International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

A Review on Feature Subset Creation Strategies

Ms. P. P. Patil[1], Prof. S. S. Banait[2]
[1]M.E. Student, [2]Assistant Professor) Department of computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik Savitribai Phule Pune University, Maharashtra, India.

***Abstract**: To reduce computational overhead in processing high dimensional dataset, dimensionality reduction is important mechanism to remove redundant and unused attributes from dataset in data analysis phase. Feature selection and feature extraction are two techniques in dimensionality reductions. This work aims to study various techniques involved in feature subset generation and reduction of data set size, its efficiency in terms of execution time and quality analysis parameters.*
***Keywords**: Dimensionality Reduction, Feature Selection, Feature Extraction, Compound Features, Classification, clustering.*

I. INTRODUCTION

Large amount of data with multiple dimensions is generated in various applications. Processing such large dataset requires high computational power and high time for execution. In high dimensional dataset important attributes are present along with noisy attributes. These noisy attributes degrades the performance of classification and clustering techniques.

To work with large amount of data in data mining, dimensionality reduction is the important aspect. In dimensionality reduction non redundant features are preserved and redundant, unnecessary noisy features are removed from the dataset. The generated sub set improves the algorithm performance, accuracy in case of classification and Normalized Mutual Information -NMI in case of classification as it removes redundant and non relevant features from dataset.

Traditional dimensionality reduction methods are classified in two basic categories: feature selection and feature extraction. Feature selection technique mine the important features from given dataset whereas in feature extraction process combination of original feature is generated as a compound feature. The feature selection and extraction techniques are again classified in two sections: supervised and unsupervised based on the nature of dataset. In supervised strategy class labels are used for selecting the feature subset whereas in unsupervised learning no class label data is provided. Features are selected based on data variance and distribution.

Paper is organized as follows: section I introduces the outlier detection techniques and methods. Section II gives the literature review. Section III concludes the paper. Following literature work focuses on different strategies proposed for feature subset generation in existing work.

II. LITERATURE WORK

The dimensionality reduction is broadly classified in following 2 sections:

A. *Feature Selection*

In feature selection process, a feature subset is generated by removing redundant features from given dataset. The feature subset includes least number of features that maximizes the classification accuracy in case of supervised learning and better Normalized Mutual Information-NMI, Jaccard coefficient, Jacc and Fowlkes-Mallows index- FM in clustering in case of unsupervised learning.

In supervised feature selection technique [4][10] [13][14] important features are extracted by evaluating the correlation of features with respective class labels and prediction performance. RELIEF algorithm family is discussed [4]. This RELIEF algorithm is used for feature selection in binary class attribute feature selection. This algorithm is best for attribute estimation and evaluating the conditional dependencies between attributes.

Graph based feature selection technique is proposed in [10]. In this, instead of comparing a selected feature with class labels, it is compared with whole selected feature subset. This technique optimizes the score of entire feature subset. This provides the global optimal solution for feature selection.

Minimum Redundancy Maximum Relevance Feature Selection – mRMR model is proposed in [14]. This model removes the redundancy occurred during sequential feature selection strategy. This method uses greedy search to select next best suited feature and hence provide the ranking score to the selected features.

In unsupervised learning technique [9][6][11] dataset features are selected using data variance and distribution. A Niching Memetic Algorithm is proposed in [9] to perform feature selection and clustering simultaneously. NMA_CFS(A Niching Memetic Algorithm for simultaneous clustering and feature selection) makes feature selection an integral part of the global clustering search procedure and attempts to overcome the problem of identifying less promising locally optimal solutions in both clustering and feature selection This technique mainly works on chromosomes dataset. X. He, D. Cai, and P. Niyogi [6] introduced a novel feature selection algorithm called Laplacian Score(LS). Laplacian Score (LS) is used to select the features based on their locality preserving power. Data variance is calculated using Laplacian Score. This technique uses Laplacian Eigenmaps and Locality Preserving Projection.

These NMA_CFS [9], Laplacian Score(LS) [6] techniques selects the feature based on the individual feature score and do not calculate the correlation among selected attributes. Due to this, these techniques are unable to produce an optimal feature subset.

Multi-Cluster Feature Selection (MCFS) technique is proposed in [11]. This technique selects those attributes which preserves multi-cluster structure of data. To capture the multi-cluster structure of data, this technique uses multiple eigenvectors of graph Laplacian to define affinity matrix of data points.

To improve the performance of unsupervised feature selection technique, semi-supervised feature selection is proposed in [12]. In this technique, small amount of labelled data is used as training data to provide additional information.

The feature selection techniques are further classified in 3 categories based on the section strategies:

1) *Filter:* Filter method evaluates the usefulness of feature using various techniques such as: mutual information, the point wise mutual information, Pearson product-moment correlation coefficient.

P. Mitra, C. Murthy, and S. K. Pal [3] described unsupervised feature selection technique based on maximum information compression index measure. This measure is used to measure feature similarity.

