{the output vector}$

for $x_i$ in $X$

for $x_j$ in $X - \{x_i\}$

if $\text{xcorr}(x_i, x_j) > \tau$

if $\text{xcorr}(x_i, O) > \text{xcorr}(x_j, O)$

$X = X - [x_j]$

else

$X = X - [x_i]$ end if

end if

end for

end for

2. 4. Stage 4: The Best K Features In the previous stages, a number of features were eliminated. From the remaining ones, $K$ features will be selected. To pick the approximatively best features, a customized GA is proposed. The implemented binary generic algorithm tries to pick at most $K$ features which minimize the proposed cost function $c(\nu)$ as presented in Equation (2).

$$c(\nu) = \begin{cases} 
1 + \text{xcorr}'(\nu) + \text{NRMSE} & \text{if } R^2 = 0 \\
(1 - R^2) + \text{xcorr}'(\nu) & \text{if } R^2 \neq 0 
\end{cases}$$

(2)

where $R^2$ is the coefficient of determination measured, as shown in Equation 3. $R^2$ is employed to compute the accuracy of the estimation by the selected features. $y_i$ and $f_i$ are the $i$th output value and estimated value respectively, and $\bar{y}$ is the average of $y$. The xcorr' measures the average cross-correlation between the selected features (Equation 4). NRMSE is the normalized root mean square error calculated, as stated in Equation (5).

$$R^2 = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y})^2}$$

(3)

$$\text{xcorr}'(\nu) = \left| \frac{\text{xcorr}(\nu_i, \nu_j)}{||\nu||} \right| \quad \forall \nu_i, \nu_j \in \nu$$

(4)
The $R^2$ output is in the range of [0-1]. $R^2 = 0$ means a completely wrong estimation, and $R^2 = 1$ indicates an exact estimation. $\text{xcorr}'$ is in range of [0,1]. $\text{xcorr}'=0$ shows no cross-correlation. In order to estimate how bad a feature set is, NRMSE term will be added to Equation 2 just when $R^2 = 0$. In other circumstances, $R^2$ will be sufficient, because it contains an approximation of NRMSE.

In the proposed GA, a chromosome includes an $N$-dimensional vector of boolean values which determines whether a feature is selected or not. The goal of the GA is to pick at most $K$ features, so a chromosome cannot have more than $K$ ones. If a newly generated chromosome has more than $K$ ones ($L$), $L-K$ values are randomly chosen and set to zero. The flowchart of the proposed GA is provided in Figure 4.

K-fold cross-validation is employed for the cost function calculation in the proposed GA as the following: I) The sample set is divided into $K$ folds, II) The cost function is evaluated $K$ times each of which utilizes $K$ folds for training and 1 fold for testing, and III) The results are averaged over $K$ as the final cost function value. The parameters configuration employed in the proposed GA is demonstrated in Table 1.

### 2.5. Optimal Minimum Number of Features

A technique is recommended to select a reasonable minimum number of features [28]. This technique divides the sample set into two training, and test sets; and then defines three criteria: training estimation accuracy (TEA), testing estimation accuracy (TAR), and training error (TE). Practically, this technique wraps around the proposed GA which, is called for different values of $K$ starting from 1 to $N$ (the number of features). In each iteration, the three criteria are evaluated and plotted, until these three lines remain almost parallel to the X-axis. TEA and TAR are calculated using $R^2$ measure on the training and test set, respectively. TE is computed by NRMSE measure on the training set.

Figure 5 displays an example with values of $K = \{1,2,\ldots,7\}$. For each value of $K$, the GA is called, and the criteria are measured and plotted, until $K = 4$ point, all the lines are approximately parallel to the X-axis. So $K = 4$ is picked as the optimal minimum number of features.

### 3. THE EXPERIMENTAL RESULTS

We have been provided two chemical datasets by Nekoei et al. [28] that were suited to be analyzed by the proposed approach. Both of these datasets have a low sample size with high-dimensional data. In the following, the two case studies based on these datasets are discussed in detail. Table 2 presents the proposed algorithm configuration used for both case studies.

![Figure 4. The flowchart of the proposed genetic algorithm. The flow order of the algorithm is presented with numbers from 1 to 5](image)

![Figure 5. An example demonstrating the three proposed criteria for finding the optimal minimum number of features [28]](image)

### TABLE 1. The proposed GA parameters configuration

| Parameter                        | Value  |
|----------------------------------|--------|
| Population size                  | 10000  |
| Termination criterion            | 1000 epochs |
| Crossover probability            | 0.7    |
| Mutation probability             | 0.2    |
| Tournament size                  | 3      |

\[ \text{NRMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \] (5)
TABLE 2. The proposed algorithm parameters configuration

| Parameter | Value           |
|-----------|----------------|
| \( \beta \) | 20 from 50 bins |
| \( \delta \) | 0.15            |
| \( \tau \)  | 0.99            |

3.1. Chemical Molecules Case Study

The first case study is focused on a chemical molecules dataset, which contains 81 molecules with 1056 physicochemical properties or theoretical molecular descriptors (Figure 6). Every molecule has a response value measured based on the descriptors. The goal is to find a linear QSAR-based model to predict the response variable with a subset of features. The descriptors and response values are all numerical and their numerical values may not be in the same range.

