Feature Selection Method To Improve The Accuracy of Classification Algorithm

Rajit Nair, Amit Bhagat

Abstract*: Today, we are living in the era of big data and it is not an easy task to process big data. Big data are also known as high dimensional data, so to reduce this high dimensional data there are two methods one is feature selection and the other one is feature extraction. This paper present the methods based on feature selection which are selectkbest and selectpercentile. This work also shows how feature selection works and how it helps during classification process. There are mainly three feature selection methods one is univariate, other one is model based feature selection and the last one is iterative method. In this paper it has been shown that how accuracy has been improved in classification algorithms used in machine learning through these feature selection methods. The proposed work has increased the classification accuracy for the algorithm like Naïve Bayes, Support Vector Machine, Logistic regression and K- Nearest Neighbor. Comparing to all other algorithms Logistic Regression has achieved higher accuracy with 96.9%.

***Index Terms*: Dimensions, Preprocessing, Datasets, Classification, Big Data.**

I. INTRODUCTION

It is already known whatever the data which has been collected by social networking websites, ecommerce websites, learning websites and many more are treated as big data. At the same time these types of data are known as high dimensional data [1]. So to process this big data or high dimensional data it is very important that we must reduce the dimensions [2]. In big data [3] world, dimensions are basically the features or attributes of the data set [4]. To reduce the dimension basically there are two approaches one is feature selection and the other one is feature extraction. Here we focus on feature selection method [5]. There are three types of feature selection method filter [6], wrapper [7] and embedded [8].

II. PREPROCESSING METHOD

Before we get into feature selection, there is another important step that is preprocessing method [9]. As we already know the data which we are receiving are raw data and we cannot apply this raw data directly into machine learning. So before applying this data to machine learning algorithm we have to preprocess it otherwise our training dataset would not provide a accurate result. Some of the methods which are applied for preprocessing are binarization, scaling, normalization, mean removal etc.

A. Binary Removal

In this method we actually convert our attribute values in the form of 0 & 1. It is acceptable when we apply numeric mining algorithms on categorical data. Sometimes we try to convert the categorical attributes in the form of binary and then apply the algorithms on the binary data. If there are different values of categorical attribute, in that case we will decide on the basis of certain parameters that which of the values has to be taken 1 and which one to be zero. Like in case of sentiment analysis we try to convert the review of any product or movie in the form of 0 & 1, where 0 is for negative review and 1 is for positive review.

B. Mean removal

This method is also known as standardization. It is again one of the important step in data preprocessing. Sometimes its performance also gets degraded when individual features do not more or less look like standard normally distributed data. Practically we ignore the shape of distribution and just transform the data to center by discarding the mean or average value of each feature, then scale by dividing non-constant features by their standard deviation.

C. Scaling

It is the another method of preprocessing, it happens that when we are dealing with raw data many features have different levels of magnitude means they are at different scales there might be algorithms like SVM, neural network which are very sensitive to different scaled data, these algorithm might not perform well. So to scale the features properly there are methods like standard scaler, min-max scaler and normalizer.

D. Normalization

It is a type of scaling techniques, through which we can find new range from existing range. Mainly it helps in prediction or forecasting. There are many ways for forecasting and prediction but most of them are varied with each other. So to maintain these variations of forecasting and predicting, the normalization techniques is used to make them approximate. It is the way by which value of each feature vector on a common scale. There are two basic method of normalization L1 & L2. There are existing normalization technique like Min-Max, Zscore & Decimal scaling and Integer Scaling technique. This technique comes from the AMZD (Advanced on Min-Max Z-score Decimal scaling).

III. FEATURE SELECTION METHOD

Higher dimensionality is never a good choice because it is very complex and it increases over fitting, that's why we

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

reduce the features. We will discuss three strategies for feature selection first one is univariate statistics, second is model based selection and last one is iterative selection.

A. *Univariate method*

This method basically shows the statistically significant relation between each features and the output or target variable. This is also known as ANOVA or Analysis of Variance [10]. After the relationship have been computed the feature having higher confidences are selected. In this method we assume that each feature are independent and they are somehow related with target variable. Now to perform selection we will use two method selectkbest and selectpercentile. The selectkbest will select the k number of features and selectpercentile will make selection on the basis of percentage of features. Now we will show how these methods work on the breast cancer dataset and we will add some noise before selection.

B. *Model based feature selection*

This method uses a supervised model [11] to determine the importance of each feature. It only keeps the most important features. It needs a measure for the importance of features (DT and RF have the feature_importances attribute). Here first we need a model that shows the important of each feature, so for that decision tree and random forest for this purpose. Now we will define some of the existing model based feature selection method namely tree based feature selection and L1 based feature selection.

