Effective Methodology of Feature Selection on Feature Consisting Group Structure

Dekate Priya A[1] Hande Kapil N[2]
[1]Student [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]PBCOE, Nagpur India

Abstract— Feature selection has become an interesting research topic in recent years. It is an effective methodology to tackle the large amount of data. It is always better to select features from group rather than selecting feature individually. This help to increases accuracy and decreases computational time. This research paper introduced a new method for feature consisting group structure called efficient group feature selection (EGFS). The experimental computations of EGFS demonstrate the advantage of using feature selection method. Also it gives good result in selecting optimal feature subset from a group of features.

Key words: Feature Selection, Group Structure, Redundant, Classification

I. INTRODUCTION

Database is a collection of large number of data and searching the hidden information from large database is the main task of data mining. High dimensionality is become an expletive for data mining which consist problem while training the data. The expletive of dimensionality can be normalized by using feature selection. The searching an optimal solution subset from original feature set is a feature selection. The main task of feature selection is to select the feature from large number of variable set. The feature selection is enforced to minimize the variable. The goal of feature selection is to select important feature that is useful for targeted output. It reduces the irrelevant and unused feature from original feature sets. The meanings of relevant feature are those that provide meaningful information and vice versa. So feature selection is an important process in effective methodology of feature data sets. There is some important features of feature selection, It is easy to understand, provides facility of data visualization, increases data predictability. It also reduces the unwanted features because of that reduces the storage requirement, reduces the training and processing time. Feature selection can be used in many applications like gene selection, intrusion detection, image retrieval, text categorization, DNA microarray analysis, information retrieval, etc. It produces the efficiency in feature selection, increases the accuracy. The feature selection uses the different algorithm such as weighted mutual algorithm and sparse group lasso algorithm. In weighted mutual algorithm, according to their weight calculated the threshold value and with the threshold value the data are reduced. Another algorithm is sparse group lasso. This algorithm consist the data from selected feature and again minimized the data from group of selected feature. These help to increases the accuracy and decreases the computational time. Because it is always better to select the data in groups rather selecting from complete database. It is more suitable to select the data from group rather than individual level. We address the problem of selecting the features from group in real world application. Its common example is multifactor analysis of variance (ANOVA). ANOVA is a set of learning model applied to test the difference among group and correlated procedure that is variation among and between the groups. Grouping can be introduced into model to take benefits of important knowledge that is significant. Example like in gene expression analysis, satellite images, etc.

In this paper we introduced new method named as efficient group feature selection. Tis consist of two stage is to select the data within the group variable selection that selects important features within the groups. In this stage each feature is evaluated individually. Another stage is between group variable selections. In this stage it select feature from already reduced feature. This is used to remove the data redundancy.

II. OBJECTIVES

1) To reduce data in intra group because it very difficult to sorting large amount of data.
2) Intra group data are further divide and combine in a inter group.
3) To perform feature selection in groups and to determine optimal subset.
4) Performs the selection in within and between the groups.
5) Provide better understanding of model to the user, and decision making.

III. METHODOLOGY

The overall Design approach is basically divided into several steps. The first step is input data sets is used which is available from UCI machine learning repository datasets for feature selection. The three datasets are used i.e Ionosphere, Wdbc, Statlog (heart).The data sets which is being used have not provide any group information. Creating the group of features is the second steps. The group of features is created by dividing the feature randomly. The size of group is depending on the user choice. This step gives the group of feature.

Next step is performing feature selection on group of features, We focus on the problem where feature possessing some group structure, to solve this problem we propose a framework for group feature selection it consist of two stages: intra group feature selection and inter group feature selection. The discriminative features are evaluated in intra group feature selection. The features are evaluated one at a time in this stage and the features are selected within the group. After intra group feature selection all the features are re-evaluated to find the correlation between the group to find an optimal subset, namely as inter group selection. This step gives the optimal subsets of features. The validation is needed on the selected feature in order to evaluate whether the features are optimal or not classification is required.

An Effective Group Feature Selection technique (EGFS) is proposed during this section for group of feature. From domain knowledge we will able to get a bunch structure or a user can provide the group size that facilitate to scale back the potency of time. Our aim is to search out Associate in nursing optimum set from a bunch. An Effective Group Feature Selection framework consists of 2 stages: among group variable selection and between group variable choice. Initially we'd like to make a bunch of options, the cluster of feature generated by at random dividing the feature area among group variable choice. Our aim to search out discriminative feature, the feature square measure determined one at a time. Once among group variable selection all the options square measure re-evaluated to search out the correlation between the teams to search out Associate in Nursing optimum set, namely as between group variable selection.

