ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Graph Clustering and Feature Selection for High Dimensional Data

K. Jaganath[1], Mr. P. Sasikumar[2]

II ME (CSE), Selvam College of Technology, Namakkal, Tamilnadu, India[1]

Assistant Professor, Selvam College of Technology, Namakkal, Tamilnadu, India[2]

Abstract: Feature selection techniques are used to select important items in the transactional data values. The features are used for the classification process. Clustering techniques are used for the feature selection process. Graph based clustering techniques are used to group up the transactional data with similarity values. Correlation similarity measures are used to identify the relevant and irrelevant features. Features And Subspace on Transactions (FAST) clustering-based feature selection algorithm is used to cluster the high dimensional data and feature selection process. FAST algorithm is divided into two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature is selected from each cluster to form a subset of features. Features in different clusters are relatively independent. The clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. Minimum-Spanning Tree (MST) clustering method is adopted to ensure the efficiency of FAST. Feature subset selection algorithm is used to identify the features from the clusters. The feature selection process is improved with a set of correlation measures. Dynamic feature intervals can be used to distinguish features. Redundant feature filtering mechanism is used to filter the similar features. Custom threshold is used to improve the cluster accuracy.

I. INTRODUCTION

Data values are analyzed with interconnectivity relationships. Network is a widespread form of data consisting of a set of vertices and a set of edges that are connections between pairs of vertices. Many real-world networks possess an intrinsic community structure, such as large social networks, Web graphs, and biological networks. A community is typically thought of as a group of vertices with dense connections within groups and relatively sparse connections between groups as well. Since clustering is an important technique for mining the intrinsic community structure in networks, it has become an important problem in a number of fields, ranging from social network analysis to image segmentation and from analyzing biological networks to the circuit layout problem. Graph cluster based methods are used to select features from the clusters.

The FAST Clustering based Feature Selection framework is composed with the two connected components of irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset. The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the clusters.

A novel clustering-based feature subset selection algorithm is used for high dimensional data. The algorithm involves 1) removing irrelevant features, 2) constructing a minimum spanning tree from relative ones, and 3) partitioning the MST and selecting representative features. In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. The performance of the proposed algorithm is compared with those of the five well-known feature selection algorithms FCBF, ReliefF, CFS, Consist and FOCUS-SF.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

The system is enhanced with different correlation measures. The dynamic threshold mechanism is also used to improve the cluster accuracy levels.

Feature selection is applied to filter a subset of useful and independent features. Features in different clusters are relatively independent. Similar features are filtered from the cluster results using redundant feature filtering method. The features strongly related to target classes is selected from each cluster to form the final subset of features. Features are compared with reference to their similarity values. Distance measures are used for the similarity estimation process. Feature interval refers the similarity between the features. Features are filtered with reference to the feature similarity intervals. The system uses the different feature intervals cluster iterations. The intervals are selected with the support of cluster similarity measures.

Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation. Cluster accuracy is estimated with reference to the precision and recall factors. The transaction assignment for the cluster is estimated using the similarity measures. The similarity level is used to select the final cluster partition state. The threshold value is customized for the accuracy levels. The system collects the threshold value from the user with reference to the accuracy level requirements.

II. RELATED WORK

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because 1) irrelevant features do not contribute to the predictive accuracy and 2) redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s) of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features [9] yet some of others can eliminate the irrelevant while taking care of the redundant features. Our proposed FAST algorithm falls into the second group.

Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function. However, Relief is ineffective at removing redundant features as two predictive but highly correlated features are likely both to be highly weighted. Relief-F extends Relief, enabling this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot identify redundant features.

However, along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms, and thus should be eliminated. FCBF [2] and CMIM [3] are examples that take into consideration the redundant features. 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. FCBF ([2], [4]) is a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis. CMIM [3] iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked. Different from these algorithms, our proposed the FAST algorithm employs the clustering-based method to choose features.

