Feature Selection using Clustering approach for Big Data

Harshali D. Gangurde
Computer Engineering, Department,
MET IOE, BKC, Adgaon, Nasik, Maharashtra, India.

ABSTRACT
Feature selection has been a productive field of research and development in data mining, machine learning and statistical pattern recognition, and is widely applied to many fields such as, image retrieval, genomic analysis and text categorization. Feature selection includes selecting the most useful features from the given data set. The feature selection involves removing irrelevant and redundant features form the data set. The feature selection can be efficient and effective using clustering approach. Based on the criteria of efficiency in terms of time complexity and effectiveness in terms of quality of data, useful features from the big data can be selected. Feature selection reduces the computational complexity of learning and prediction algorithms and saves on the cost of measuring non selected features. The feature selection can be done using the graph clustering approach based on theoretic graph. The most relevant features are selected from the cluster for the relevant target class. The features in every cluster are different and independent of the other**.**

Keywords
Feature selection, Clustering

1. INTRODUCTION
Data mining refers to extraction of hidden predictive information from voluminous data. Different data mining functionalities are used for selecting the most relevant data from the big data set. Clustering analysis is the task of grouping of objects (data) with similar features or attributes. Whereas, data pre-processing is used to improve efficiency in mining process i.e. to extract data from the voluminous data with required features and removing irrelevant, redundant feature subset. With the aim of choosing good feature subset, many feature subset selection algorithms are proposed like the Embedded, Wrapper, Filter and Hybrid approaches. Feature selection is an approach of identifying subset of features that are mostly related to the target class. The main objective of feature selection is to remove irrelevant and redundant features from big data and to increase the level of accuracy and reduce dimensionality of the data. The Embedded method is concerned with feature selection is usually specific to given learning algorithms. This method may be more efficient than the other categories. The Wrapper methods determine the goodness of the selected features by predetermined learning algorithm. The Filter methods do not depend on learning algorithms. These methods have low computational complexity. The Hybrid methods are fusion of filter and wrapper methods. It uses a filter method to reduce search space. The Wrapper methods are computationally expensive and over fit on small training sets. The Filter methods are usually a good option for dataset having large number of features. Filter method is used in the clustering approach. The graph will be constructed for clustering as follows: A neighborhood graph of instances/feature is computed. Any edge in the graph that is much longer or shorter determined by some criterion, than its neighbors are deleted. The result obtained after deletion of edge is a forest and each tree in the forest represents a cluster. In the clustering approach minimum spanning tree (MST) based clustering algorithms is used, since it does not assume that data points that are grouped around centers or distinguished by a regular geometric curve. Feature selection is used in data mining to determine the tools and techniques available for reducing data size so that when it is provided as input it must be manageable for processing and analysis. Feature selection is concerned with cardinality reduction of data set, which involves a predefined cutoff on the number of features. It also implies selection of the choice of features, that either the data analyst may select or discard features based on their usefulness.

2. RELATED WORK
Large number of features can adversely affect the performance of inductive learning algorithms. Feature selection involves removing irrelevant data and redundant features. Removal of irrelevant and redundant features is important since they do not contribute to the anticipated accuracy. A feature selection technique forms a subset of the field of feature extraction. Feature extraction generates new features from the original features, whereas feature selection generates a subset of the features. The proposed Clustering approach falls into the second group. Feature selection research is mainly concerned with searching for relevant features. In Relief method each feature is weighed according to its liability to differentiate instances under different targets which is based on distance-based criteria function. Relief method is not effective in removing redundant features but it is highly correlated features are likely to be weighed highest. [2]. But Relief is an easy to use, fast and accurate algorithm even with dependent features and noisy data. Relief works by measuring the ability of an attribute in separating similar instances. Relief-F (RFF) is an extension to relief algorithms. It was extended by Kononenko so that it can deal with multi-class problems and missing values. The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature weighing vector to give more weight to features that discriminate the instance from neighbors of different classes. It is also improved to deal with noisy data and can be used for regression problems[15].Content Feature Selection(CFS) is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other [6].Fast Correlation Based Feature (FCBF) selection is a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis[16][17].Conditional Mutual Information Maximization(CMIM) iteratively selects features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked[18].Filter methods include Relief method and Focus methods; they normally evaluate the statistical

performance of the features over the data without considering the proper classifiers. The irrelevant features are filtered out before the classification process. Their main advantage is low computational complexity which makes them very fast. Their main drawback is that they are not optimized to be used with a particular classifier as they are completely independent of the classification stage. Butterworth et al. suggested to cluster features using a Barthelemy-Montjardet distance, and then makes use of the dendogram from the results of cluster hierarchy to choose the most relevant feature[12].Wrapper methods on the contrary evaluate feature subsets with the classification algorithm in order to measure their efficiency according to the correct classification[14]. Among algorithms widely used, we can mention Genetic Algorithm (GA) and Sequential Forward Selection (SFS) methods[14]. The computational complexity is higher than the one of filter methods but selected subsets are generally more efficient, even if they remain sub-optimal[14].