Figure 7 shows the minimum number of descriptors suggested by the proposed technique. The average values of \( R^2 \) measure over training and test sets for different numbers of selected features and the selected features itself are summarized in Table 3.

![Figure 6](image_url)  
Figure 6. The chemical molecules dataset with 81 samples (m), 1056 features (a), and a response value (f)

![Figure 7](image_url)  
Figure 7. The optimal number of descriptors in chemical molecules dataset

It is evident from Table 3 that by selecting more than four features, there will be slight variations in \( R^2 \) response values. Therefore, the suggested optimal minimum number of features is four. It is worth noting that the feature number 188 must have a significant contribution to the linear model, as it is selected in all six suggested features set.

The linear model found by the proposed GA for \( K = 4 \) using multiple linear regression (MLR) is given by Equation (6). The model was used to predict the response variable, and the average result measured by K-fold cross-validation is compared with a feed-forward neural network (NN) once trained with the initial features set, and once trained with the reduced features set. The comparison results are presented in Table 4. Additionally, the regression plot is demonstrated in Figure 8.

\[
f = -0.403 (V_{EA1}) - 0.454 (BElv1) + 0.022 (GATS2e) + 0.880 (H^5p) + 1 \tag{6}
\]

3.2. Chemical Drugs Case Study

This case study includes a chemical drugs dataset with 103 samples of 1482 dimensional data. Like the previous dataset, each sample has a response value. The goal is to build a linear model to predict the response variable employing just a subset of descriptors.

| Num of features | Selected feature indices | Average \( R^2 \) |
|-----------------|--------------------------|-------------------|
| 1               | [188]                    | 0.880             |
| 2               | [75,188]                 | 0.919             |
| 3               | [188,413,682]            | 0.928             |
| 4               | [188,281,413,936] [V_{EA1}, BElv1, GATS2e, H^5p] | 0.942             |
| 5               | [188,413,684,778,990]    | 0.945             |
| 6               | [188,384,412,650,684,741] | 0.946             |

3.2. Chemical Drugs Case Study

This case study includes a chemical drugs dataset with 103 samples of 1482 dimensional data. Like the previous dataset, each sample has a response value. The goal is to build a linear model to predict the response variable employing just a subset of descriptors.

| Set type       | Method       | \( R^2 \) | NRMSE |
|----------------|--------------|-----------|-------|
| Training       | Proposed     | 0.96      | 0.04  |
| Test           | Proposed     | 0.92      | 0.05  |
| Training       | NN + reduced features | 0.86 | 0.07 |
| Test           | NN + reduced features | 0.58 | 0.11 |
| Training       | NN           | 0.70      | 0.10  |
| Test           | NN           | 0.33      | 0.17  |
The optimal minimum number of features suggested for this dataset, as depicted in Figure 9 is six. Moreover, the average values of $R^2$ term over training and test sets for 1 to 6 selected features are presented in Table 5.

The linear model suggested by the proposed GA for $K = 6$ using the MLR is given by Equation (7). Similar to the previous section, the average prediction accuracy and error of the model is compared with NN, and the results are provided in Table 6. Also, the regression plot is shown in Figure 10.

\[
f = 5.333 \left( T(O..F) \right) + 0.021(GGI7) + 0.700(JG7) - 12.126 (MATS7v) - 37.334 (PCWTe) - 0.234 (RDF065m)
\]

\(7\)

3.3. Discussion Applying linear regression on the initial high dimensional data leads to poor results because there are a lot of noisy and irrelevant features. NN operates a little better by acquiring the average accuracy
of around 70% on training data (Tables 4 and 6). Although, due to the small number of samples, NN overfits and thereupon, shows a sudden accuracy decrease over the test data. By reducing the number of input features drastically, the performance of NN grows significantly on both datasets. Still, the best result is earned by the proposed approach, which utilizes multiple linear regression internally. As can be seen in Tables 4 and 6, the accuracy of the proposed method on the training and test sets are very close together. It indicates that overfitting has not happened when training the model, and the built model is robust and accurate.

4. CONCLUSION

In this work, a heuristic hybrid approach for feature selection is proposed. The approach reduces the number of features significantly in four consecutive stages. In the early stages, some of the irrelevant and less discriminative features are omitted. In the final stages, the approximately best feature subset of length $K$ is chosen by a GA which uses a customized cost function. The proposed cost function maximizes the prediction accuracy and minimizes the prediction error, and the cross-correlation between the selected features subset simultaneously.