Recently, hierarchical clustering has been adopted in word selection in the context of text classification. Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words by Pereira et al. or on the distribution of class labels associated with each word by Baker and McCallum. As distributional clustering of words are agglomerative in nature, and result in suboptimal word clusters and high computational cost, Dhillon et al. [5] proposed a new information-theoretic divisive algorithm for word clustering and applied it to text classification. Butterworth et al. [8] proposed to cluster features using a special metric of Barthelemy-Montjardet distance, and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluation measure based on Barthelemy-Montjardet distance does not

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

identify a feature subset that allows the classifiers to improve their original performance accuracy. Furthermore, even compared with other feature selection methods, the obtained accuracy is lower.

Hierarchical clustering also has been used to select features on spectral data. Van Dijck and Van Hulle [7] proposed a hybrid filter/wrapper feature subset selection algorithm for regression. Krier et al. [6] presented a methodology combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information. Their feature clustering method is similar to that of Van Dijck and Van Hulle [7] except that the former forces every cluster to contain consecutive features only. Both methods employed agglomerative hierarchical clustering to remove redundant features. Quite different from these hierarchical clustering-based algorithms, our proposed FAST algorithm uses minimum spanning tree-based method to cluster features. Meanwhile, it does not assume that data points are grouped around centers or separated by a regular geometric curve. Moreover, our proposed FAST does not limit to some specific types of data.

III. MINIMUM SPANNING TREE

There may be several minimum spanning trees of the same weight; in particular, if all weights are the same, every spanning tree is minimum. If each edge has a distinct weight then there will only be one, unique minimum spanning tree. The proof to this fact is trivial and can be done by induction. This is true in many realistic situations, such as the cable TV company example above, where it's unlikely any two paths have exactly the same cost. This generalizes to spanning forests as well. If the weights are non-negative, then a minimum spanning tree is in fact the minimum-cost subgraph connecting all vertices, since subgraphs containing cycles necessarily have more total weight.

For any cycle C in the graph, if the weight of an edge e of C is larger than the weights of other edges of C, then this edge cannot belong to a MST. Indeed, assume the contrary, i.e., e belongs to a MST T1. If its deleted it, T1 will be broken into two subtrees with the two ends of *e* in different subtrees. The remainder of *C* reconnects the subtrees, hence there is an edge *f* of *C* with ends in different subtrees, i.e., it reconnects the subtrees into a tree T2 with weight less than that of T1, because the weight of *f* is less than the weight of *e*, q.e.d. For any cut C in the graph, if the weight of an edge e of C is smaller than the weights of other edges of C, then this edge belong to all MSTs of the graph. Indeed, assume the contrary, i.e., *e* does not belong to a MST T1. Then adding *e* to T1 will produce a cycle, which must have another edge *e2* from T1 in the cut *C*. Replacing *e2* with *e* would produce a tree T1 of smaller weight.

The first algorithm for finding a minimum spanning tree was developed by Czech scientist Otakar Borůvka in 1926. Its purpose was an efficient electrical coverage of Moravia. There are now two algorithms commonly used, Prim's algorithm and Kruskal's algorithm. All three are greedy algorithms that run in polynomial time, so the problem of finding such trees is in P. Another greedy algorithm not as commonly used is the reverse-delete algorithm, which is the reverse of Kruskal's algorithm. The fastest minimum spanning tree algorithm to date was developed by Bernard Chazelle, and based on Borůvka's. Its running time is O(e α(e,v)), where *e* is the number of edges, *v* refers to the number of vertices and α is the classical functional inverse of the Ackermann function. The function α grows extremely slowly, so that for all practical purposes it may be considered a constant no greater than 4; thus Chazelle's algorithm takes very close to O(e) time.

