Journal of Artificial Intelligence Practice (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2022.050103
ISSN 2371-8412 Vol. 5 Num. 1

A Novel Classification Model based on Ensemble Feature Selection and Hyper-parameter Optimation

Liguo Zhang[1,a,*], Jian Zhang[1,b] and Yan Wang[2,c]

[1]*College of Information Science & Technology, Agricultural University of Hebei, Hebei, Baoding, China*
[2]*Dept of Computer, North China Electric Power University, Hebei, Baoding, China*
[a]*zhangliguo2006@126.com*, [b]*xxzj@hebau.edu.cn*, [c]*Wangyan1206@126.com*
**corresponding author*

Keywords: Feature Selection, PCA, HPO, Recursive Feature Elimination, Random Forest Classification.

Abstract: Background and purpose: the central problem in machine learning is identifying a representative set of features from which to construct a classification model and setting the best Hyper-Parameters of the model for a particular task. Materials and Methods: Feature selection aims to reduce the dimensionality of patterns for classificatory analysis by selecting the most informative instead of irrelevant and/or redundant features. Here, RFE is used for selecting the most useful predictive features, PCA extract the characteristics of all the original variables. And also the Grid Search is used for model HPO. Experiments are applied on Breast cancer dataset. Findings and Conclusion: The experiments show that, the number of features is reduced from 30 to 15 features and the classification accuracy increase from 95.6% to 97.02%. And the proposed Feature selection and hyper-parameter optimation method can be used to other pattern classification problems.

1. Introduction

Nowadays in many pattern classification problems it is not uncommon that we are confronted with a very high dimensional variable space, with hundreds to tens of thousands of attributes or features. In this case, not only the “curse of dimensionality” caused by a very high ratio of number of features to number of data samples is problematic but also the irrelevant features in the feature set that undermine the performance of a given learning algorithm.

In machine learning Feature Selection (FS) is known as variable selection, is the process of selecting a subset of relevant features for use in model construction [1-5]. Feature subset selection is a useful way for reducing dimensionality, removing irrelevant data, increasing learning accuracy. Numerous feature subset selection methods have been planned and studied for machine learning applications.

However, different Feature Selection algorithms use different criteria to select representative features, making it difficult to find the best algorithm for different domain datasets [6]. Both Feature Selection algorithm and Feature Extraction algorithm, in the paper, are used to reduce dimension, which overcomes the limitations of single feature selection methods. And also, Grid Search

technology is used for model Hyper-parameter Optimation (HPO). Based on mentioned above, a classification model was built. At last, Breast cancer dataset is used to validate the proposed model.

2. Data preprocessiong

2.1. Feature dimensionality reduction

Selecting the appropriate features to achieve the best result in data classification has been one of the most challenging topics in recent decades. Data reduction is a preprocessing step for classification. It aims to improve the classification performance through the removal of redundant features. Data reduction can be achieved by Recursive feature elimination and Principal component analysis.

2.1.1. Recursive Feature Elimination

RFE seeks to improve generalization performance by removing the least important features whose deletion will have the least effect on training errors [7-9]. In addition, RFE is closely related to support vector machines (SVMs) which have been shown to generalize well even for small sample classification. While it has shown great promise in small-sample feature selection problems, RFE tends to remove redundant and weak features and retains independent features. As pointed out in [8], (1) presumably redundant features may provide better class separation, and (2) two weak features that are useless by themselves can provide a significant performance improvement when used together. Thus, simply removing redundant or weak features may degrade classification performance. In the Recursive Feature Elimination with Cross-Validation (RFECV), the “step” parameter indicates the number of features to remove at each iteration. Here, using feature_importances_attribute of Random Forest Classifier to calculate the significance of the set of features in the iteration

2.1.2. PCA

The main idea of PCA is to map the original features to the new space, and the features in the new space will be expressed as the linear combination of the original features [10]. The premise of using PCA is that there should be a strong linear correlation between the variables of the original data. If the linear correlation between the original variables is very small and there is no simplified data structure between them, principal component analysis is actually meaningless. Therefore, when applying principal component analysis, its applicability should be statistically tested first.

