DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5496

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

RANDOM SUBSET FEATURE SELECTION FOR CLASSIFICATION

Lakshmi Padmaja Dhyaram
Associate Professor
Anurag Group of Institutions
Hyderabad, Telangana, India

Dr. B. Vishnuvardhan
Professor,
JNTUH,
Hyderabad, Telangana, India

***Abstract*:** Feature selection has been the focus of interest in the recent past. Large data sets are collected from scientific experiments and many times features are outnumbered the observations. This demands for new approaches to minimize the data set without compromising the latent knowledge. This is also called dimensionality reduction. In this paper, we have presented a detailed review of methods used in minimizing the datasets. We have selected papers which are published last 10 years in the field of dimensionality reduction using Random Subset Feature Selection (RSFS). We have concentrated mainly on random subset feature selection methods used in the dimensionality reduction. The feature subset selection methods are classified into two 4 categories-Embedded, Filter, Wrapper and Hybrid. The data mining task flow from pre-processing, feature subset selection using random forest, random subset feature selection and classification. This survey is a comprehensive overview on random subset feature selection used in various applications.

Keywords: Feature Selection, Feature Subset Selection, Random Forests, Random Subset Feature Selection

INTRODUCTION

Feature Subset Selection (FSS) is a process to select subset of relevant features, where as Random Feature Selection(RSFS) process, randomly selects the subset of relevant features from the data set to avoid the bias and over fitting selected subsets, for classifying the features from high dimensional data sets. Before starting the process of RSFS the origin for feature selection is briefly introduced in the following.

A. Feature Selection

Feature selection is a supervised inducting learning algorithm to find the relevant features and to eliminate the irrelevant features or redundant features for better accuracy. Feature selection for classification discussed by author[1]. Feature selection to improve performance (in terms of speed, predictive power, simplicity of the model),to visualize the data for model selection and to reduce dimensionality and remove noise. Basically feature selection is a process to select optimal subset of features from the total number of features. Sequential forward selection(SFS)[2] by adding the new feature with the existing feature relevancy in every iteration, initially it starts with one feature, where as back- ward selection removing the irrelevant features from the total number of features based on its relevancy. The drawback of the above method is only adding the new feature every iteration, removal of feature is not possible which leads to nesting problem. The nesting problem is resolved with Sequential Floating Forward Selection(SFFS) by adding the relevant feature and removing the irrelevant feature from the total data set. set[3].

Measuring the accuracy of learning process done by various measures like distances, dependence or consistency. The filter methods are usually faster, it does not rely on a particular learning bias, in such a way that the selected features can be used to learn different models from different Data Mining techniques, for high dimensional data can handle, due to the simplicity and low time complexity of the evaluation measures.

Predictive performance can be achieve by wrappers and which control the over fitting usage of internal statistical validation to control the over fitting, filter models cannot allow a learning algorithm to fully exploit its bias, whereas wrapper methods do.

In Embedded feature selection, features during the learning process. In this process training and testing of data sets splitting is not required. It avoids re-training of a predictor for each subset gives the solution in faster. By using these measures feature subset selection can be processed.

1) Feature Subset selection: Feature subset selection is to select the features with a specified subset size that optimizes the evaluation measure. Subset selection has two advantages, the lower cost and data collection is faster than measuring the entire data set. Especially in scientific data sets like cancer data sets, gene classification, micro array data sets etc., have more features than the observations. The pre-processing task in data mining is a great challenge. For cancer, or heart disease data sets for identification of uninformative features, removing from the data sets gives a great work for pre screening to diagnose whether patient effected or not for the disease[4], [5]. The disadvantage of traditional feature subset selection is the subset of features are fixed, but in random subset feature selection, in every iteration randomly generates new subset with new features. In turn, this avoids the bias. The random subset feature selection gives mores accuracy compared to traditional method SFS, SFFS, SBS. The random subset feature selection follow the random forest algorithm discussed in the following section.

B. Random Forest

The main applications of random forests are medical diagnosis and document retrieval where there are number of features and less number of observations. Random forests is an ensemble classifier and well developed technique for classification. Random forests are applied to substantially totally different areas of application, where as the analysis of micro arrays of organic phenomenon , there are thirteen genome-wide association studies and also the analysis of gene interactions, for a review see the prediction of protein interactions, weather forecasting, land cowl classification.

