Print ISSN: 2321-6379
Online ISSN: 2321-595X
DOI: 10.17354/ijssl/2017/20

Original Article

An Approach to Feature Selection Using Random Projection Method

Ramin Nasiri, Hessam Abjam

Department of Computer, Faculty of Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran

Abstract

In This Paper, By introducing an approach to feature selection via use of random projection, we aim to reduce data dimensions by a more advanced method and also achieving an accurate prediction of knowledge inherent in data. In the proposed approach, reducing dimensions of data by random projection method after preprocessing is applied on the data. Then finding the threshold will be investigated and after finding the threshold, the classification operations is done on the original data and the data that have been reduced by the above method. At the end, results are compared and evaluated with each other.

Key words: Feature selection, Random projection, Reducing dimensions

INTRODUCTION

Many algorithms have been proposed for feature selection, which often have problems such as the need for high computations. In machine learning techniques, in the case of working with high dimensionsof data and having a number of top features, we will be faced with problems such as the complexity of computations in time of time and memory and understanding of the problem and the difficulty of extracting knowledge. Therefore, using a variety of methods as ways of reducing the dimensions, we imagine data to a space with less dimension in a way that, if possible, the properties of initial space not to be lost.[1] By creating an approach in feature selection where using random projection that is considered one of the ways of reducing the dimensions, can be achieved a method for feature selection of data and thus, more accurate prediction of knowledge required in the data.

Related Works

Kevin Fernandes, Pedro Winger and Paulo Gortez provided a method of feature projection in the form of supporting system of decision in which analyzes the contents of an online news site and predicts the popularity of online news among the users. In the results of this study, random tree method provides accuracy and better results.[1] Mazart andcolleaguesprovided an article where a method of feature selection using analysis method of main components to reduce dimensions of data has been designed andtwo dataset is used for evaluating machine learning algorithms, genetic algorithm performed best among them.[2] JasnaKhozepublished an article about the feature selection methods in the data in which investigated describing methods of reducing dimension of data and its impact on data mining. Issues such as improving forecasting performance, increasing the speed of calculations and reducing the costs of predictionexpressed in this study.[3]

Access this article online

www.ijss-sn.com

Month of Submission : 06-2017
Month of Peer Review : 06-2017
Month of Acceptance : 07-2017
Month of Publishing : 07-2017

Database

In this study, the Mashable data set in the database UCI Machine Learning is used. This database using a set of features available detects the number of share in a context in social networks. This database contains heterogeneity summary information of the articles posted on the website Mashable within two years. Number of properties of dataset is 61 that show the last feature of class of data or in other words the number of share related articles. The material presented in this chapter is based on this database. The total number of records is 39797. Features of database do not show the actual texts published but displaysome statistical information about them. Table 1 shows Mashable Statistics dataset.[1]

The method used to fill missing values in this study is the use of average values of feature in cells missing. This means

Corresponding Author: Hessam Abjam, Department of Computer, Faculty of Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran. E-mail: hessam.abjam@gmail.com

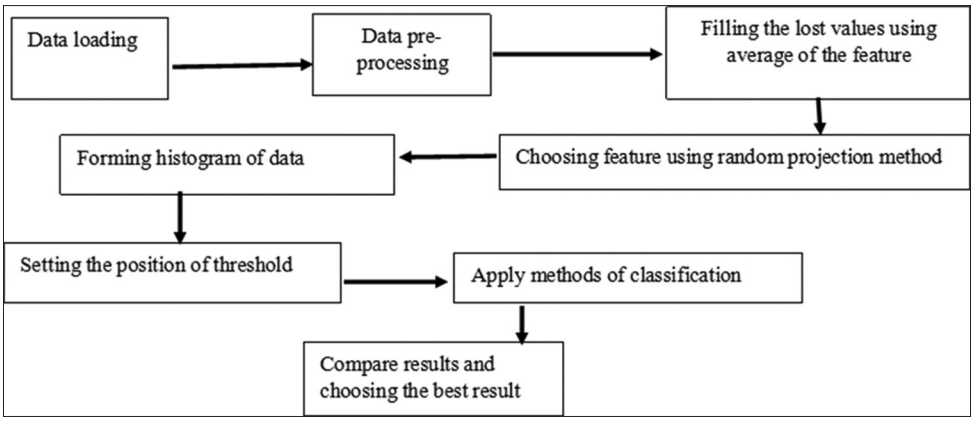

Figure 1: Flowchart of research

Table 1: Statistics of the dataset used Mashable[1]

