www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 5 May 2017, Page No. 21307-21308
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i5.19

Effective Feature Selection for Feature Possessing Group Structure

Yasmeen Sheikh, Prof. S. V. Sonekar

(yaso.yasmeen@gmail.com)

J.D College of Engineering, Nagpur.

Abstract— Feature selection has become an interesting research topic in recent years. It is an effective method to tackle the data with high dimension. The underlying structure has been ignored by the previous feature selection method and it determines the feature individually. Considering this we focus on the problem where feature possess some group structure. To solve this problem we present group feature selection method at group level to execute feature selection. Its objective is to execute the feature selection in within the group and between the group of features that select discriminative features and remove redundant features to obtain optimal subset. We demonstrate our method on data sets and perform the task to achieve classification accuracy.

I. INTRODUCTION

Searching hidden information and pattern from very large database is the task of data mining. High dimensionality has made data mining a tedious work which. This curse of dimensionality can be minimizing by using feature selection. The method of searching a variable subset from actual feature set is a feature selection. The application in which there are large numbers of variable the feature selection is enforced to minimize the variable. The actual aim of feature selection is to search a relevant feature that is useful for target output. It removes the irrelevant and redundant feature from original feature sets. Relevant feature are those features that provide useful information and redundant feature are those that is not useful. So feature selection is an important process in efficient learning of large multi feature data sets. There are some potential advantages of feature selection. It facilitate data visualization, it also increases data predictability and understanding. Feature selection also helps to reduce the measurement and storage requirement, reduces processing time. Feature selection can be used in many applications such as gene selection, intrusion detection, image retrieval, DNA microarray analysis etc. It enhances the literature efficiency, increases anticipating certainty and help to minimizing learned result complexity. The feature selection algorithm generates an output as a subset of feature or by measuring their utility of feature with weights. The assessment of features in feature selection method can be in various forms such as consistency, dependency, separability, information and training model which are generally occurred in wrapper model.

Previously feature selection methods were evaluating or selecting feature individually and avoids selecting feature from groups. It is good to select features from group rather than selecting features individually. This increases accuracy and decreases computational time of data. Therefore in some situation finding a vital feature equivalent to the evaluating a group of feature. The group of variable must take an advantage of group structure while selecting important variable.

Features can be selected from the available feature set through many feature selection methods. However, they always tend to select features at individual level with small percentage and more preferably than the group structure. When group structure exists, it is more convenient to select features with small percentage at a group level rather than individual level. We address the problem of selecting the features from groups so we consider the problem that feature possesses some group structure, which is potent in many real world application and its common example is Multifactor Analysis of Variance (ANOVA). It is a set of learning model applied to examine the difference among group and correlated procedures that is variation among the groups and between the groups

Group structure can appears in different modelling goal for multiple reasons. Grouping can be introduced to take benefits of prior knowledge that is significant. Example like in gene expression analysis, the matches to the same categories can be known as group. In data analysis it is convenient to consider about the group structure. In some conditions, the individual features in group may or may not be much useful, if this features are useful then we are not interested in selecting an important feature in this case group selection is our objective. But if individual features are useful then we are interested in selecting an important features and important group.

This paper develops an efficient group feature selection methods, the main thing is that they are with group structure. In this paper, we propose a new group feature selection method named as efficient group variable selection (EGVS). This consists of two stages, within group variable selection stage that select discriminative features within the group. In this stage each feature is evaluated individually. After an estimation and sparsity an error of prediction of groups within group selection all the features are re-evaluated so far to remove redundancy this stage is known as between group variable selection.

The paper is constructed as follow, section II describe various feature selection approaches and provides review on existing literature on underlying group structure such as group lasso.

II. FEATURE SELECTION METHODS

The feature selection method is divided into three category based on their label information and label information method is used most commonly used. In supervised feature selection technique there are difficulties in acquiring the data label. In recent year unsupervised feature selection has more attention.

Unsupervised feature selection generally selects features that preserves the data similarity of multiple structure whereas semi supervised feature selection makes use of label information and multiple structures related to labelled data and unlabelled data. There are 3 types of methods for feature selection, filter method, wrapper method, and embedded method. Filter method does not use any learning algorithms for measuring feature subsets. This method is fast and efficient for computations. Filter method may fail to select the features that are not beneficial for themselves but can be very beneficial when unite with other features. Wrapper method use learning algorithms and search for optimal attribute subset from original attribute set which discover relationship between relevance and optimal data subset selection. The embedded method is a combination of wrapper methods. This decreases the computational cost than wrapper method and captures dependencies. It searches locally for features that allow better discrimination and the relationship between the input feature and the targeted feature. It involves the learning algorithm which is used to select optimal subset among the original subset with different cardinality. Many analysts have focuses on a feature that contain certain group structure such as group lasso. The group lasso applies L2 norm of the coefficient joined in the penalty function by a collection of features. An extended form of Lasso is group lasso. It simplifies the standard lasso technique. Many authors have studied the various property of group lasso structure by building the many approaches of lasso. Yuan and Lin have demonstrated the group Lasso used to solve the problem of convex optimization that consider for size of group and applied Euclidean norm. This process acts as a lasso at group level, whereas if the sizes of group are same, then it is reduced to the lasso. The author has proposed the method for adjusting the group lasso that considers the model matrices in each groups are orthonormal. Whereas in non-orthonormal case, it uses the rigid regression to handle the groups of variable. Mieere [9] proposed the method for logistic regression to extend the group lasso. Suhrid Balakrishnan and David Madigan [10] unite the idea from group lasso Yaun and Lin [8] and fused Lasso. The Bakin [11] proposed the group Lasso and computational algorithm. This method related group selection method and algorithm are further developed by Yuan and Lin [8]. Composite absolute penalty (CAP) approach developed by Zhao Rocha [12] is same as group lasso but instead of using L2 norm it uses L1 norm the group information in CAP method consider the group lasso and combine the group penalty for Lr0 norm. It does not imply any information but the grouping information. CAP method includes the group Lasso as special case.

V. CONCLUSION

We have presented efficient group variable selection for group of features. Method focuses on the problem where feature comprise some group structure. We also provide the literature reviews on existing method. We divided the efficient group variable selection into two stages, i.e., within group variable selection and between group variable selections. In within group variable selection uses mutual information and introduces the sparse group lasso to minimize the redundancy in between group variable selection. The within group variable selection effectively select discriminative feature, in this step each feature is evaluated individually. Between group selection controls the compactness and revaluate the features. We have also demonstrated the experiment on several UCI benchmark data sets. This increases the classification accuracy and shows the effectiveness of our method.

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] Suhrid Balakrishnan and David Madigan, "Finding predictive runs with LAPS" 7TH IEEE conference on Data mining, 2007.

[11] S.Bakin. "Adaptive regression and model selection in data mining problems," Ph.D. thesis, Australian National Univ., Canberra. 1999.

[12] Zhao, P., Rocha, G. and Yu, B. "The composite ab-solute penalties family for grouped and hierarchical variable selection," Annals of Statistics , Vol. 37, No. 6A, 3468-3497.2009.