C. *Iterative selection*

Iterative method [12] in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

IV. PROPOSED METHOD

In supervised learning we always go for classification and to perform classification there are various algorithms like Naïve Bayes [13], SVM [14], logistic regression [15], random forest [16] and many more. Out of these algorithms we cannot say which is better because it totally depend upon the condition and the dataset on which we are performing classification. In this paper we have shown how dimensionality reduction or feature selection increases the performance or accuracy of classification algorithm. Methods used for feature selections are selectkbest and selectpercentile. Advantages of using these methods are as follows:

- Reduces Overfiting: Less redundant data means less possibility of making decisions based on redundant data/noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: Less data means that algorithms train faster.

A. *SelectkBest*

Score function is used which takes parameters and applied on the (X, y). This function returns an array of scores. Scores are computed for each feature X[:,i] of X. Sometimes it returns the p value also which is neither needed nor required. This feature selection method basically returns the first k features of X with highest scores.

Suppose you pass chi2 as a score function. This method will compute chi2 statistic between each features of X and y i.e. assumed to be class labels. If the computed value is less this indicates that feature is independent of class label and high value indicates that the feature is non-randomly related to class label. This way we only select the highest k features. This function can be called as:

select = SelectKBest() and

then call select.fit_transform(X, y)

B. *Selectpercentile*

SelectPercentile, selects a fixed percentage of features. This method is imported from scikit-learn, which actually computes how many numbers of features have to be selected from the given dataset contain with feature scores generated by selectkbest method.

The dataset which we are using is breast cancer dataset which contains 30 features after reducing these features the classification algorithm classifies the data in to two target variable one is malignant and the other one is benign.

C. *Datasets*

In this work we use breast cancer dataset which contains 30 features and noise features are also added and that is 50 so total of 80 features are there. So to perform feature selection we first split the data into training and testing data. X_train and y_train are the parameters for training and X_test and y_test are for testing parameters. We can use any of the method like SelectKBest or SelectPercentile method. Some of the features are namely mean radius, mean texture, mean perimeter, mean area mean smoothness, mean compactness, mean concavity, mean concave points, mean symmetry, mean fractal dimension, radius error, texture error, perimeter error, area error, smoothness error, compactness error, concavity error, concave points error, symmetry error and many more. The target variables are malignant and benign.

V. RESULTS

In this section we will show how feature selection improves the efficiency of classification algorithm. Our first step is to load the breast cancer dataset which already mentioned in the above section which have 30 features and after considering noise features it becomes 80 and we will show how this features will reduce using the SelectKBest and SelectPercentile method. The SelectPercentile method when apply on the cancer dataset and it reduce the features from 80 to 40. As you can see below how SelectPercentile method perform on the breast cancer dataset. The implementation

is done in Pyhton 3.6 and the output is given:
X_train.shape is: (284, 80)
X_train_selected shape is: (284, 40)
Here we can also see what features they have taken by display them
[True True True True True True True True True False True False
True True True True True True False False True True True True
True True True True True True False False False True False True
False False True False False False False True False False True False
False True False True False False False False False False True False
True False False False False True False True False False False False
True True False True False False False False]

The variables which are having true identity are the taken features and the false are not selected feature.
In the below fig 1. black color shows the taken feature and white shows the deselected feature. After performing feature selection accuracy has been improvised, the implementation of all the classifiers with and without feature selection is given in the form of table 1. this is shown below:

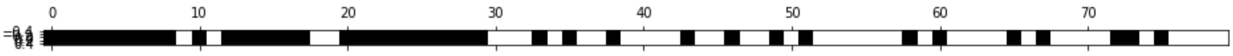

Fig 1. Selected and non selected features

Table 1. Comparison table of the classifiers

Classifiers	Accuracy(Before Feature Selection)	Accuracy(After Feature Selection)
Naïve Bayes(NB)	0.940351	0.959158
Support Vector Machine(SVM)	0.936842	0.955579
Logistic Regression(LR)	0.929825	0.969895
K-Nearest Neighbor	0.929825	0.948421