		Average per day			
Articles number	Number of days	Average	Standard deviation	Minimum	Maximum
39.797	709	55.00	22.65	12	105

for the feature that have continuous values is the average values of the feature. Innovation of this study is the use of random projection. In this way that before selection of features unlike similar approach, using a randomprojection and reducing dimensions of data, minor features are deleted.This improvement of the performance of features selection and improve forecast accuracy causesto improve and provide a new approach in feature selection to predict the popularity of online news. Figure 1 shows the research process.

Determine the Location of Threshold

To determine the number of thresholds and specify a numeric value of thresholds, the histogram related to thecolumn of the number of shares is used. The local maximum points of this chart will be used to divide the range of values of this index in different categories. So our question becomes a problem of classification and can be used the algorithms of this field.[4] Figure 2 shows a histogram related to the number of article sharing. It is observed that there is only a clear maximum in data and in other cases, the frequency for the number with shares not has a significant difference compared to the points of neighbor. Figure 2 shows frequency of sharing articles on the neighboring of maximum number 1000.

According to the results in Figure 2, we consider the number of share 1000 as a threshold. So articles that the number of their share is less than 1,000 are initialized in the category of "low" share and articles with more shares than 1,000 in the category of "high" share are initialized.

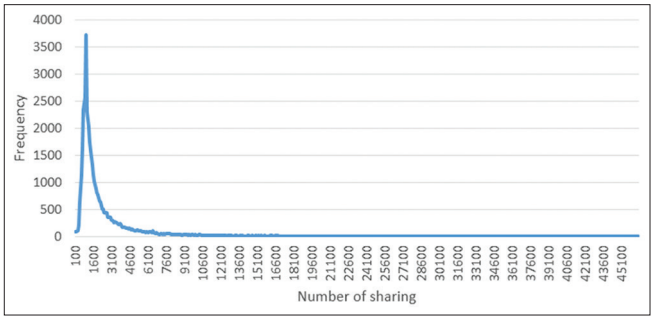

Figure 2: Histogram of number of sharing articles

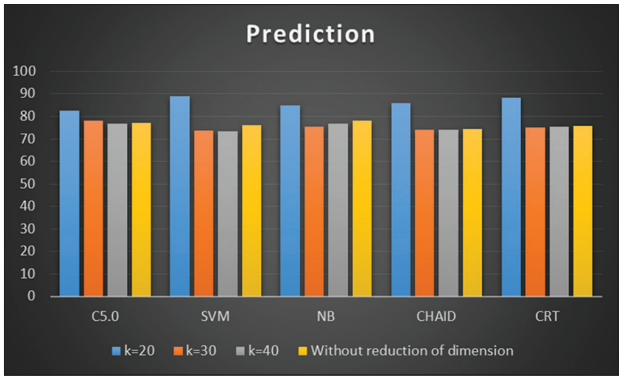

Figure 3: Comparison of the prediction obtained value for different values s of K