KMO statistic compares the sample correlation coefficient with the sample partial correlation coefficient. It is used to test whether the sample is suitable for principal component analysis. The value of KMO is between 0 and 1. The larger the value, the more suitable the sample data is for principal component analysis and factor analysis. It is generally required that the value is greater than 0.5 before principal component analysis or correlation analysis. KMO calculation formula is as follows:

$$\mathrm{KMO} = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} \alpha_{ij}^2} \tag{1}$$

Here, r_{ij} denotes Simple correlation coefficient, $\alpha^2_{ij.1,2,3,\ldots,k}$ denotes Partial correlation coefficient.

The main idea of PCA is reduce the dimensionality of a dataset, while preserving as much ‘variability’ (i.e. statistical information) as possible, which means ‘preserving as much variability as possible’ translates into finding new variables that are linear functions of those in the original dataset, that successively maximize variance and that are uncorrelated with each other.

Given $X = (X_1, X_2, \cdots, X_5)$ is a set of observable variables, the dimensionality of the dataset can be reduced to two main components. The fig.1, formula 2 and formula 3 can show the principle of PCA.

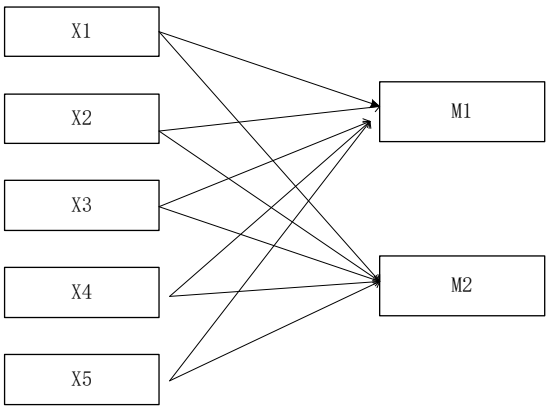

Figure 1: Principle of PCA

$$M1 = a1 * X1 + a2 * X2 + a3 * X3 + a4 * X4 + a5 * X5 \quad (2)$$

$$M2 = b1 * X1 + b2 * X2 + b3 * X3 + b4 * X4 + b5 * X5 \quad (3)$$

2.2. Hyper parameter optimization

Generally, there are two kinds of parameters in the learner model. One kind can be learned and estimated from the data, which is called parameter. There is another type of parameter that cannot be estimated from the data and can only be designed and specified by human experience. We call it hyper parameter. A hyper-parameter is a parameter whose value is set before starting the learning process. On the contrary, the values of other parameters are obtained through training.

2.2.1. Grid Search

Grid Search is a kind of Exhaustive search to adjust parameters. Among all the candidate parameters, through the loop traversal, try every possibility. The best performance parameter is the final result. It works like finding the maximum value in an array. The main disadvantage of this method is that it is time-consuming.

So grid search works for three or four （ Or less ） Super parameter （ When the number of super parameters increases, The computational complexity of grid search will increase exponentially , In this case, random search is used ）, The user lists a small super parameter value field, The Cartesian product of these super parameters （ Permutation and combination ） For a set of super parameters. The grid search algorithm uses each group of super parameter to train the model and selects the super parameter combination with the least error of the verification set.

2.2.2. Cross-Validation

Cross-validation is one of the most widely used data resampling methods to estimate the true prediction error of models and to tune model parameters. There are two common method Leave-one-out cross-validation and K-fold Cross Validation. In this paper, we only introduce K-fold Cross Validation. The main steps of K-fold Cross Validation are as follows:

1) Split dataset s into K equal subsets s_k
2) Do K times:

a) Estimate the model $\beta_{(k)}$ on $s_k = \{s_1, s_{k-1}, s_{k+1}, \dots, s_k\}$
b) Compute the prediction error PE(k) between the test sample s_k and the predicted model by $\beta_{(k)}$

3) Compute the average of those K prediction errors as the overall estimate of prediction error

$$\mathrm{CV} = \frac{1}{K}\sum_{k=1}^{K} PE(k) \tag{4}$$

2.3. Train Flow

During the process of model train, both filter methods (chi-square) and wrapper approaches (Recursive Feature Elimination) are used to features select, namely reducing dimensionality and removing irrelevant data. And also PCA is use to extract the overall features of all original data. Then, combine the obtained features to train Random Forest Classifier. At last, the parameters of each algorithm are optimized by grid search to obtain the optimal model. The whole training process is shown in Figure 2.

