Stratified Feature Ranking Method with Clustering For High-Dimensional Data

Dipali K. Pawar, Archana S. Vaidya

Abstract — Different feature selection methods have been used for handling high-dimensional data. But, mostly these feature selection methods, the selected features are highly correlated. These features are redundant and have similar properties. Here proposed method for rank features. In this approach first identify feature clusters by using a subspace feature clustering algorithm and then calculating weight for each attribute in each cluster for ranking. The ranked feature list is generated which contains diverse and informative high rank features.

Index Terms —Clustering algorithms, Data Mining, feature selection methods

I. Introduction

Feature selection plays important role in handling high-dimensional data. It is also known as attribute or variable selection. This is method of select set of important feature from original set. Without losing of information it could remove redundant and irrelevant features. Overfitting and curse of dimensionality are the big challenges in supervised learning algorithms.

The filter method is process of selecting features that independent of machine learning algorithm. In this methods correlation estimation approach used to select relevant features. Wrapper methods select optimal feature subset by using given predictor. But these methods are time-consuming because training base classifier used for selecting features. Embedded methods are combination of both methods which is generally related to specific learning algorithms.

Here proposed method used for rank features. In this approach first identify feature clusters by using a subspace feature clustering algorithm and then calculating weight for each attribute in each cluster for ranking. The ranked feature list is generated which contains diverse and informative high rank features.

II. Related Work

H. Hotelling [3], presented feature selection methods. It is also known as attribute or variable selection. This is method of select set of important feature from original set. Without losing of information it could remove redundant and irrelevant features [4]-[6].

X. He et al. [7] described method which based on observations in classification problems. A Laplacian score method evaluating the features by locality preserving power. This approach performed in both supervised and unsupervised way.

T. Wu et al. [8], they introduced a tensor spectral co-clustering method. This method based on random walk model. Here simultaneously cluster rows, columns and three modes of non-negative tensor data i.e. sparse, non-square, and asymmetric.

Peng et al. [9] presented principle of redundancy called it as mRMR. By using maximal statistical dependency it select optimal features. That result in improvement of classification and feature selection accuracy.

A. Das et al. proposed spectral regularizes [10], for selecting diverse and noiseless feature sets. The local search algorithms and greedy algorithms are used for efficient approximation of features sets. These construct submodular spectral regularizers for selection of diverse features that capture diversity.

Kong et al. [11], they proposed efficient re-weighting method. This method analyses convergence property and solve new formulation. This system extended to handle non-smooth loss function and non-convex property. This uncorrelated feature selection method is very time consuming process.

T. George et al. [12] presented weighted co-clustering algorithm. This approach used with collaborative filtering frameworks. These frameworks build by designing parallel and incremental version of co-clustering. It based on loss function other than squared error function.

Hartigan [13] presented new clustering method. This method performs direct clustering of data matrix. Partitional co-clustering algorithms used for clustering of data matrix. This is an iterative partition process which clusters a data matrix into disjoint co-clusters.

Banerjee et al. [14] presented principle of minimum Bregman information (MBI) for coclustering. In BBAC simultaneously generalize the standard least squares and the maximum entropy. This algorithm based on optimal matrix approximation. But it cannot recognize noisy values that exist in high-dimensional data disadvantage of this method.

I. S. Dhillon et al. [15] introduced ITCC algorithm (information theoretic co-clustering). The joint probability distribution values of rows and column used data matrix.

III. System Overview

Figure 1, show proposed system design. Here important task is select top ranked feature from feature clusters. First clustering performed on dataset. And then weight is calculated for each attribute to the class of clusters. Then sort the features in descending manner in each feature cluster. And that arrange features in a descending manner from all feature clusters. Finally a select top ranked feature from all of this is an ultimate result. Here stratified subspace clustering and stratified feature ranking algorithm used for selecting top ranked features.

Here important task is select top ranked feature from feature clusters. First clustering performed on dataset. And then weight is calculated for each attribute to the class of clusters. Then sort the features in descending manner in each feature cluster. And that arrange features in a descending manner from all feature

clusters. Finally a select top ranked feature from all of this is an ultimate result. Here stratified subspace clustering and stratified feature ranking algorithm used for selecting top ranked features.