Before Brieman, Random forests paper explained[6]. Some other authors has strives for different ideas when large number of features for high dimensional data sets as show in below.

- Irrelevant features and subset selection problem by Ron Kohaivi et. al. 1994
- An early example is bagging (Breiman [1996]), Each tree growing using random selection (without replacement) is made in the training set is bagging.
- In an important paper on written character recognition, Amit and Geman [1997] define a large number of geometric features and search over a random selection of these for the best split at each node.
- Amit and Geman [1997] analysis to show that the accuracy of a random forest depends on the strength of the individual tree classifiers and a measure of the dependence between them
- Another example is random split selection (Dietterich [1998]) where at each node the split is selected at random from among the K best splits.
- Ho [1998] has written a number of papers on "the random subspace" method which does a random selection of a subset of features to use to grow each tree.
- Breiman [1999] generates new training sets by randomizing the outputs in the original training set. Another approach is to select the training set from a random set of weights on the examples in the training set.
- Voting by the most popular classes for large number of generated trees technique is called random forests. Brieman 2001.

C. Random KNN-Feature Selection(RKNN-FS)

For large number of features and less number of observations RKNN-FS is more faster and robust with Random KNN classifier(RKNN)is an alternate to random forests. RKNN-FS is easy to implement and stability for high dimensional data sets. 2011 Li. et. al [7]

D. Random Subset Feature Selection

Random subset selection is a useful approach to find the variability of static data such as location and scaling estimates. When no outliers or data anomalies the random subset results provides a useful measure of the inherent variability of the data characterization of interest. The main goal of the random subset selection is to generate q subsets to generate from the data set D each of the same size M. The simple version of random subset selection is probably random selection with replacement in which each of the M elements of the subset Si is randomly drawn from the size N data set independently with probability 1/N[8].

E. Random Subset Feature Selection Method Application areas:

1. Random Approximated Greedy Search (RAGS) is a sequential approach developed in this paper and apply it to the feature subset selection for regression based on GRASP/Super-heuristics Method for combinatorial optimization problems, where estimation of performance is costly. The main idea of RAGS are from Ordinal Optimization (OO) technique. The aim of the paper to get the optimality for crude estimation model and for finding the estimation of performance error. After applying many times the running of RAGS better solutions than greedy search under the computational effort at the same time. Here evaluation procedure is time consuming. The authors experimented with feature subset selection technique[9].
2. Joseph DePasquale, Student Member IEEE and Robi Polikar, Member IEEE 2007 This paper developed based on ensemble of classifiers based algorithm for the missing feature problems with an inspiration of random subspace method. To classify an observation with missing features only those classifiers whose training data did not include the currently features are used. Basically it is an incremental learning algorithm. This algorithm tried up to 30% of missing data without significant drop in performance. For smaller dimensionality the classifier trained faster compared to high dimensional data sets with missing features[10].
3. Prediction of Heart Disease using Random Forest and Feature Subset Selection. In this paper the authors proposed a new method uses random forest algorithm for predicting the heart disease, to identify the un informative features and for their removal chi square feature subset selection algorithm is used. And also chi-square test used to improve the accuracy in predicting heart disease[11].
4. Random Feature Subset selection in a non stationary environment: Application to textured image segmentation-This algorithm used as an application to textured image segmentation for 2 class labels. Here the new feature subset selection intends to optimize or performances maintaining of decisional system in case of deviation or loss of information. Experimentation results to determine the optimum features subsets to improve accuracy of the classification for textured image segmentation. Proposed method is for non stationary environment method[12].
5. Conditional Random Fields feature subset selection based on Genetic Algorithms- for phosphorylation site prediction Conditional Random Fields (CRFs) are undirected probabilistic graphical models that were introduced for solving sequence labelling and segmenting problems based on genetic algorithms to find the performance. Basically is a part of bio informatics for protein modification mechanism.
6. Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech- In this paper the speech recognition of a child, classification of

different age groups and lastly recognizing the deviation from social conflicts occurred during speech. Authors proposed a method for identifying the performance of the classifier for this work and to eliminate the over fitting problem. Speech recognition is a challenging task, having more number of features to classify or to reduce the dimensionality[13].

7. Feature Selection Methods and Their Combinations in High-Dimensional Classification of Speaker Likability, Intelligibility and Personality Traits- This study concentrated on recognition of speaker likability, intelligibility and five personality traits by using RSFS algorithm to find the performance using kNN classifier for classification of features in speech. Dimensionality reduced without impairing its originality. The performance is calculated or RSFS on this area, with combi- nation of different feature selection methods comparison made using greedy hill climbing technique[14].

