International Journal of Engineering

Journal Homepage: www.ije.ir

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

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

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

[*] Corresponding Author Email: *mbt925@shahroodut.ac.ir* (M. Biglari)

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 β, is marked as a non-discriminative feature and will be removed. $\beta = floor(\frac{2}{5}\, bins\ count)$ 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.

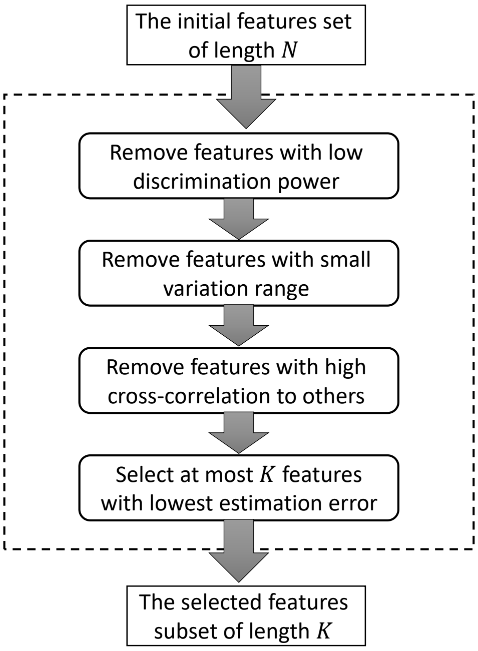

The initial features set of length N
Remove features with low discrimination power
Remove features with small variation range
Remove features with high cross-correlation to others
Select at most K features with lowest estimation error
The selected features subset of length K

Figure 1. Flowchart of the proposed approach

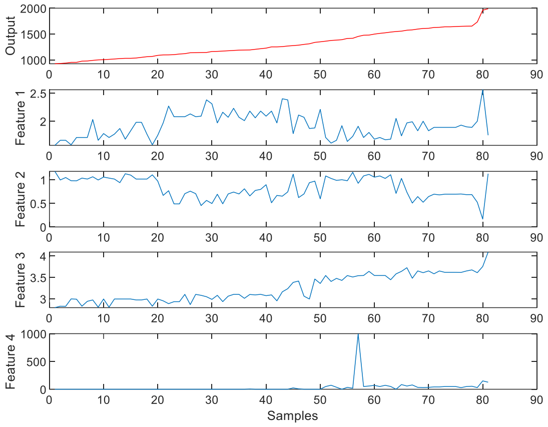


Figure 2. A sample plot of features and the output value across all samples

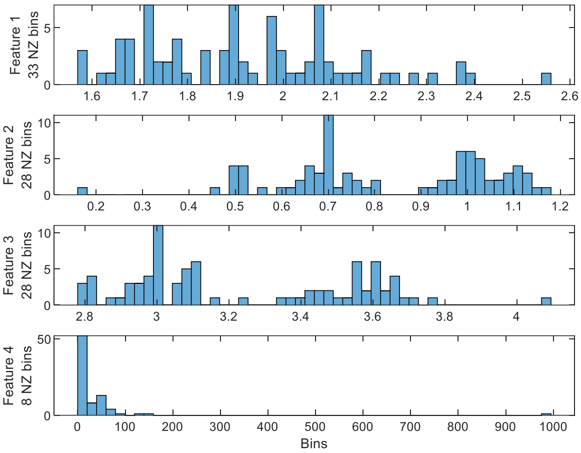


Figure 3. The histogram of features provided in Figure 2. For a better representation, 50 bins are considered for all features. *NZ* means Non-Zero

2. 2. Stage 2: Variation Range

To compute features with a small variation range, the first step is to normalize the features set. Min-Max normalization is used, as stated in Equation (1).

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \qquad (1)$$

where $x_i = (x_1, x_2, \dots, x_M)$ is the i$^{\text{th}}$ feature, M is the number of samples, and z_i is the i$^{\text{th}}$ normalized feature.

In the second step, the features with a standard deviation smaller than a threshold δ will be omitted. We utilized $\delta = 0.15$ as a good empirical threshold.