What is the fastest possible algorithm for this problem That is one of the oldest open questions in computer science. There is clearly a linear lower bound, since the system must at least examine all the weights. If the edge weights are integers with a bounded bit length, then deterministic algorithms are known with linear running time, O(e). For general weights, David Karger exhibited a randomized algorithm whose expected runtime is linear. Whether there exists a deterministic algorithm with linear running time for general weights is still an open question. However, Seth Pettie and Vijaya Ramachandran have found a provably optimal deterministic minimum spanning tree algorithm, the computational complexity of which is unknown. More recently, research has focused on solving the minimum spanning tree problem in a highly parallelized manner. For example, the pragmatic 2003 Fast Shared-Memory Algorithms for Computing the Minimum Spanning Forest of Sparse Graphs" by David A. Bader and Guojing Cong demonstrates an algorithm that can compute MSTs 5 times faster on 8 processors than an optimized sequential algorithm. Typically, parallel algorithms are based on Boruvka's algorithm Prim's and especially Kruskal's algorithm do not scale as well to additional processors.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

IV. FEATURE SUBSET SELECTION FOR HIGH-DIMENSIONAL DATA

With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories [1]. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high.

However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. The wrapper methods are computationally expensive and tend to overfit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, we will focus on the filter method in this paper. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms. Pereira et al., Baker and McCallum and Dhillonet al. employed the distributional clustering of words to reduce the dimensionality of text data.

In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Their results have, sometimes, the best agreement with human performance. The general graph-theoretic clustering is simple: compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter than its neighbors. The result is a forest and each tree in the forest represents a cluster. In our study, we apply graph-theoretic clustering methods to features. In particular, we adopt the minimum spanning tree (MST)-based clustering algorithms, because they do not assume that data points are grouped around centers or separated by a regular geometric curve and have been widely used in practice.

Based on the MST method, we propose a Fast clustering bAsed feature Selection algoriThm (FAST). The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form the final subset of features. Features in different clusters are relatively independent the clustering based strategy of FAST has a high probability of producing a subset of useful and independent features. The proposed feature subset selection algorithm FAST was tested upon 35 publicly available image, microarray, and text data sets. The experimental results show that, compared with other five different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of the four well-known different types of classifiers.

V. PROBLEM STATEMENT

Fast clustering-based feature selection algorithm (FAST) is used to cluster the high dimensional data and feature selection process. FAST algorithm is divided into two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature is selected from each cluster to form a subset of features. Features in different clusters are relatively independent. The clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. Minimum-Spanning Tree (MST) clustering

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

method is adopted to ensure the efficiency of FAST. Feature subset selection algorithm is used to identify the features from the clusters. The following drawbacks are identified from the existing system. They are correlation measures are not optimized, limited clustering accuracy, feature relevance is low and threshold is not optimized.

VI. FEATURE SELECTION WITH CORRELATION MEASURES

The feature selection process is improved with a set of correlation measures. Dynamic feature intervals can be used to distinguish features. Redundant feature filtering mechanism is used to filter the similar features. Custom threshold is used to improve the cluster accuracy. The graph based clustering algorithm is designed with Minimum Spanning Tree (MST). Correlation measures are optimized to improve the cluster results. Feature selection process is enhanced with dynamic threshold values. The system is designed with five major modules. They are data preprocess, irrelevant filtering, MST construction, cluster process and feature selection. Data preprocess is designed to perform data cleaning with missing value assignment Process. Irrelevant filtering module is designed to filter irrelevant features with correlation analysis. MST construction module is designed to construct Minimum Spanning Tree (MST) with transactions. Cluster process module is designed to partition the MST with boundaries. Feature selection module is designed to fetch features from cluster results.

6.1. Data Preprocess

Noisy data remove and missing data update operations are carried out under the data preprocess. Redundant data values are removed from the transactional data collection. Aggregation based data substitution mechanism is used for missing data update process. Dimensionality analysis is performed for high dimensional data values. The system can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. We achieve this through a new feature selection framework which composed of the two connected components of irrelevant feature removal and redundant feature elimination. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset.