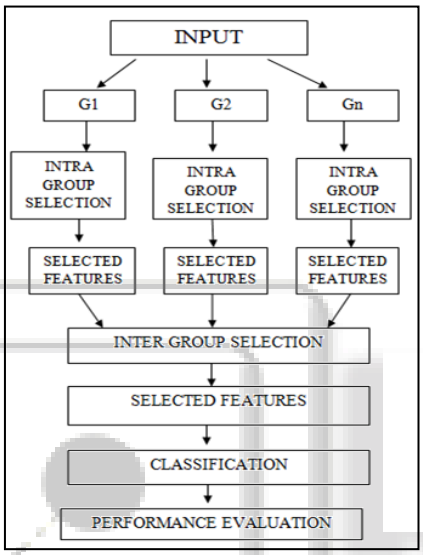

Fig. 1: Flow of work for efficient group variable selection approach.

IV. EXPERIMENTAL RESULT

A. Datasets Description

Our first step is input of data sets. For the feature selection three data set is used to further verify the effectiveness of our method, the datasets are ionosphere, Wdbc, Statlog (heart) are available from UCI datasets[12]

- Ionosphere data set is a radar dataset, which consist of 351 instance, 34 attribute and 2classes this dataset shows some structure of good ionosphere and bad ionosphere. Mostly in numeric form.
- wdbc is referred as Wisconsin diagnostic breast cancer consist 30 attributes and 569 instance. There are two classes malignant and benign.
- The Stat-log dataset is a heart disease data set consists of 13 attribute and 270 instances there are two classes. In all three dataset there is no any grouping information is given.

Data sets	No of classes	No of instance	No of features
Ionosphere	2	351	34
Wdbc	2	569	30
Statlog(heart)	2	270	13

Table 1: Data sets & No of classes

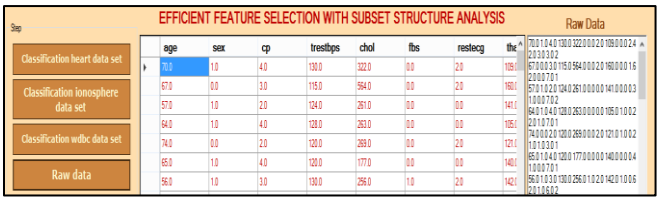

Fig. 2: Preprocessing Of Data

The first step is to collect the information from datasets which is available in UCI machine learning repository. In above figure we have done preprocessing where the raw data are arrange in proper format. Proper format means the white space, non-numeric data, redundant data are removed and data are arrange properly.

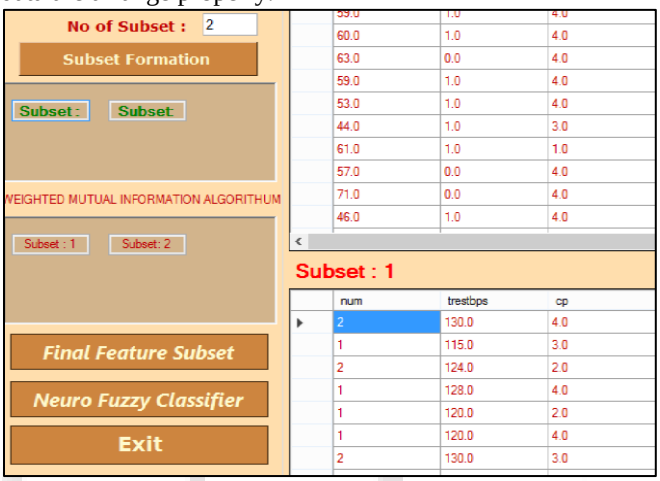

Fig. 3: Formation of subset

In above fig shows the formation of subset. In this case the entry get from user how many groups he want to make. According to user entry the subsets are formed.