3. FEATURE SELECTION

The benefit of feature selection is that the identity of the selected features can provide insights into the nature of the problem at hand. The objective of using a feature selection technique is that the big data contains many redundant or irrelevant features and do not contribute to the efficient data mining. Irrelevant features provide no useful information with reference to any context and redundant features provide same information for the currently selected features. Feature Selection is useful in data analysis process, since it shows features which are important for prediction and relevant to the target class. Feature selection should provide useful results from the big data and must be able to detect and discard as much of the redundant and irrelevant information. Unlike other algorithms which use a given static modeling parameters to find clusters. The clustering approach algorithm generates clusters by dynamic modeling.

3.1 Clustering based Feature Selection

Irrelevant features and redundant features affect the accuracy of the learning machines. The feature selection using clustering includes following steps(see Figure1):

(i) Data preprocessing: concerned with redundant feature removal and finding relevant feature to the target class.

(ii)Construction of minimum spanning tree for the graph constructed for the data set after preprocessing it.

(iii) Formation of the clusters of the features.

(iv)Selection of features which are more relevant to the target class.

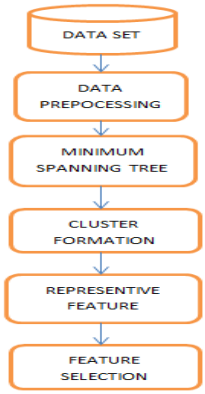

Fig1: Feature Selection Process.

3.2 Data Pre-Processing Technique

The data preprocessing techniques includes removal of redundant and relevant features. In order to more precisely introduce a clustering of features for feature selection, the proposed feature selection process involves irrelevant feature removal and redundant feature elimination. The traditional definitions of relevant and redundant features then provide the definitions based on variable correlation as follows. John et al. presented a definition of relevant features [2**].**

3.2.1 Symmetric Uncertainty

Mutual information measures how much the distribution of the feature values and target classes are different in terms of statistical independence .It is a nonlinear estimation of correlation between feature values and target classes. The *symmetric uncertainty* (*SU*) is derived from the mutual information by normalizing it to the entropies of feature values and target classes [4]. It has also been used to compute the goodness of features for classification by a number of researchers[5][6][7][8][9][10]. Therefore, symmetric uncertainty is the measure of correlation between either two features or a feature and the target concept chosen.

Symmetric uncertainty: It is defined as

$$SU(A, B) = 2 \times Gain(A|B)/E(A) + E(B) \quad (1)$$

Where,

1)E(A) is the entropy of a discrete random variable A. Suppose p(a) is the prior probabilities for all values of A,

E(A) is defined by:

$$E(A) = -\sum_{a \in A} p(a) \log_2 p(a) \quad (2)$$

2) Gain (A|B) is the amount by which the entropy of B decreases. It reflects the additional information about B provided by A and is called the information gain which is given by [11]

$$Gain(A|B) = E(A) - E(A|B)$$

$$= E(B) - E(B|A) \quad (3)$$

Where, E(A|B) is the conditional entropy which quantifies the remaining uncertainty of a random variable A given that the value of another random variable B is known. Suppose p(a) is the prior probabilities for all values of A and p(a|b) is the posterior probabilities of A given the values of B, E (A|B) is defined by:

$$E(A|B) = -\sum_{b \in B} p(b) \sum_{a \in A} p(a|b) \log_2 p(a|b) \quad (4)$$

After calculation of the entropy of features in the data set, n undirected graph is constructed. For big data set the graph is heavily dense. So the decomposition of this heavily dense graph is NP-complete problem. Hence, for given SU (A, B) the symmetric uncertainty of variables A and B, the Target-Relevance between a feature and the target concept C, the Feature-Correlation between a pair of features, the Feature Redundancy and the Representative Feature of features in cluster can be defined as follows:

3.2.2 Target-Relevance

The target relevance between the feature F_and the target concept $C$is denoted by $SU(F_a, C)$. If the SU between the feature F_a and C is greater than predetermined threshold than that feature is said to have strong relevance to the target class

3.2.3 Feature-Correlation

The correlation between any pair of features F_a and $F_{b(}F_a$, F_b

$\in F$and a≠b) is called the Feature-Correlation between two features and denoted by $SU(F_a, F_b)$.

3.2.4 Feature-Redundancy

Let $S = \{F_1, F_2, F_i ... < |F|\}$ be a cluster of features. If there exist a feature F_b, then,

$SU(F_b, C) \geq (F_a, C) \wedge (Fa, F_b) > SU(F_a, C)$ is always corrected and then Fa are redundant features with respect to the given F_b.

