International Journal of Technology and Engineering System (IJTES)
Vol 7. No.1 2015 Pp. 76-80

available at : www.ijcns.com
ISSN: 0976-1345

ENHANCED MINING OF HIGH DIMENSIONAL DATA USING CLUSTERING BASED FEATURE SUBSET SELECTION ALGORITHM

K.BABU[1], P.CHARLES[2]
Department of Computer Science MRK Institute of Technology
Kattumannarkoil-608 301
[1]babukumarit@gmail.com
[2]Unicorn.charl@yahoo.com

ABSTRACT

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). The many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features 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.
Key words: Feature subset selection, filter method, feature clustering, graph-based clustering.

I. INTRODUCTION

The aim of choosing a feature subset selection is an effective way for reducing dimensionality, removing unwanted data, increasing learning accuracy. The feature subset selection methods have been studied for machine learning applications. They can be divided into four categories such as

- ✓ Embedded
- ✓ Wrapper
- ✓ Filter
- ✓ Hybrid.

(I) Embedded Method:

The embedded method is more efficient than other three categories. Decision trees (or) artificial neural network are examples of embedded approaches.

Decision tree is a tree shaped diagram. It is mainly used to determine a course of action. Each branch of the decision tree represents a possible decision. The tree structure is used to shows how one choice leads to the next and the use of branches indicates that each option is mutually exclusive.

Artificial neural networks Non-linear predictive models that learn through training and resemble biological neural networks in structure.

(Ii) Wrapper Method:

The wrapper method is mainly used to determine the goodness of the selected subsets, the accuracy of the learning algorithm is high and computational complexity also high. The main objective function is a pattern classifier, which evaluates feature subsets by their predictive accuracy by cross validation.

Advantages

- ✓ Accuracy
- ✓ Ability to generalize.

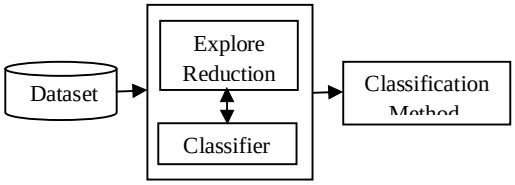


Figure 1.1 Wrapper Architecture

(Iii) Filter Method:

The filter methods are independent of learning algorithm and computational complexity is low, but the accuracy of learning algorithm is not guaranteed. The filter method is mainly used to evaluation is independent of the classification algorithm.

Advantages

- ✓ Fast Execution
- ✓ Generality.

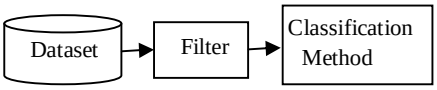


Figure 1.2: Filter Architecture.

(IV) Hybrid Method:

The hybrid methods are combination of filter and wrapper methods. The filter methods to reduce search space that will be considered by the subsequent wrapper. It is mainly used to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods.

II.RELATED WORK

R. Battiti, "Using Mutual Information for Selecting features in Supervised Neural Net Learning," [1] has investigate the application of the mutual. In for criterion to evaluate a set of candidate features and to select an informative subset to be used as input data for a neural network classifier. Nonetheless, the use of the mutual information for tasks characterized by high input Dimensionality requires suitable approximations because of the Prohibitive demands on computation and samples. An algorithm is proposed that is based on a "greedy" selection of the features and that takes both the mutual information with respect to the Output class and with respect to the already-selected features into account. Finally the results of a series of experiments are discussed.

D.A. Bell and H. Wang, "A Formalism for Relevance and Its Application in Feature Subset Selection," [2] The notion of relevance is used in many technical fields. In the areas of machine learning and data mining, for example, relevance is frequently used as a measure in feature subset selection (FSS). In previous studies, the interpretation of relevance has varied and its connection to FSS has been loose. In this paper a rigorous mathematical formalism is proposed for relevance, which is quantitative and normalized. To apply the formalism in FSS, a characterization is proposed for FSS: preservation of learning information and minimization of joint entropy. Based on the characterization, a tight connection between relevance and FSS is established: maximizing the relevance of features to the decision attribute and the relevance of the decision attribute to the features. This connection is then used to design an algorithm for FSS. The algorithm is linear in the number of instances and quadratic in the number of features. The algorithm is evaluated using 23 public datasets, resulting in an improvement in prediction accuracy on 16 datasets, and a loss in accuracy on only 1 dataset. This provides evidence that both the formalism and its connection to FSS are sound