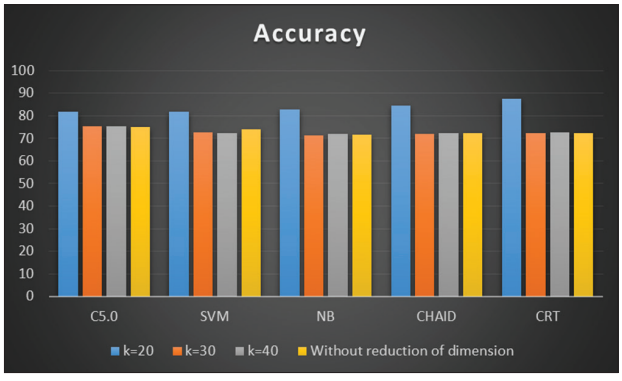

Figure 4: Comparison of the accuracy obtained value for different values s of K

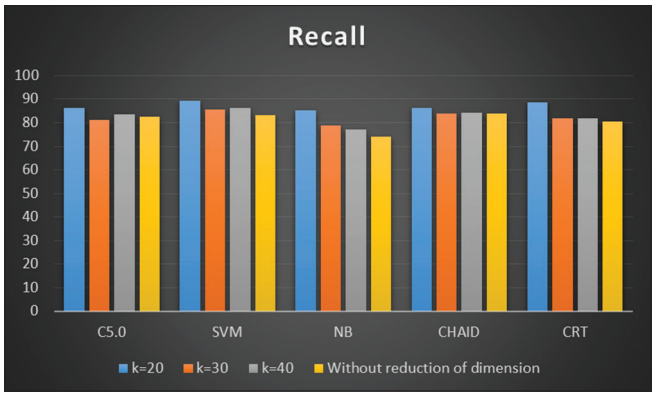

Figure 5: Comparison of the recall obtained value for different values s of K

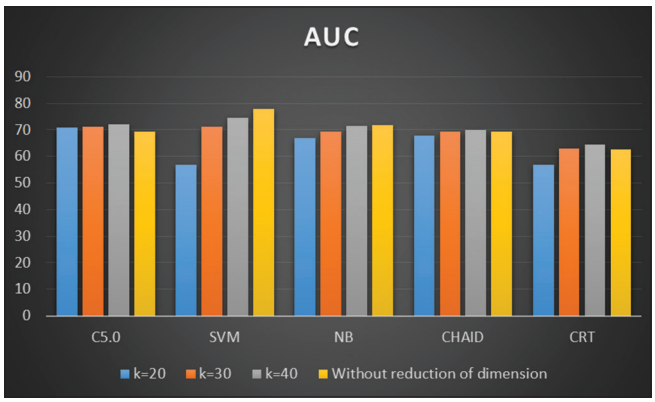

Figure 6: Comparison of the AUC obtained value for different values s of K

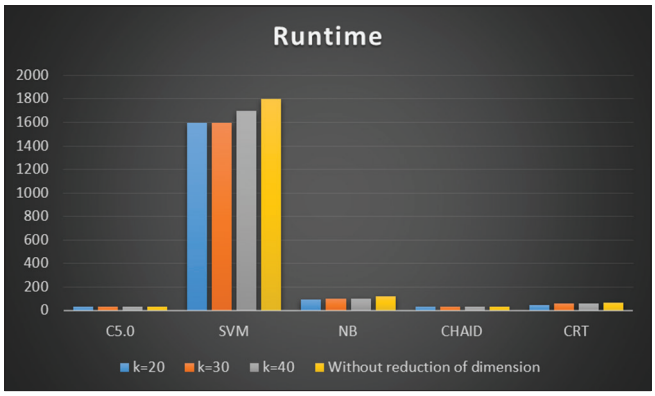

Figure 7: Compare the value of run time in seconds for different values of K

Applying classification methods and Comparison based on different values of the number of features.

In order to use different methods to classify and take advantage of each one, several ways to classify the data of pre-processing and feature selection were performed. It is observed that the criteria compared of applied methods on the data are improved in the case that random projection feature selection method is used. Variable K is the number of properties obtained after random projection. This chart shows improving accuracy in methods that have used feature selection. In this section, to evaluate the proposed method, different values obtained from number of features are considered and the results are compared together.[5-18]

As seen in the Figures, the highest accuracy is obtained with the amount of %89.01 and calling amount with amount %89.29 percent by SVM classifier. Also as shown in Figure, the best prediction accuracy rate was obtained % 87.78 with CRT classifier. All the percentages mentioned are obtained if the dimensions of features decreased to 20 features. However, in AUC value obtained, the best value for SVM classifier was with value %77.9% while the value was obtained that after feature selection with method provided, the AUC decreased. In another sense, the proposed method in improving standard AUC in all classifiers has not performed well. The next thing is the runtime. The considered point is high time of implementation of SVM classifier with high levels of precision and recall than other classifiers. The best run time is related to C5.0 classifier that the runtime takes only 10 seconds (Figures 3-7).