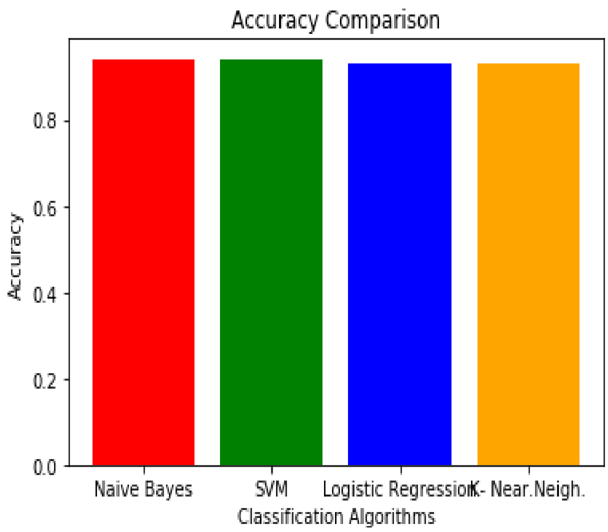

Fig 2. Accuracy comparison of classifiers before feature selection

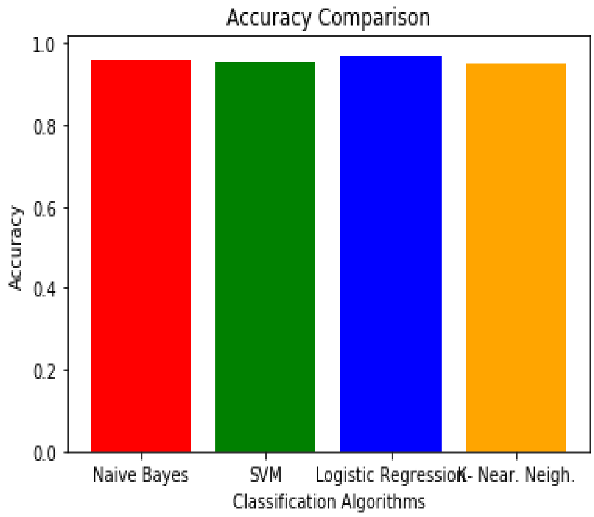

Fig 3. Accuracy comparison of classifiers after feature selection

VI. Conclusion and Future Work

As it is observed that dimensional reduction will increase the performance of classification algorithms but in our previous section we have seen some of the classification like SVM and KNN has not been improved. So it is not necessary that always dimensional reduction will improve the performance of classification algorithm. Considering this work, future research will be on classification of genes expression by selecting different features and

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

through this classification we will try to find the genes which are performing normal and which are not. In the future work we will implement deep learning methods for classification and check whether they need dimensionality reduction methods for improved accuracy. It is assumed that deep learning methods have inbuilt dimensionality reduction approach.

REFERENCES

1. D. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," AMS Conf. Math Challenges 21st Century, 2000.
2. M. Sugiyama, "Linear Dimensionality Reduction," in Introduction to Statistical Machine Learning, 2016.
3. M. Chen, S. Mao, and Y. Liu, "Big data: A survey," in Mobile Networks and Applications, 2014.
4. R. Nair and A. Bhagat, "A Life Cycle on Processing Large Dataset - LCPL Rajit Nair," vol. 179, no. 53, pp. 27–34, 2018.
5. V. Bolón-Canedo and A. Alonso-Betanzos, "Feature selection," in Intelligent Systems Reference Library, 2018.
6. G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Comput. Electr. Eng., 2014.
7. R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artif. Intell., 2002.
8. P. Yang, W. Liu, B. B. Zhou, S. Chawla, and A. Y. Zomaya, "Ensemble-based wrapper methods for feature selection and class imbalance learning," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013.
9. L. Ladha and T. Deepa, "FEATURE SELECTION METHODS AND ALGORITHMS."
10. J. A. Cuesta-Albertos and M. Febrero-Bande, "A simple multiway ANOVA for functional data," Test, vol. 19, no. 3, pp. 537–557, 2010.
11. R. Gupta, "MACHINE LEARNING Raghav Agarwal," no. 11, pp. 1342–1347, 2015.
12. T. M. Khoshgoftaar, K. Gao, and A. Napolitano, "Exploring an iterative feature selection technique for highly imbalanced data sets," Proc. 2012 IEEE 13th Int. Conf. Inf. Reuse Integr. IRI 2012, pp. 101–108, 2012.
13. K. P. Murphy, "Naive Bayes classifiers," Bernoulli, 2006.
14. C. Chang, C. Lin, and T. Tieleman, "LIBSVM : A Library for Support Vector Machines," ACM Trans. Intell. Syst. Technol., 2008.
15. S. Sperandei, "Understanding logistic regression analysis," Biochem. Medica, 2014.
16. A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in Ensemble Machine Learning: Methods and Applications, 2012.

AUTHORS PROFILE

Rajit Nair - I am having 12 years of teaching experience. I am pursuing my Ph.D. from NIT Bhopal, India. I have presented papers in various national and international conferences. Published my papers in various national and international journals. Written 2 book chapters in international book publication.

Dr. Amit Bhagat - Currently working as a Assistant Professor in MANIT, Bhopal, having expertise in the field of Data Science, Machine Learning, Big Data and many more. Published many papers in international journals which includes Scopus indexed journals. Presented many papers in international conferences.