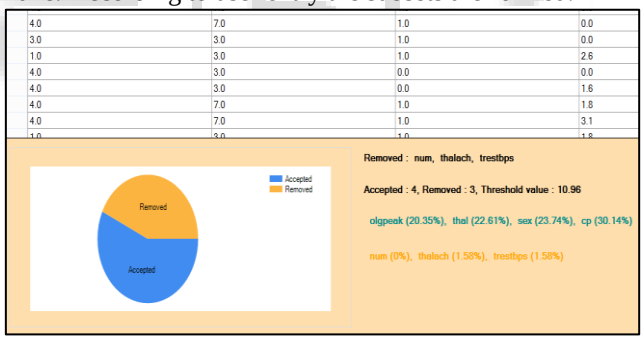

Fig. 4: Intra group selection method

It finds the correlation among the features and selects the discriminative features. In this stage each features is evaluated individually and assign the scores to the feature. . For intra group feature selection the weighted mutual information is applied.

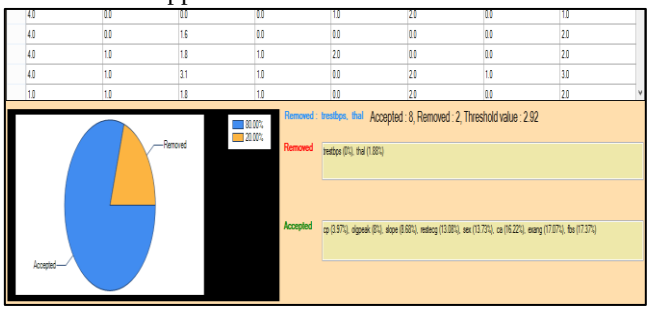

Fig. 5: Inter Group Selection method

In inter group selection it select relevant feature from every group but there may be probability of consisting

redundant feature so to remove redundant feature the between group feature selection re-evaluated the entire feature and find the optimal subset. In above fig we show that the from the original features how many features are selected and how many are removed. And find out the percentage of threshold.

V. RESULT EVALUATION AND ANALYSIS

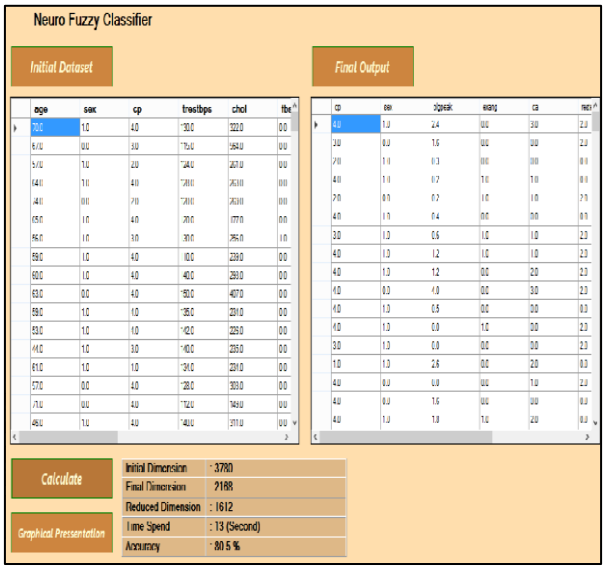

Fig. 6: Classification by NFC

After performing the group feature selection, we get the optimal feature subset. The validation is required for this we perform the classification. For classification we apply Neuro-fuzzy classifier (NFC). The neuro-fuzzy classifier is the combination of both neural network and fuzzy system. With NFC we can see that the initial dimension is3780, the final dimension is 2168 and reduced dimension is 1612. The accuracy is 80.5% according to NFC.

Data Sets	No. of instance	No of features	Selected features by EGFS
Inosphere	351	34	18
Wdbc	569	30	12
Statlog(Heart)	270	13	08

Table 2: Experimental Result by Performing Efficient Group Feature Selection Method on UCI Data Sets

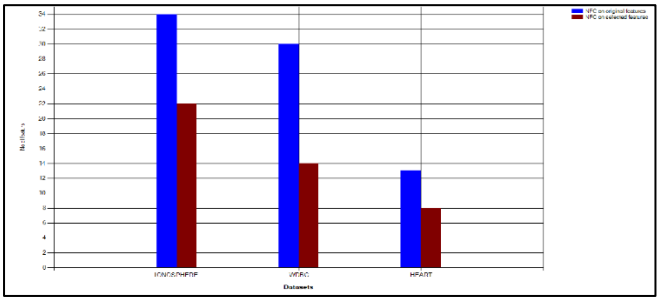

Fig. 7: Graphical representation of Dataset

In above fig shows the graphical representation of all three dataset. We can see that from all of three dataset like ionosphere, wdbc and heartdata. The ionosphere give the the more accuracy than the two dataset.