8. Comparative Study of Feature Subset Selection Methods for Dimensionality Reduction on Scientific Data. In this study the existing feature selection methods compared to RSFS algorithm particularly on scientific data. Scientific data mainly having more number of features and less number of observations eg. Bio informatics, medical diagnosis, genes analysis etc. The RSFS algorithm performance is better to compare with other methods. The accuracy vs features graph show that RSFS has more accuracy and it removes over fitting problem[15], [16]

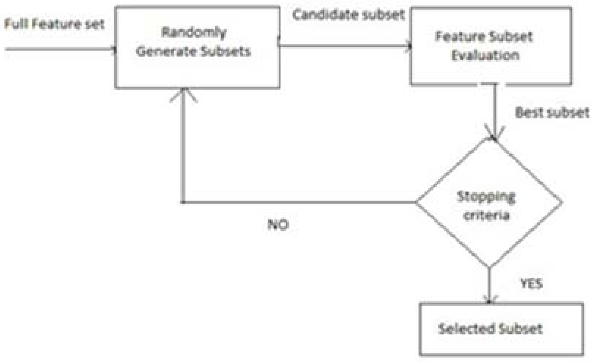

Fig. 1. Random Subset Feature Selection Algorithm

CONCLUSION

This paper describes the overview of Random subset feature selection, techniques used in and its application areas in different fields. The main goal of data mining techniques are to discover the knowledge from active data. Out of all techniques classification in one of the prominent area to find relevant features from large amounts of data. In future work we can develop various learning algorithms for random subset feature selection method.

REFERENCES

1. M. Dash and H. Liu, “Feature selection for classification,” *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.
2. A. W. Whitney, “A direct method of nonparametric measurement selection,” *IEEE Transactions on Computers*, vol. 100, no. 9, pp. 1100– 1103, 1971.
3. P.Pudil, J.Novovicova, and J.Kittler, “Floating Search methods in feature selection,” *Elsevier science*, vol. 1994, no. 15, pp. 1119 – 1125,
4. Nov. 1994.
5. R. Kohavi and G. H. John, “Wrappers for feature subset selection, ”*Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.
6. H. John, R. Kohavi, K. Pfleger *et al.*, “Irrelevant features and the subset selection problem,” in *Machine learning: proceedings of the eleventh international conference*, 1994, pp. 121–129.
7. L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
8. Li, S, Harner, J., and Adjeroh, D, “Random kNN feature selection a fast and stable alternative to random forests. BMC Bioinformatics,” Dec. 2011.
9. R. K. Pearson, *Mining imperfect data: Dealing with contamination and incomplete records*. SIAM, 2005.
10. Gao and Y.-C. Ho, “Random approximated greedy search for feature subset selection,” *Asian Journal of Control*, vol. 6, no. 3, pp. 439–446, 2004.
11. J. DePasquale and R. Polikar, “Random feature subset selection for ensemble based classification of data with missing features,” in *International Workshop on Multiple Classifier Systems*. Springer, 2007, pp. 251–260.
12. M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, “Prediction of heart disease using random forest and feature subset selection,” in *Innovations in Bio-Inspired Computing and Applications*. Springer, 2016, pp. 187–196.
13. X. He, P. Beauseroy, and A. Smolarz, “Random feature subset selection in a nonstationary environment: Application to textured image segmentation,” in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 3028–3031.
14. O. Ra¨sa¨nen and J. Pohjalainen, “Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech.” in *INTERSPEECH*, 2013, pp. 210–214.
15. J. Pohjalainen, O. Ra¨sa¨nen, and S. Kadioglu, “Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits,” *Computer Speech & Language*, vol. 29, no. 1, pp. 145–171, 2015.
16. D. L. Padmaja and B. Vishnuvardhan, “Comparative study of feature subset selection methods for dimensionality reduction on scientific data,” in *Advanced Computing (IACC), 2016 IEEE 6th International Conference on*. IEEE, 2016, pp. 31–34.
17. D. Lakshmi Padmaja and Dr.Vishnuvardhan, “Survey of dimensional- ity reduction and mining techniques on scientific data,” *IJCSET*, vol. 5, no. 11, pp. 1062–1066.