2. 3. Stage 3: Cross-correlation

The relevancy of a feature is measured based on the characteristics of the data, not by its value. There are some statistical measures to show the relations between the features [1, 27]. Usually, there are some features that have a high correlation with each other. There are different types of 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 $x_i, i \in \{1, \dots, N\}$ and other features $x_j, j \in \{1, \dots, N\} - \{i\}$ is measured, and if the result was greater than a threshold τ, 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. 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

```
$X$ = feature set for all samples
$O$ = the output vector
for $x_i$ in $X$
    for $x_j$ in $X - \{x_i\}$
        if xcorr($x_i$,$x_j$) > $\tau$
            if xcorr($x_i$,$O$) > 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 approximately 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 $c(v)$ as presented in Equation (2).

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

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$^{\text{th}}$ 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).

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad (3)$$

$$xcorr'(v) = \left| \frac{xcorr(v_i, v_j)}{|v|} \right| \ \forall\, v_i, v_j \in v \qquad (4)$$

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_i (y_i - f_i)^2}}{\bar{y}} \tag{5}$$

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. xcorr' is in range of [0,1]. 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-1 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 $K = \{1,2,\dots,7\}$. For each value of K, the GA is called, and the criteria are measured and plotted. After $K = 4$ point, all the lines are approximately parallel to the X-axis. So $K = 4$ is picked as the optimal minimum number of features.

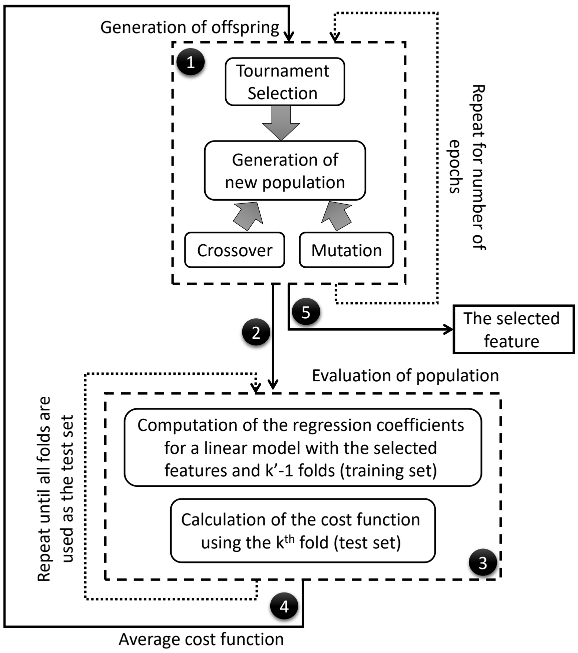

Generation of offspring
1
Tournament Selection
Generation of new population
Crossover
Mutation
Repeat for number of epochs
5
The selected feature
2
Evaluation of population
Computation of the regression coefficients for a linear model with the selected features and k'-1 folds (training set)
Calculation of the cost function using the kth fold (test set)
3
Repeat until all folds are used as the test set
4
Average cost function

Figure 4. The flowchart of the proposed genetic algorithm. The flow order of the algorithm is presented with numbers from 1 to 5

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

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.

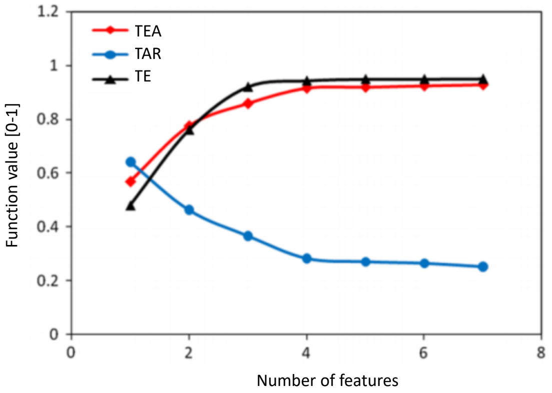


Figure 5. An example demonstrating the three proposed criteria for finding the optimal minimum number of features [28]