C. Cardie, "Using Decision Trees to Improve Case-Based Learning," [3] The decision trees can be used to improve the performance of case based learning (CBL) systems. We introduce a performance task for machine learning systems called semi-flexible prediction that lies between the classification task performed by decision tree algorithms and the flexible prediction task performed by conceptual clustering systems. In semi-flexible prediction, learning should improve prediction of a specific set of features known a priori rather than a single known feature (as in classification) or an arbitrary set of features (as in conceptual clustering). We describe one such task from natural language processing and present experiments that compare solutions to the problem using decision trees, CBL, and a hybrid approach that combines the two. In the hybrid approach, decision trees are used to specify the features to be included in k-nearest neighbor case retrieval

P. Chanda, Y. Cho, A. Zhang, and M. Ramanathan, "Mining of Attribute Interactions Using Information Theoretic Metrics," [4] Knowledge of the statistical interactions between the attributes in a data set provides insight into the underlying structure of the data and explains the relationships (independence, synergy, redundancy) between the attributes. In a supervised learning problem, normally, small subsets of the classifying attributes are actually associated with the class label. Interaction information among the attributes captures the multivariate dependencies (synergy and

redundancy) among the attributes and the class label. Mining the significant statistical interactions that contain information about the class label is a computationally challenging task - the number of possible interactions increases exponentially and most of these interactions contain redundant information when a number of correlated attributes are present. In this paper, we present a data mining method (named IM or Interaction Mining) to mine non-redundant attribute sets that have significant interactions with the class label. We further demonstrate that the mined statistical interactions are useful for improved feature selection as they successfully capture the multivariate inter-dependencies among the attributes.

III.EXISTING SYSTEM

The embedded method is more efficient than other three categories. Decision trees (or) artificial neural network are examples of embedded approaches. Decision tree is a tree shaped diagram. It is mainly used to determine a course of action. Each branch of the decision tree represents a possible decision. The tree structure is used to shows how one choice leads to the next and the use of branches indicates that each option is mutually exclusive. Artificial neural networks Non-linear predictive models that learn through training and resemble biological neural networks in structure. The wrapper method is mainly used to determine the goodness of the selected subsets, the accuracy of the learning algorithm is high and computational complexity also high. The filter methods are independent of learning algorithm and computational complexity is low, but the accuracy of learning algorithm is not guaranteed. The filter methods to reduce search space that will be considered by the subsequent wrapper.

A. Disadvantages of Existing System

- ✓ The generality of the selected features is limited and computational complexity is large.
- ✓ The hybrid methods are a combination of filter an wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper.

IV.NATURE OF WORK

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.

A. Advantages of Proposed System

- ✓ Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.
- ✓ The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.
- ✓ Generally all the six algorithms achieve significant reduction of dimensionality by selecting only a small portion of the original features.
- ✓ The null hypothesis of the Friedman test is that all the feature selection algorithms are equivalent in terms of runtime.

V. METHODOLOGY

- ✓ Distributed clustering
- ✓ Subset selection algorithm
- ✓ Time complexity
- ✓ Microarray data
- ✓ Data resource
- ✓ Irrelevant feature.

A. Distributed Clustering

The 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. Proposed to cluster features using a special metric of distance, and then makes use of the of the resulting cluster hierarchy to choose the most relevant attributes. Unfortunately, the cluster evaluation measure based on distance does not 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.

B. Subset selection algorithm

The 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 (predictive of) the class, yet uncorrelated with (not predictive of) each other. Keeping these in mind, we develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

C. Time complexity

The major amount of work for Algorithm 1 involves the computation of SU values for TR relevance and F-Correlation, which has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity in terms of the number of features m. Assuming features are selected as relevant ones in the first part, when k ¼ only one feature is selected.

D. Microarray data

The proportion of selected features has been improved by each of the six algorithms compared with that on the given data sets. This indicates that the six algorithms work well with microarray data. FAST ranks 1 again with the proportion of selected features of 0.71 percent. Of the six algorithms, only CFS cannot choose features for two data sets whose dimensionalities are 19,994 and 49,152, respectively.