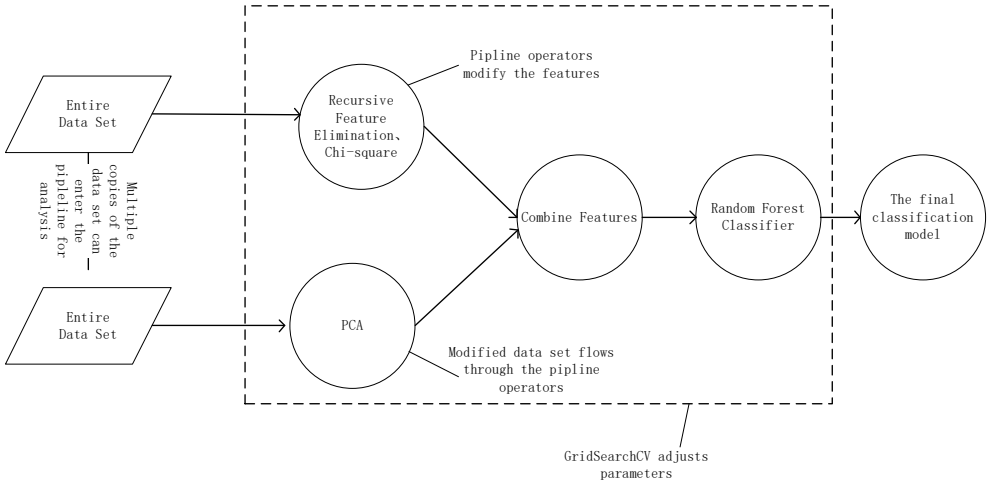

Figure 2: Training process

3. Experiment

In this paper, Breast cancer dataset is used to validate the proposed model. Breast cancer dataset has 569 records with 30 independent variables (features) and binary classes for the dependent variables [9]. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

3.1. REFCV feature select

In REFCV, feature importance is calculated based on the estimator selected, and one/few features are dropped in each iteration. Here, we drop one feature in each iteration to identify the right set of features which have maximum influences on the dependent variable. To split data into train/test sets and calculate the feature importance we have mentioned Five-fold cross-validation with Stratified K-Folds cross-validator. Based on feature importance, RFECV recommended 15 features are significant in predicting the class of cancer. The figure 3 shows that for the combination of 15 most important features may get the cross-validation scores tops. Increasing the dimension of the training dataset further doesn't improve the prediction accuracy further.

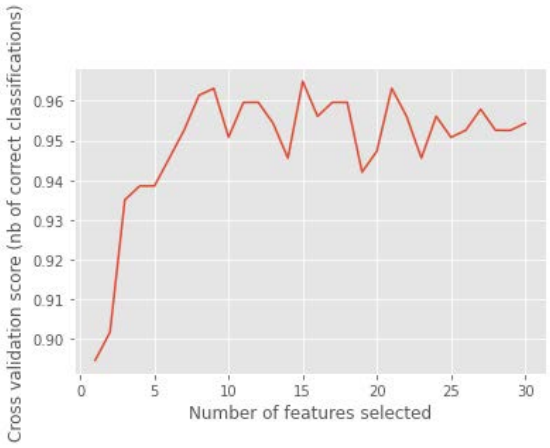

Figure 3: Number of features and cross-validation scores

3.2. PCA reducing dimensionality

3.2.1. The Necessity of PCA

With calculating the correlation coefficient matrix of the original 30 eigenvectors, it is found that there is a great degree of correlation between eigenvectors, and many values are more than 0.98; Kaiser-Meyer-Olkin (KMO) value is 0.8317, and according to the criteria (KMO>0.6) given by statistician Kaiser, the original eigenvectors are suitable for FA. Meanwhile, the corresponding Bartlett sphericity test value is 0.000 (far less than the significance level of 0.05), so the null hypothesis should be refused, which means it is necessary of FA.

3.2.2. Confirm the Number of components

Here, use PCA function to calculate the relationship between Number of components and cumulative explained variance ratio, shown in figure 4. As figure shows, only two main components could explain 99.75% variance of the original data.