TABLE 2. The proposed algorithm parameters configuration

Parameter	Value
β	20 from 50 bins
δ	0.15
τ	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.

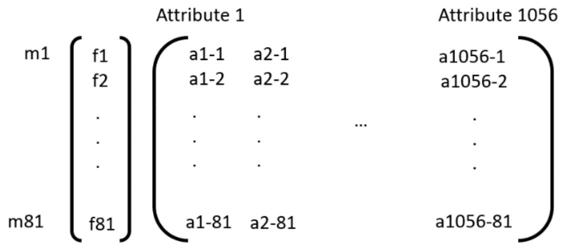


Figure 6. The chemical molecules dataset with 81 samples (m), 1056 features (a), and a response value (f)

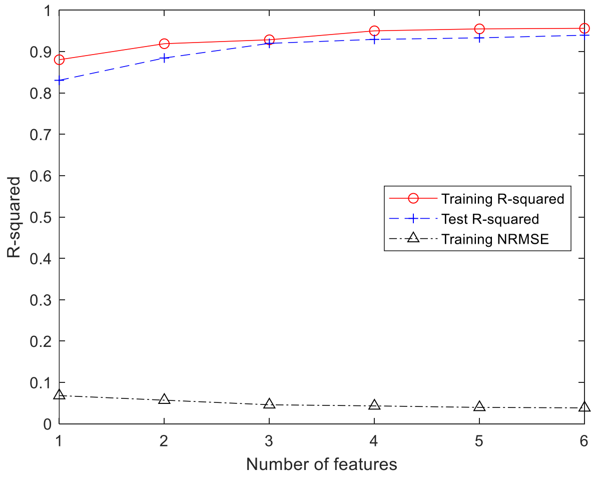


Figure 7. The optimal number of descriptors in chemical molecules dataset

TABLE 3. The average values of R^2 measure, and the selected descriptors indices for different numbers of descriptors of chemical molecules dataset. The features name are displayed for the selected descriptors (fourth row)

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] [$VEA1$, BELv1, GATS2e, H5p]	0.942
5	[188,413,684,778,990]	0.945
6	[188,384,412,650,684,741]	0.946

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.

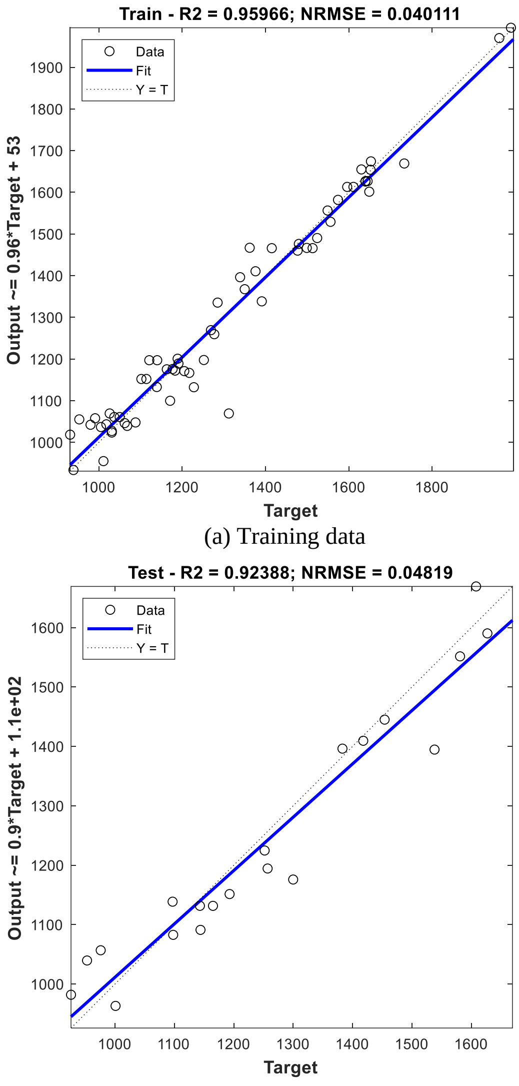
$$f = -0.403\,(\boldsymbol{VEA1}) - 0.454\,(\mathbf{BELv}) + 0.022\,(\mathbf{GATS2e}) + 0.880\,(\mathbf{H5p}) + 1 \qquad (6)$$

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.

TABLE 4. The average R^2 and NRMSE measures of the built model of the chemical molecules case study using K-fold cross-validation, compared against NN

Set type	Method	R^2	NRMSE
Training	Proposed	**0.96**	**0.04**
Test		**0.92**	**0.05**
Training	NN + reduced features	0.86	0.07
Test		0.58	0.11
Training	NN	0.70	0.10
Test		0.33	0.17



(a) Training data

(b) Test data

Figure 8. The regression plot of the built model of chemical molecules case study on (a) training and (b) test sets

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.

$$f = 5.333\,\big(T(O..F)\big) + 0.021(\mathrm{GGI7}) + 0.700\,(\mathrm{JGT}) - 12.126\,(\mathrm{MATS7v}) - 37.334\,(\mathrm{PCWTe}) - 0.234\,(\mathrm{RDF065m}) \qquad (7)$$

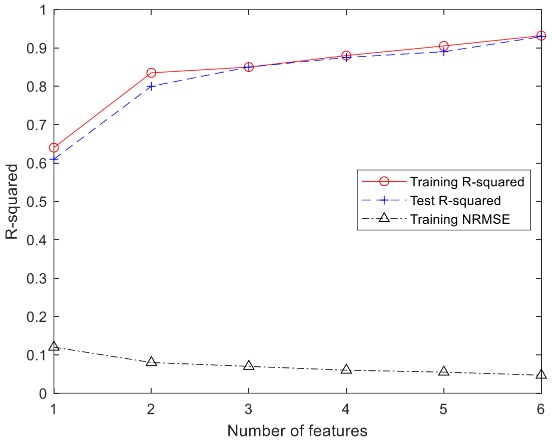


Figure 9. The optimal number of descriptors in chemical drugs dataset

TABLE 5. The average values of R^2 measure, and the selected descriptors indices for different numbers of descriptors of chemical drugs dataset. The features' names are displayed for the selected descriptors (sixth row).

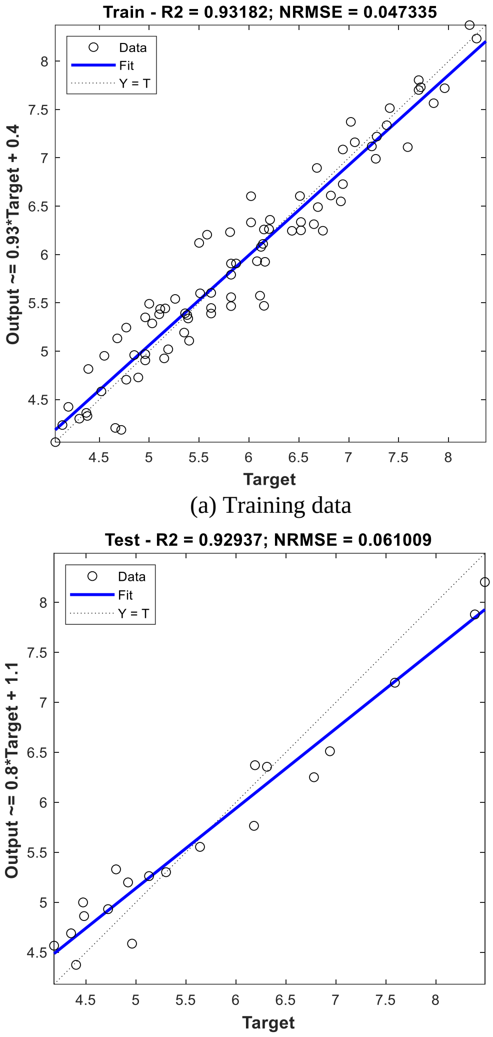
Num of features	Selected feature indices	Average R^2
1	[103]	0.633
2	[61,550]	0.825
3	[136,569,892]	0.849
4	[415,524,788,1220]	0.877
5	[415,524,788,901,1355]	0.898
6	[285,401,415,462,524,670] [T(O..F), GGI7, JGT, MATS7v, MATS7v, PCWTe, RDF065m]	0.930

TABLE 6. The average R^2 and NRMSE measures of the built model of the chemical drug case study using K-fold cross-validation, compared against NN

Set type	Method	R^2	NRMSE