E. Data resource

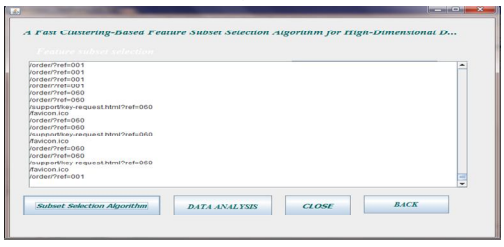
The purposes of evaluating the performance and effectiveness of our proposed FAST algorithm, verifying whether or not the method is potentially useful in practice, and allowing other researchers to confirm our results, 35 publicly available data sets1 were used. The numbers of features of the 35 data sets vary from 37 to 49, 52 with a mean of 7,874. The dimensionalities of the 54.3 percent data sets exceed 5,000, of which 28.6 percent data sets have more than 10,000 features. The 35 data sets cover a range of application domains such as text, image and bio microarray data classification in the corresponding statistical information that for the data sets with continuous-valued features, the well-known off-the-shelf MDL method was used to discredit the continuous values.

F. Irrelevant feature

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 3. The selection of representative features from the clusters.

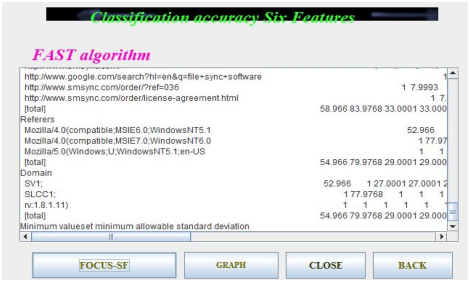
VI. RESULTS

(i)Subset selection Algorithm



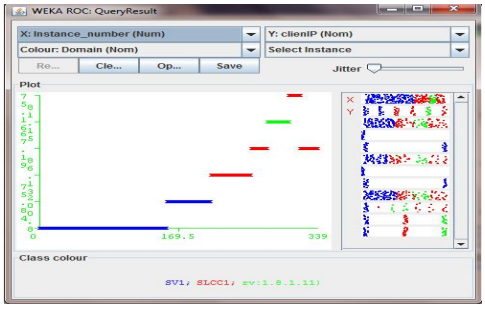
(ii)FAST Algorithm

This page will be displayed FOCUS-SF algorithm.



(iii) Query Result

This page will be displayed the query results



VII.CONCLUSION

The novel clustering-based feature subset selection algorithm 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. Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy confirmed the conclusions. We have presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves removing irrelevant features, constructing a minimum spanning tree from relative ones, and 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. We have compared the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, CFS, Consist, and FOCUS-SF on the publicly available image, microarray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record.

REFERENCES

[1] *H. Almuallim and T.G. Dietterich, "Algorithms for Identifying Relevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.*

[2] H. Almuallim and T.G. Dietterich, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305, 1994.

[3] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[4] L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.

[5] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.

[6] J. Biesiada and W. Duch, "Features Election for High-Dimensional data a Pearson Redundancy Based Filter," Advances in Soft Computing, vol. 45, pp. 242-249, 2008.

[7] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 581-584, 2005.

[8] C. Cardie, "Using Decision Trees to Improve Case-Based Learning," Proc. 10th Int'l Conf. Machine Learning, pp. 25-32, 1993.

[9] P. Chanda, Y. Cho, A. Zhang, and M. Ramanathan, "Mining of Attribute Interactions Using Information Theoretic Metrics," Proc. IEEE Int'l Conf. Data Mining Workshops, pp. 350-355, 2009.

Babu *is an Assistant Professor in the Dept. of Computer Science and Engineering at M.R.K.Institute of Technology, Kattumannarkoil, India from 2013. He received his B.Tech degree at E.S.College of Engineering and Technology from Anna University in 2011, Chennai and the M.Tech degree at Anna University Regional Centre, Coimbatore from Anna University, and Chennai in 2013.*

Charles *is an Assistant Professor in the Dept. of Computer Science and Engineering at M.R.K.Institute of Technology, Kattumannarkoil, India from 2013. He received his B.Tech degree at Annai Teresa College of Engineering and Technology from Anna University in 2009, Chennai and the M.Tech degree at Prist University Pondicherry in 2012.*