3.2.5 Representative-Feature

A feature $F_a \in S = \{F_1, F_2, Fk\}$ ($k<|F|$) is a representative feature of the cluster (i.e. F_a is a *Representative Feature)* if and only if: $F_a = \text{argmax}_{F_j \in S} SU(F_b, C)$. This means the feature, which has the strongest Target-Relevance, can act as a Representative-Feature for all the features in the cluster.

4. METHODOLOGY FOR FEATURE SELECTION USING CLUSTERINGAPPROACH

1) Compute the Target-Relevance $SU(F_a, C)$ value for each feature.

2) Calculate the Feature-Correlation $SU(F_a, F_b)$ value for each pair of features F_a and F_b ($F_a, F_b \in F$and a≠b). The Feature Correlation $SU(F_a, F_b)$ is the weight of edges.

3) Build a minimum spanning tree(MST), then first remove the edges whose weights are smaller than both of the Target-Relevance $SU(F_a, C)$ and $SU(F_b, C)$, from the MST.

4) Check for redundant features in MST by the property that for each pair of nodes (F_a, F_b), $SU(F_a, F_b) \geq SU(F_a, C) \vee (F_a, F_b) \geq (F_b, C)$.

5) By removing all the unnecessary edges, a forest is obtained. Each tree in a Forest represents a cluster.

6) The features in each cluster are redundant, so for each cluster we choose a representative feature whose Target-Relevancies the greatest.

5. CONCLUSION

A novel clustering approach is proposed for feature selection from big data. The formation of clusters reduces the dimensionality and helps in selection of the relevant features for the target class. The data preprocessing involved removes the redundant and irrelevant features. The formation of clusters obtained from minimum spanning tree reduces the complexity for the computation of feature selection. The proposed feature selection approach forms clusters of features and a representative feature can be chosen from it but this does not guarantee that the particular feature from the subset is more relevant to the target class. The classification of features which belongs to particular class can be done in this case. Thus, the feature subset obtained from can be given to Supervised Learners for classification purpose. Ensemble methods can be used for voting that gives the best feature for target relevant class. The main advantage of feature selection is that the identity of the selected features can provide insights into the nature of the problem at hand. Therefore, the feature selection is an important step in efficient learning of large multi-featured datasets.

5. REFERENCES

[1] QinbaoSong, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE transaction on Knowledge and Data Engineering 2013

[2] John G.H., Kohavi R. and Pfleger K., "Irrelevant Features and the Subset Selection Problem", Proceedings of the Eleventh International Conference on Machine Learning, pp 121-129, 1994.

[3] Koller D and SahamiM.,"Toward optimal feature selection", Proceedings of International Conference on Machine Learning, pp 284-292, 1996.

[4] Yu L. and Liu H.," Redundancy based feature selection for microarray data", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 737-742, 2004

[5] Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T., "Numerical recipes in C". Cambridge University Press, Cambridge, 1988.

[6] Hall M.A., "Correlation-Based Feature Subset Selection for Machine Learning," Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.

[7] Hall M.A. and Smith L.A., "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper", Proceedings of the Twelfth international Florida Artificial intelligence Research Society Conference, pp. 235-239, 1999

[8] Yu L. and Liu H., "Feature selection for high-dimensional data: a fast correlation-based filter solution", Proceedings of 20th International Conference on Machine Leaning, 20(2), pp. 856-863, 2003.

[9] Yu L and Liu H."Efficient feature selection via analysis of relevance and redundancy", Journal of Machine Learning Research, 10(5), pp. 1205-1224, 2004.

[10] Zhao Z. and Liu H., "Searching for interacting features", Proceedings of the 20th International Joint Conference on AI, 2007

[11] Zhao Z. and Liu H., "Searching for Interacting Features in Subset Selection" ,Journal Intelligent Data Analysis, 13(2), pp. 207-228, 2009.

[12] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., "On Feature Selection through Clustering", Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.

[13] Quinlan J.R., C4.5: Programs for Machine Learning. San Mateo, Calif: Morgan Kaufman, 1993

[14] Zhongzhe Xiao, Emmanuel Dellandrea, Weibei Dou, Liming Chen., "ESFS: A new embedded feature selection method based on SFS", Department of Electronic Engineering, Tsinghua University, Beijing, 100084, P.R.China.

[15] Kononenko I., Estimating Attributes:.,"Analysis and Extensions of RELIEF", Proceedings of the 1994 European Conference on Machine Learning, pp 171-182, 1994.,

[16] Pereira F., Tishby N. and Lee L.,"Distributional clustering of EnglishWords", Proceedings of the 31st Annual Meeting on Association forComputationalLinguistics, pp 183-190, 1993.

[17] Dash M., Liu H. and Motoda H., "Consistency based feature Selection"Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp.98-109,2000.

[18] Fleuret F., "Fast binary feature selection with conditional mutual Information",Journal of Machine Learning Research, 5, pp 1531-1555, 2004.