VI. CONCLUSION

We have presented effective methodology of feature selection on feature consisting group structure. Mainly we divided the effective group structure in two stages. The first stage is within group variable selection and another stage is between the group variable. In intra group variable selection it uses the weighted mutual information algorithm. It finds correlation among the features and selects the discriminative features and assign scores to the features. In inter group variable selection it selects relevant feature from every group. It uses the sparse group lasso algorithm in inter group selection. The datasets are collect from the UCI repository. The three datasets are used Inosphere, Wdbc, and Statlog(heart). Feature selection is used to select the relevant feature from the targeted output. This increases the accuracy and decreases the computational time.

REFERENCE

[1] X. Wu, X. Zhu, G.Q. Wu, and W. Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2014.
[2] Guyon and A. Elisseeff. "An introduction to variable and feature selection," Journal of Machine Learning Research, 3:1157–1182, 2003.
[3] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," The Journal of Machine Learning Research, vol. 5, pp. 1205–1224, 2004.
[4] Haiguang Li, Xindong Wu, Zhao Li, Wei ding"Group feature selection with streaming features," IEEE 13th international conference on data mining. 2013.
[5] Jennifer G. Dy, Carla E. Brodley "Feature Selection for Unsupervised Learning," Journal of Machine Learning Research, 845–889.2004.
[6] H. Liu and H. Motoda, "Computational methods of feature selection," CRC Press, 2007.
[7] Daphne Koller, Mehran Sahami, "Toward Optimal Feature Selection," Computer Science Department, Stanford University, Stanford, CA 94305-9010.1996.
[8] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," Journal of the Royal Statistical Society, vol. 68, no. 1, pp. 49–67, 2006.
[9] Meier L., Van De Geer, S., & Buhlmann P. "The Group Lasso for Logistic Regression," J. Roy. Stat. Soc.B, 70, 53–71.2008
[10] Bach F. R. "Consistency of the group lasso and multiple kernel learning," Journal of Machine Learning Res. 9 1179–1225.2009.
[11] N. Meinshausen and P. Buhlmann "High-dimensional graphs an variable selection with the lasso," Annal of Statistic.,34 1436–1462.2006.
[12] Zhao.P. and Yu.B. "On model selection consistency of Lasso,"Journal of Machine Learning," Res. 7 2541–2563. 2006
[13] H. Zou "The adaptive lasso and its oracle properties".J. Amer. Statist. Assoc.2006.
[14] Zhang, C.-H. and Huang, J. "sparsity and bias of the LASSO selection in high-dimensional linear regression," The Annals of Statistic. 36 1567–1594.2008.
[15] Lei Yuan, Jun Liu, and Jieping Ye, "Efficient method for overlapping group lasso," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 9, September 2013.
[16] S. Xiang, X. T. Shen, and J. P. Ye, "Efficient sparse group feature selection via nonconvex optimization," in ICML, 2012.

[17] Seyoung Kim, Eric P. Xing, "Tree-Guided Group Lasso for Multi-Task Regression with Structured Sparsity," in ICML, 2010.
[18] Roth.V. and Fischer. B. "The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms," In ICML,pp. 848–855, 2008.
[19] Hanchuan Peng, Fuhui Long, and Chris Ding, "Feature selection based on mutual information criteria of Max dependency, max relevance and min redundancy," IEEE transactions on pattern analysis and machine intelligence, vol. 27, no. 8, august 2015.
[20] Erik Schaffernicht and Horst-Michael Gross "Weighted Mutual Information for Feature Selection"21 international conference on artificial neural network(ICANN 2011),Espoo, Finland, LNCS 6792,pp. 181-188, Springer 2011.
[21] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "A sparse-group lasso," Journal of Computational ssand Graphical Statistics. May 2011.
[22] S. Xiang, X. T. Shen, and J. P. Ye, "Efficient sparse group feature selection via non convex optimization," in ICML, 2012
[23] UCI Machine Learning Repository [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html.
[24] Prashant Borkar, M.V.Sarode,Latesh Malik, "Modality of adaptive neuro-fuzzy classifier for acoustic signal-based traffic density state estimation employing linguistic hedges for feature selection" , Int. J. Fuzzy Syst, Springer 2015