2) *Wrapper:* This technique is applied in supervised learning method. This technique selects the features based on given classifier. As per distinct classifier and dataset feature subset selection strategy is evaluated in [1]. In [1] a tailored approach is proposed which can be customized with respect to the classification algorithm and dataset in any domain. These wrapper methods are computationally expensive and time consuming method. But this technique provides high classification accuracy as compared to the filter technique.
3) *Embedded:* Embedded Approach is specific learning algorithm and performs feature selection in the training process. It uses model construction process. The model construction is done using LASSO algorithm. This is linear model construction technique. Least Angle Regression (LARS) technique is proposed for model construction [5]. LARS calculates all possible Lasso estimates for a given dataset with less computer time.

B. Feature Extraction

Feature extraction technique transforms the high dimensional dataset to non-redundant low dimensional dataset. This technique transforms the D dimensional space to relevant reduced d dimensional space using linear or nonlinear combinations of features.

The feature extraction is also classified in supervised and unsupervised categories based on availability of information of class labels.

Maximum Margin Criterion (MMC) is the supervised feature extraction technique discussed in [7]. This includes linear as well as non linear feature extraction. MMC maximizes the average margin between classes after feature extraction. This technique is also applicable small sized Sample dataset .

To work with small sample sized dataset, Angular Liner Discriminant Embedding (ALDE) is proposed in [15]. This is again a supervised learning strategy. This technique calculates the cosine of angles to generate new within-class and between-class scatter matrices. Using these matrices, this technique performs balancing in within class and between-class scatters for feature extraction.

In unsupervised feature selection technique Principal Component Analysis (PCA) [8] is important techniques for linear space transformation. PCA is statistical process. It uses an orthogonal transformation to convert set of attributes to the correlated variables. In PCA each generated variable is combination of all original attributes and hence it is difficult to analyze the results of PCA. Hence, sparse PCA is used is to overcome this problem.[8] This technique evaluates the coefficients using regression-type optimization problem. It uses lasso constraint on the coefficients.

R. O. Duda, P. E. Hart, and D. G. Stork described Independent Component Analysis (ICA) technique in[2].This is a statistical technique to transform the data linearly to generate components that maximally independent from each other. ICA technique uses estimation of a latent variable model.

The study of dimensionality reduction techniques such as feature selection and feature extraction are done independently. If theses 2 techniques applied in symmetric faction, system may lead to better result generation. Feature set may include both types of feature like individual features and compound features complementary to each other. Sreevani and C. A. Murthy[1] proposed a technique to bridge the gap between these 2 techniques. This technique provides simultaneous study of feature selection and feature extraction techniques. For supervised feature extraction process Normalized Projection Error -Npe technique is used. It uses pair of features to remove redundancy. In this technique, to remove redundant features, only the combinations of two features are used to clubbed together.

III. CONCLUSION

In most of the existing work strategies feature selection and feature extraction are studied independently. To achieve better result accuracy simultaneous study of these two techniques is required. To bridge the gap between these two techniques and generate a symmetric way of execution is required. Such technique will be able to provide feature subset containing combination of selected and extracted features. Feature extraction strategy should be able to generate unique features by combining two or more features simultaneously.

IV. ACKNOWLEDGMENT

Authors would like to thanks Prof. Dr. K. N. Nandurkar, Principal and Prof. Dr. S. S. Sane, Head of Department of Computer Engineering, K.K.W.I.E.E.R., Nashik for their kind support and suggestions. We would also like to extend our sincere thanks to all the faculty members of the department of computer engineering and colleagues for their help.

REFERENCES

[1] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial intelligence, vol. 97, no. 1, pp. 273–324, 1997.
[2] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," Neural networks, vol. 13, no. 4, pp. 411–430, 2000.
[3] P. Mitra, C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 3, pp. 301–312, 2002.
[4] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," Machine learning, vol. 53, no. 1-2, pp. 23–69, 2003.
[5] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani et al., "Least angle regression," The Annals of statistics, vol. 32, no. 2, pp. 407–499, 2004.
[6] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in Advances in neural information processing systems, 2005, pp. 507–514.
[7] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," Neural Networks, IEEE Transactions on, vol. 17, no. 1, pp. 157–165, 2006.
[8] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," Journal of computational and graphical statistics, vol. 15, no. 2, pp. 265–286, 2006.
[9] W. Sheng, X. Liu, and M. Fairhurst, "A niching memetic algorithm for simultaneous clustering and feature selection," Knowledge and Data Engineering, IEEE Transactions on, vol. 20, no. 7, pp. 868–879, 2008.
[10] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection." in AAAI, vol. 2, 2008, pp. 671–676.
[11] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010, pp. 333–342.
[12] Z. Xu, I. King, M. R.-T. Lyu, and R. Jin, "Discriminative semi supervised feature selection via manifold regularization," Neural Networks, IEEE Transactions on, vol. 21, no. 7, pp. 1033–1047, 2010.
[13] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 3, pp. 619–632, 2013.
[14] D. Wang, F. Nie, and H. Huang, "Global redundancy minimization for feature ranking," Knowledge and Data Engineering, IEEE Transaction on, 2015.
[15] S. Liu, L. Feng, and H. Qiao, "Scatter balance: an angle-based supervised dimensionality reduction," Neural Networks and Learning Systems, IEEE Transactions on, vol. 26, no. 2, pp. 277–289, 2015.
[16] Sreevani and C.A. Murthy, "Bridging Feature Selection and Extraction: Compound Feature Generation," IEEE Transactions on Knowledge and Data Engineering, Vol. 29, No. 4, pp. 757 - 770 , April 2017.