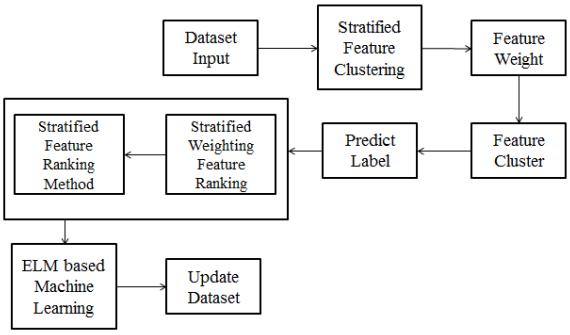

Figure 1. System Design

Stratified subspace clustering used for cluster features into feature cluster. Using stratified feature weighting method calculating weight for each attribute of that cluster. Then rank it in each feature cluster by using stratified feature raking method.

In Stratified Feature Clustering Algorithm, first initialize co-cluster centroid randomly. Then directly construct row cluster binary matrix with given class labels associated each data points in with the nearest centroid. This will divide the points into column clusters. Recalculate positions of centroids. Then calculate weight matrix which contain weight for each object of cluster. Repeat steps until there are no more changes in the membership of the data points.

In Stratified Feature ranking algorithm, compute norm of feature weight matrix and sort the features in ascending order. Then compute stratified feature ranking vector and sort in descending manner. Finally select top rank features from feature list as result.

IV. Results And Implementation

Here the given dataset table contain the three datasets. These four datasets are SRBCT, Leukemia, Cancer Dataset of SRBCT includes the 2308 genes with 83 instances. Dataset of Leukemia includes the 4026 genes with 62 instances. Dataset of Cancer includes the 24481 genes with 97 instances.

There are three evaluation indices i.e. Recall, precision, F1-measure are used to evaluate clustering results.

In figure 2, figure 3 and figure 3 show the performance evaluation on clustering results on SRBCT dataset.

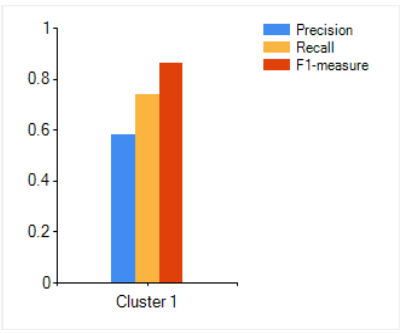

Figure 2. Precision, Recall, F1-measure calculation for cluster 1

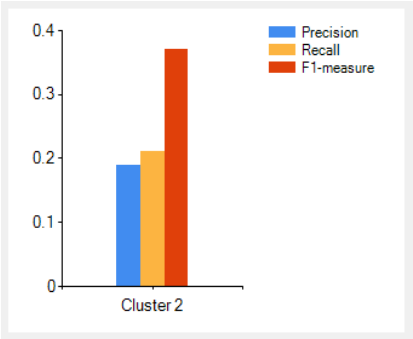

Figure 3. Precision, Recall, F1-measure calculation for cluster 2

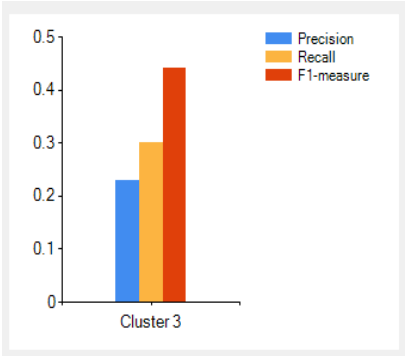

Figure 4. Precision, Recall, F1-measure calculation for cluster 3

In figure 5 show the accuracy comparison graph between proposed and existing system.

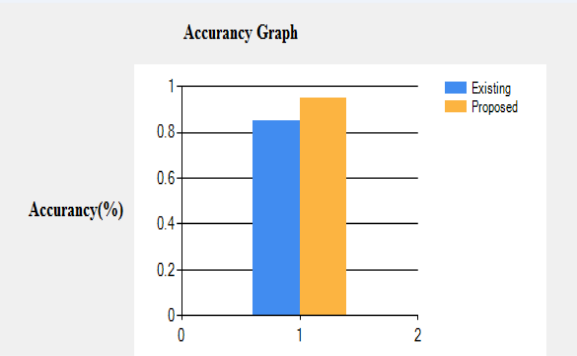

Figure 5. Accuracy Comparison graph between proposed and existing system

Conclusions

In this system the proposed method select top ranked features. Here, stratified subspace clustering algorithm used for clustering features into feature clusters. And importance of each feature is also identifying with respect to their cluster by using assign weights to that feature. Stratified feature weighting method calculating weight for each in feature in each cluster. After that, SFR used for rank the feature in the each cluster. Finally select top most features from all clusters i.e. final ranked feature list. Ensemble learning approach is used to train the cluster. It will increase clustering performance.