Two case studies with high-dimensional data were analyzed to indicate the performance of the proposed approach. Firstly, the proposed method was applied to a chemical molecules dataset and reduced the number of features from 1056 to 4 with a prediction accuracy of $R^2 = 0.92$. Secondly, a similar configuration is used for the next dataset that led to the reduction of 99.6 percent of the features with a prediction accuracy of $R^2 = 0.93$. The experimental results indicate that the proposed method is better suited to be used for small sample sets with high dimensional data than neural networks. Additionally, our approach can be employed as a preprocessing step in other methods. As we demonstrated the performance boost of NN when injected with our reduced features set.

5. REFERENCES

1. Venkatesh, B., Anuradha, J., ‘A Review of Feature Selection and Its Methods’, Cybernetics and Information Technologies, Vol. 19, No. 1, (2019), 3–26.
2. Li, J., Cheng, K., Wang, S., et al., ‘Feature Selection: A Data Perspective’, ACM Computing Surveys, Vol. 50, No. 6, (2018), 94.
3. Cai, J., Luo, J., Wang, S., Yang, S., ‘Feature selection in machine learning: A new perspective’, Neurocomputing, Vol. 0, (2018), 1–10.
4. Li, J., Liu, H., Science, C., ‘Challenges of Feature Selection for Big Data’, IEEE Intelligent Systems, Vol. 32, No. 2, (2017), 9–15.
5. Jovi, A., Brki, K., Bogunovi, N., ‘A review of feature selection methods with applications’, in ‘38th International Convention on Information and Communication Technology, Electronics and Microelectronics’ (2015), 1200–1205.
6. Asir, D., Appavu, S., Jebamalar, E., ‘Literature Review on Feature Selection Methods for High-Dimensional Data’, International Journal of Computer Applications, Vol. 136, No. 1, (2016), 9–17.
7. Mazimpaka, J.D., Timpf, S., ‘Trajectory data mining: A review of methods and applications’, Journal of Spatial Information Science, Vol. 13, (2016), 61–99.
8. Hamidi, H., Daraei, A., ‘Analysis of Pre-processing and Post-processing Methods and Using Data Mining to Diagnose Heart Diseases’, International Journal of Engineering-Transactions A: Basics, Vol. 29, No. 7, (2016), 921–930.
9. Kumar, S., Sahoo, G., ‘A Random Forest Classifier based on Genetic Algorithm for Cardiovascular Diseases Diagnosis’, International Journal of Engineering-Transactions B: Applications, Vol. 30, No. 11, (2017), 1723–1729.
10. Liu, H., Motoda, H., ‘Computational Methods of Feature
feature selection for small sample sets with high dimensional data using heuristic hybrid approach

M. Biglari, F. Mirzaei, H. Hassanpour

Computer Engineering and IT Department, Shahrood University of Technology, Shahrood, Iran

Paper history:
Received 28 August 2019
Received in revised form 16 January 2020
Accepted 16 January 2020

Keywords:
Feature Selection
Data Mining
Regression
High Dimensional Data
Evolutionary Methods

Abstract:

In this paper, we propose a new hybrid feature selection method for small sample sets with high dimensional data. The proposed method combines recursive feature elimination (RFE) with a genetic algorithm (GA). The RFE method is used to initial screening of the data and GA is used to select the best subset of features. The proposed method is applied to two practical problems and the results show that the proposed method outperforms the compared methods.

Introduction

Feature selection is a crucial step in data mining and machine learning tasks. It is used to reduce the number of features in a dataset, which can improve the performance of machine learning algorithms and reduce the computational complexity. However, feature selection becomes more challenging when dealing with high dimensional data and small sample sizes. In this paper, we propose a new hybrid feature selection method for small sample sets with high dimensional data. The proposed method combines recursive feature elimination (RFE) with a genetic algorithm (GA).

Methodology

The proposed method consists of two main steps: initial screening and feature selection.

1. Initial Screening:

In the initial screening step, the recursive feature elimination (RFE) method is used to remove features that have low correlation with the target variable. The RFE method recursively eliminates features based on a linear regression model.

2. Feature Selection:

In the feature selection step, the genetic algorithm (GA) is used to select the best subset of features. The GA is a population-based optimization algorithm that mimics the process of natural selection. The GA starts with a population of candidate solutions (initial feature subsets), and iteratively selects the best solutions (feature subsets) based on fitness functions (evaluation criteria). The GA process is repeated until a stopping criterion is met (e.g., a fixed number of generations).

Results

The proposed method is applied to two practical problems: predicting polycythemia and forecasting polyclausal B-lymphocytes. The results show that the proposed method outperforms the compared methods in terms of classification accuracy and computational complexity.

Conclusion

In this paper, we proposed a new hybrid feature selection method for small sample sets with high dimensional data. The proposed method combines recursive feature elimination (RFE) with a genetic algorithm (GA). The results show that the proposed method outperforms the compared methods in terms of classification accuracy and computational complexity. Therefore, the proposed method can be used as a powerful tool for feature selection in practical applications.