Training	Proposed	**0.93**	**0.05**
Test		**0.93**	**0.06**
Training	NN + reduced features	0.87	0.07
Test		0.65	0.10
Training	NN	0.73	0.1
Test		0.40	0.16

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



(a) Training data

(b) Test data

Figure 10. The regression plot of the built model of chemical drugs case study on (a) training and (b) test sets

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 pre-processing 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

Selection', (CRC Press, 2007)

11. Liu, H., Setiono, R., 'A Probabilistic Approach to Feature Selection - A Filter Solution', , in 'Proceedings of 13th International Conference on Machine Learning' (1996), 319–327
12. Fayyad, U.M., Irani, K.B., 'The attribute selection problem in decision tree generation', , in 'AAAI' (1992), 104–110
13. Kalpana, P., Mani, K., 'A New Hybrid Framework for Filter based Feature Selection using Information Gain and Symmetric Uncertainty', ***International Journal of Engineering-Transactions B: Applications***, Vol. 30, No. 5, (2017), 659–667.
14. Ch, V., Asvestas, P.A., Delibasis, K.K., Matsopoulos, G.K., 'A classification system based on a new wrapper feature selection algorithm for the diagnosis of primary and secondary polycythemia', ***Computers in Biology and Medicine***, Vol. 43, (2013), 2118–2126.
15. Kohavi, R., John, H., 'Wrappers for feature subset selection', ***Artificial Intelligence***, Vol. 97, (1997), 273–324.
16. Dy, J.G., Brodley, C.E., 'Feature Subset Selection and Order Identification for Unsupervised Learning', , in 'Proceedings of 17th International Conference of Machine Learning' (2000), 247–254
17. Yang, Y., Pedersen, J.O., 'A Comparative Study on Feature Selection in Text Categorization', , in 'Proceedings of 14th International Conference on Machine Learning' (1997), 412–420
18. Mohsenzadeh, Y., Sheikhzadeh, H., Member, S., Reza, A.M., Member, S., Kalayeh, M.M., 'The Relevance Sample-Feature Machine : A Sparse Bayesian Learning Approach to Joint Feature-Sample Selection', ***IEEE Transactions on Cybernetics***, Vol. 43, No. 6, (2013), 2241–2254.
19. Yan, K., Zhang, D., 'Feature Selection and Analysis on Correlated Gas Sensor Data with Recursive Feature Elimination', ***Sensors & Actuators: B. Chemical***, Vol. 212, (2015), 353–363.
20. Jain, A., Zongker, D., 'Feature Selection Evaluation, Application, and Small Sample Performance.pdf', ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 19, No. 2, (1997), 153–158.
21. Talavera, L., 'Feature Selection as a Preprocessing Step for Hierarchical Clustering', , in 'Proceedings of 25th International Conference of Machine Learning' (1999), 389–397