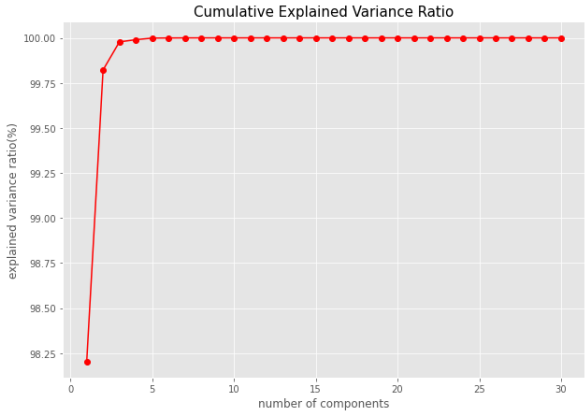

Figure 4: Number of components and explained variance ratio

3.3. Experiment analysis

During the test, many algorithms were adopted to train the model. Such as, Recursive Feature Elimination and Variance filtering were used to features select, Principal Component Analysis was used to extract the overall characteristics of original features and Random Forest Classifier was use

to diagnosis the Breast Cancer [11]. At last, using Grid SearchCV (GSCV) or HalvingGridSearchCV (HGSCV) was used to optimize the Hyper-Parameters of the above algorithms. The optimal parameters of the model are obtained by combining various algorithms. Such as 'n_features_to_select' of Recursive Feature Elimination is 16, the 'n_estimators' of RFC is 15, 'the max_features' of RFC is log2, etc. Diagnosis accuracy of each combination are shown in Table 1. Compared with other methods, this method improves the diagnostic accuracy.

Table 1: Compare with different methods.

Methods	Accuracy
RFC+Chi2	0.9157%
Ref[11]	95.6%
RFC	0.9543%
RFE+RFC	0.9648%
RFE+RFC+HGSCV	0.9613%
RFE+RFC+GSCV	0.9683%
RFE+PCA+RFC+GSCV	0.9701%

4. Conclusions

Feature selection is a very important step in classification since the inclusion of irrelevant and redundant features often degrade the performance of a classification algorithm both in speed and prediction accuracy. However, each feature contains more or less useful information. In this paper, the combination of Feature Selection algorithm (REF) and Feature Extraction algorithm (PCA) is used to make full use of the characteristics of each method and Grid search is used to model Hyper-Parameters Optimization. The test results show the method is effective and can obtain high accuracy. Next, according to the actual situation, different dimensionality reduction methods, feature extraction methods and parameter optimization methods can be combined for modelling.

Acknowledgements

This work was supported by Overseas Study Project of Young and Middle-aged Key Teachers in Hebei Agricultural University.

References

[1] Eid, H. F., Hassanien, A. E., Kim, T. H., & Banerjee, S. (2013). Linear Correlation-Based Feature Selection for Network Intrusion Detection Model. Springer Berlin Heidelberg. Springer Berlin Heidelberg.

[2] Kumaravel., V. , Raja., K. , Kumaravel., V. , & Raja., K. . (2014). Feature Subset Selection Algorithm for High-Dimensional Data by using FAST Clustering Approach.

[3] Akadi, A. E., Ouardighi, A. E., & D Aboutajdine. (2008). A powerful feature selection approach based on mutual information. International journal of computer science & network security.

[4] Hall, M. A. (2000). Correlation-based feature selection for machine learning. Phd Thesis Waikato Univer Sity.

[5 Chaves, R., J Ramírez, JM Górriz, M López, & Segovia, F. (2009). Svm-based computer-aided diagnosis of the alzheimer's disease using t-test nmse feature selection with feature correlation weighting. Neuroscience Letters, 461(3), 293-297.

[6] Upadhyay, D., Manero, J., Zaman, M., & Sampalli, S. (2021). Intrusion detection in scada based power grids: recursive feature elimination model with majority vote ensemble algorithm. IEEE Transactions on Network Science and Engineering, PP (99), 1-1.

[7] Kim, J. "A Comparative Study of Sequential Feature Selection Methods for Support Vector Machine." (2008).

[8] Naing, T. T., & Khaing, K. T. (2010). Enhanced features ranking and selection using recursive feature elimination(rfe) and k-nearest neighbor algorithms in support vector machine for intrusion detection system. International journal of computer science issues.

[9] Chen, X. W., and J. C. Jeong. "Enhanced recursive feature elimination." Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on IEEE, 2008.
[10] Zhang, L., Zhi, L., & Yi, L. I. (2013). Research on diagnosis method of breast tumor based on factor analysis and ga-bp. Journal of Convergence Information Technology.
[11] Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 36(2p2), 3465-3469.