6.2. Irrelevant Filtering

Irrelevant filtering process is carried out to remove irrelevant features. Correlation measures are used in the relevancy analysis process. Relevancy is analyzed for all features. A threshold value is used to filter the feature values. Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, "good feature subsets contain features highly correlated with the class, yet uncorrelated with each other".

6.3. MST Construction

Graph theoretic method is applied for the tree construction. Minimum Spanning Tree (MST) is constructed with the neighborhood information. Shorter/longer edges are removed with reference to its neighbors. The MST produces a forest with a set of trees. The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature elimination is a bit of sophisticated. In our proposed FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the clusters.

6.4. Cluster Process

Features are divided into clusters by using graph-theoretic clustering methods. Fast clustering algorithm is used for the data partitioning process. The Minimum Spanning Tree is used in the clustering process. Trees under the MST are separated with interval values as clusters. Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

6.5. Feature Selection

Feature selection is applied to filter a subset of useful and independent features. Features in different clusters are relatively independent. Similar features are filtered from the cluster results using redundant feature filtering method. The features strongly related to target classes is selected from each cluster to form the final subset of features. In order to more precisely introduce the algorithm, and because our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination, we first present the traditional definitions of relevant and redundant features, then provide our definitions based on variable correlation.

VII. CONCLUSION

The high dimensional data values are grouped using the clustering technique. Feature selection methods are used to select key elements in the transactions. FAST algorithm is used to select features from high dimensional data values. Correlation measures are used to improve the feature selection process. The system achieves high feature selection quality. Process time is low in the feature selection scheme. The selected features can be applied for classification process. Cluster accuracy is high in the correlation measures based feature selection process.

REFERENCES

[1] Yanqing Ji, Hao Ying, John Tran, Peter Dews, Ayman Mansour, and R. Michael Massanari, "A Method for Mining Infrequent Causal Associations and Its Application in Finding Adverse Drug Reaction Signal Pairs", IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 4, April 2013.
[2] C. Marinica and F. Guillet, "Knowledge-Based Interactive Postmining of Association Rules Using Ontologies," IEEE Trans. Knowledge and Data Eng., June 2010.
[3] L. Troiano, G. Scibelli, and C. Birtolo, "A Fast Algorithm for Mining Rare Itemsets," Proc. Ninth Int'l Conf. Intelligent Systems Design and Applications, 2009.
[4] H. Jin, J. Chen, C. Kelman, and C. O'Keefe, "Mining Unexpected Temporal Associations: Applications in Detecting Adverse Drug Reactions," IEEE Trans. Information Technology in Biomedicine, July 2008.
[5] G.N. Nore´n, J. Hopstadius, A. Bate, K. Star, and I.R. Edwards, "Temporal Pattern Discovery in Longitudinal Electronic Patient Records" Data Mining and Knowledge Discovery, vol. 20, pp. 361- 387, 2010.
[6] Y. Ji, H. Ying, P. Dews, M.S. Farber, A. Mansour, J. Tran, R.E. Miller, and R.M. Massanari, "A Fuzzy Recognition-Primed Decision Model-Based Causal Association Mining Algorithm for Detecting Adverse Drug Reactions in Postmarketing Surveillance," Proc. IEEE Int'l Conf. Fuzzy Systems, 2010.
[7] Y. Ji, H. Ying, P. Dews, A. Mansour, J. Tran, R.E. Miller, and R.M. assanari, "A Potential Causal Association Mining Algorithm for Screening Adverse Drug Reactions in Postmarketing Surveillance," IEEE Trans. Information Technology in Biomedicine, vol. 15, no. 32, pp. 428-437, May 2011.
[8] J. Pearl, Causality: Models, Reasoning and Inference, second ed. Cambridge Univ. Press, 2009.
[9] L. Szathmary and A. Napoli, "Finding Minimal Rare Itemsets and Rare Association Rules," Proc. Fourth Int'l Conf. Knowledge Science, Eng. and Management, 2010.