22. Das, S., 'Filters , Wrappers and a Boosting-Based Hybrid for Feature Selection', , in 'Engineering' (2001), 74–81
23. Biesiada, J., Duch, W., 'Feature Selection for High-Dimensional Data – A Pearson Redundancy Based Filter', , in 'Advances in Soft Computing' (2007), 242–249
24. Estévez, P.A., Member, S., Tesmer, M., Perez, C.A., Member, S., Zurada, J.M., 'Normalized Mutual Information Feature Selection', ***IEEE Transactions on Neural Networks***, Vol. 20, No. 2, (2009), 189–201.
25. Vinh, L.T., Thang, N.D., Lee, Y., 'An Improved Maximum Relevance and Minimum Redundancy Feature Selection Algorithm Based on Normalized Mutual Information', , in 'Proceedings of 10th IEEE/IPSJ International Symposium on Applications and the Internet' (2010), 395–398
26. Quinlan, J.R., 'Bagging, Boosting, and C4.5', ***AAAI/IAAI***, Vol. 1, (2006), 725–730.
27. Gheyas, I.A., Smith, L.S., 'Feature Subset Selection in Large Dimensionality Domains', ***Pattern Recognition***, Vol. 43, No. 1, (2010), 5–13.
28. Nekoei, M., Mohammadhosseini, M., Pourbasheer, E., 'QSAR study of VEGFR-2 inhibitors by using genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm-support vector machine (GA-SVM): A comparative approach', ***Medicinal Chemistry Research***, Vol. 24, No. 7, (2015), 3037–3046.

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

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

چکیده

در مواجهه با داده‌های با ابعاد بالا، به ویژه اگر تعداد نمونه‌ها کم باشد، انتخاب ویژگی از اهمیت بالایی برخوردار است. روش‌های استخراج ویژگی در چنین شرایطی، عملکرد قابل قبولی نخواهند داشت. با وجود تعداد نمونه‌های کم و تعداد ویژگی‌های زیاد، کاوش فضای جستجوی بزرگ دشوار شده و یادگیری با تعداد نمونه‌ها کم، به سادگی انجام‌پذیر نیست. از این رو، شبکه‌های عصبی و روش‌های دسته‌بندی بر روی این نوع داده‌ها عملکرد ضعیفی دارند. در این مقاله، یک روش انتخاب ویژگی ترکیبی جدید برای کاهش شدید تعداد ویژگی‌ها ارائه شده که تنها منجر به کاهش جزئی دقت پیش‌بینی خروجی مورد انتظار خواهد شد. روش پیشنهادی در چند مرحله عمل می‌کند. در ابتدا ویژگی‌های نامرتبط با قدرت تمایز کم حذف شده و سپس ویژگی‌هایی که دارای بازه تغییرات محدود هستند، خط می‌خورند. در مرحله بعد، از بین هر مجموعه ویژگی‌ها با همبستگی بالا، تنها یک ویژگی که داری بیشترین همبستگی با مقدار خروجی است، نگه داشته می‌شود. در گام نهایی، یک الگوریتم ژنتیک با تابع هزینه سفارشی به ویژگی‌های باقیمانده اعمال شده تا کوچکترین مجموعه ویژگی‌های کارا را برگزیند. برای نمایش کارایی روش پیشنهادی، دو مسئله متفاوت با تعداد تقریبا ۱۰۰ نمونه و تعداد ویژگی‌های بیشتر از ۱۰۰۰، مورد مطالعه قرار گرفته‌اند. نتایج آزمایشات انجام شده، کاهش بیش از ۹۹ درصدی تعداد ویژگی‌ها را نشان می‌دهد، در حالی که دقت پیش‌بینی خروجی بیش از ۹۲ درصد می‌باشد.

doi*: 10.5829/ije.2020.33.02b.05*