Acknowledgment

I have a tremendous pleasure in presenting the project "Stratified Feature Ranking Method with Clustering for High-dimensional Data" under the guidance of Prof. A. S. Vaidya and PG coordinator Prof. A. S. Vaidya. I am really obligated and appreciative to Head of the

Department Dr. D. V. Patil for their significant direction and consolation. I might likewise want to thank the Gokhale Education Society's R. H. Sapat College of Engineering, Management Studies Research, Nashik-5, India for giving the required offices, Web get to and vital books. At last I must express my sincere heartfelt gratitude to all the Teaching Non-teaching Staff members of Computer Department of GESRHSCOE who helped me for their important time, support, remark, thoughts.

REFERENCES

[1] R. Chen, N. Sun, X. Chen, M. Yang, and Q. Wu, "Stratified Feature Ranking method for High-Dimensional data," IEEE Access, vol. 6, 2018.

[2] Dipali K. Pawar, Archana S.Vaidya, "A Review on Stratified Feature Ranking Method," IJSRCSAMS vol. 8, issue 1, Jan. 2019.

[3] H. Hotelling, "The Selection of variates for use in prediction with some comments on the general problem of nuisance parameters," Ann. Math.Statist., vol. 11, no. 3, pp. 271-283, 1940.

[4] G. Chandrashekar and F. Sahin, "A survey on feature selection methods,"Comput.Elect. Eng., vol. 40, no. 1, pp. 16-28, Jan. 2014.

[5] Y S. H. Huang, "Supervised feature selection: A tutorial," Artif. Intell. Res., vol. 4, no. 2, p. 22, 2015.

[6] Y. Saeys, I. Inza, and P. Larraaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19,pp. 2507-2517, 2007.

[7] X. He, D. Cai, and P. Niyogi, ``Laplacian score for feature selection,'' in Proc. Adv. Neural Inf. Process. Syst., 2005, pp. 507-514.

[8] T. Wu, A. R. Benson, and D. F. Gleich, "General tensor spectral coclustering for higher-order data,'' in Advances in Neural Information Processing Systems 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2016, pp. 2559-2567.

[9] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226-1238, Aug. 2005.

[10] Das, A. Dasgupta, and R. Kumar, "Selecting diverse features via spectral regularization,'' in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1583-1591.

[11] D. Kong, J. Liu, B. Liu, and X. Bao, "Uncorrelated group lasso,'' in Proc.30th AAAI Conf. Artif. Intell., 2016, pp. 1765-1771.

[12] T. George and S. Merugu, "A scalable collaborative filtering framework based on co-clustering,'' in Proc. 5th IEEE Int. Conf. Data Mining, Nov. 2005, p. 4.

[13] J. A. Hartigan, "Direct clustering of a data matrix,'' J. Amer. Statist. Assoc., vol. 67, no. 337, pp. 123-129, 1972.

[14] Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha, "A generalized maximum entropy approach to bregman co-clustering and matrix approximation,'' J. Mach. Learn. Res., vol. 8, pp. 1919-1986, Aug. 2007.

[15] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic coclustering,'' in Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2003, pp. 89-98.

AUTHORS' DETAIL

Dipali K. Pawar
Student,
Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering Management Studies and Research, Nashik-5, India
Email: dkpawar5@gmail.com

Archana S. Vaidya
Assistant Professor**,**
Department of Computer Engineering, Gokhale Education Society's R. H. Sapat College of Engineering Management Studies and Research, Nashik-5, India
Email: archana.s.vaidya@gmail.com

CITE THIS ARTICLE AS :

Dipali K. Pawar, Archana S. Vaidya, " Stratified Feature Ranking Method with Clustering For High-Dimensional Data", International Journal of Technology and Science, vol. 6, Issue. 2, pp. 6-8, 2019