CONCLUSION

In this study, random projection feature selection method was used to select the appropriate features and dimension reduction of data. In the database used with 60 features, by reducingthedimensions of data and reaching the number of features between the ranges of 20 to 40, it has investigated obtained results. Then it was investigated to apply various algorithms to classify. Techniques such as support vector machines, decision tree, etc. were applied to the data. In the end, by doing the operations of feature selection after data dimension reduction, it was referred to classification of data compared to before and after data dimension reduction. In future research,one can change method of feature selection to the ways such as INTERACT in which the relationship between features also is effective in the importance of a feature and compare results obtained in these ways with each other. Also one can use other methods of reducing dimensions of data such as PCA and NMF to improve results in the classification and compared with RP.

REFERENCES

1. K. Fernandes, P. Vinagre, P. Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News", Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, Vol. 50, pp. 535-546, 2015.
2. Q. Guo, W. Wu, D. Massart, C. Boucon, S. De Jong, "Feature selection in principal component analysis of analytical data", Chemo metrics and Intelligent Laboratory Systems, Vol. 24, pp. 119-125, 2002.
3. J. Jose, "A Survey on Feature Selection Techniques", International Journal of Engineering Research and General Science, Vol. 40, pp. 16-28, 2014.

4. P. Vehviläinen, "Data Mining for Managing Intrinsic Quality of Service in Digital Mobile Telecommunications Networks," Doctoral thesis, Tampere University of Technology, 2004.
5. A. Blum, "Random Projection, Margins, Kernels, and Feature-selection", in Proceedings of the 2005 International Conference on Subspace, Latent Structure and Feature Selection, pp. 52-68, 2006.
6. X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach" in ICML, vol. 3, pp. 186-193, 2003.
7. A. Sahu, D. W. Apley, G. C. Runger, "Feature selection for noisy variation patterns using kernel principal component analysis", Knowledge-Based Systems, Vol. 72, pp. 37-47, 2014.
8. M. Adedoyin-Olowe, M. M. Gaber, F. Stahl, "Survey of Data Mining Techniques for Social Network Analysis", International Journal of Research in Computer Engineering and Electronics, Vol. 3, Issue: 2, pp. 19-39, 2014.
9. F. Song, Z. Guo, D. Mei, "Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining", Fifth International Conference on Communication Systems and Network Technologies, Vol. 67, pp. 119-125, 2015.
10. A. Khan, H. Farooq, "Principal component analysis-linear discriminant analysis feature extractor for pattern recognition", Vol. 8, pp. 2-6, 2012.
11. V. S. Verma, R. K. Jha, A. Ojha, "Digital watermark extraction using support vector machine with principal component analysis based feature reduction", Journal of Visual Communication and Image Representation, Vol. 60, pp. 250-256, 2015.
12. S. Moro, P. Cortez, P. Rita, "A data-driven approach to predict the success of bank telemarketing", Decision Support Systems, Vol. 62, pp. 22-31, 2014.
13. S. Moro, R. Laureano, P. Cortez, "Using data mining for bank direct marketing: An application of the crisp-dm methodology", in Proceedings of European Simulation and Modelling Conference-ESM, Vol. 32, pp. 21-26, 2015.
14. M. Kesharaju, R. Nagarajah, "Feature selection for neural network based defect classification of ceramic components using high frequency ultrasound", Vol. 62, pp. 271-277, 2015.
15. I. Ahmad, M. Hussain, A. Alghamdi, A. Alelaiwi, "Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components", Neural Computing and Applications, Vol. 50, pp. 500-505, 2014.
16. B. Liu, E. Blasch, Y. Chen, D. Shen, G. Chen, "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier", IEEE International Conference on Big Data, Vol. 62, pp. 124-130, 2013.
17. H. Hasan, N. M. Tahir, "Feature selection of breast cancer based on principal component analysis", in Signal Processing and Its Applications, Vol. 6, pp. 1-4, 2010.
18. F. Song, Z. Guo, D. Mei, "Feature selection using principal component analysis", Engineering Design and Manufacturing Informatization (ICSEM), Vol. 6, pp. 27-30, 2010.

How to cite this article: Nasiri R, Abjam H. An Approach to Feature Selection Using Random Projection Method. Int J Sci Stud 2017;5(4):140-143.

Source of Support: Nil, **Conflict of Interest:** None declared.