FEATURE SUBSET SELECTION ALGORITHM OVER MULTIPLE DATASET

PRIYANKA M G

PG Scholar, Department of Computer Science and Engineering, TKM Institute of Technology, Karuvelil, Kollam (Dist), Kerala, India

Abstract - Feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Here a fast clustering based feature subset selection algorithm is used. The algorithm involves (i) removing irrelevant features, (ii) constructing clusters from the relevant features, and (iii) removing redundant features and selecting representative features. It is an effective way for reducing dimensionality. This FAST algorithm has advantages like efficiency and effectiveness. Efficiency concerns the time required to find a subset of features and effectiveness is related to the quality of the subset of features. It can be extended to use with multiple datasets.

Index Terms: Feature Subset Selection, Feature Clustering, Graph-Based Clustering

I. INTRODUCTION

In machine learning and statistics, feature selection is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points).

The feature subset selection algorithm A Fast clustering-bAsed feature Selection algorithm (FAST) works in two steps. In the first step, features are divided into cluster 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.

Of 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. The FAST algorithm falls into the second group. Naive Bayes classification algorithms are employed to classify data sets before and after feature selection. Naive Bayes utilizes a probabilistic method for classification by multiplying the individual probabilities of every feature-value pair. This algorithm assumes independence among the features and even then provides excellent classification results.

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because: (i) irrelevant features do not contribute to the predictive accuracy, and (ii) 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. So there is a need for such an algorithm that eliminates both irrelevant and redundant features.

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 this in mind, develop a novel algorithm which can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. This can be achieving 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.

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 (i) the construction of the minimum spanning tree (MST) from a weighted complete graph; (ii) the partitioning of the MST into a forest with each tree representing a cluster; and (iii) the selection of representative features from the clusters. The rest of the paper is organized as follows: a brief review of the existing techniques for feature selection in section II, proposed system in section III , result analysis in section IV followed by conclusion.

II. REVIEW OF EXISTING TECHNIQUES

Feature selection is frequently used as a preprocessing step to machine learning. It is a process of choosing a subset of original features so that the feature space is optimally reduced according to a certain evaluation criterion. In recent years, data has become increasingly larger in both number of instances and number of features in many applications such as genome projects, text categorization, image retrieval, and customer relationship management. This enormity may cause serious problems to many machine learning algorithms with respect to scalability and learning performance. For example, high dimensional data (i.e., data sets with hundreds or thousands of features) can contain high degree of irrelevant and redundant information which may greatly degrade the performance of learning algorithms. Therefore, feature selection becomes very necessary for machine learning tasks when facing high dimensional data nowadays. However, this trend of enormity on both size and dimensionality also poses severe challenges to feature selection algorithms.

A. Feature weighting algorithms

Feature weighting algorithms assign weights to features individually and rank them based on their relevance to the target concept. A feature is good and thus will be selected if its weight of relevance is greater than a threshold value. A well-known algorithm that relies on relevance evaluation is Relief. The key idea of Relief is to estimate the relevance of features according to how well their values distinguish between the instances of the same and different classes that are near each other.

Relief randomly samples a number of instances from the training set and up-dates the relevance estimation of each feature based on the difference between the selected instance and the two nearest instances of the same and opposite classes. However, Relief does not help with removing redundant features. As long as features are deemed relevant to the class concept, they will all be selected even though many of them are highly correlated to each other. Many other algorithms in this group have similar problems as Relief does. They can only capture the relevance of features to the target concept, but cannot discover redundancy among features.

However, empirical evidence from feature selection literature shows that, along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms and thus should be eliminated as well. Therefore, in the context of feature selection for high dimensional data where there may exist many redundant features, pure relevance-based feature weighting algorithms do not meet the need of feature selection very well.

B. Subset search algorithms

Subset search algorithms search through candidate feature subsets guided by a certain evaluation measure which captures the goodness of each subset. An optimal (or near optimal) subset is selected when the search stops. Some existing evaluation measures that have been shown effective in removing both irrelevant and redundant features include the consistency measure, and the correlation measure. Consistency measure attempts to find a minimum number of features that separate classes as consistently as the full set of features can.

An inconsistency is defined as two instances having the same feature values but different class labels. Different search strategies, namely, exhaustive, heuristic, and random search, are combined with this evaluation measure to form different algorithms. The time complexity is exponential in terms of data dimensionality for exhaustive search and quadratic for heuristic search. The complexity can be linear to the number of iterations in a random search, but experiments show that in order to find best feature subset, the number of iterations required is mostly at least quadratic to the number of features.

A correlation measure is applied to evaluate the goodness of feature subsets based on the hypothesis that a good feature subset is one that contains features highly correlated to the class, yet uncorrelated to each other. The underlying algorithm, named CFS, also exploits heuristic search. Therefore, with quadratic or higher time complexity in terms of dimensionality , existing subset search algorithms do not have strong scalability to deal with high dimensional data.

To overcome the problems of algorithms in both groups and meet the demand for feature selection for high dimensional data, A Fast Correlation-Based Filter Solution algorithm is used which can effectively identify both irrelevant and redundant

features with less time complexity than subset search algorithms. Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features.

A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, a fast clustering -based feature selection algorithm is generated.

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 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. To ensure the efficiency of FAST, clustering method is used.

The feature selection framework is composed of two connected components such as 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 from different feature clusters, and thus produces the final subset .

Disadvantages of Existing System are this is less useful for image data compared to other algorithms and Use only single dataset.

III. PROPOSED SYSTEM

The proposed system uses multiple dataset for feature selection instead of single dataset. It is an effective way for reducing dimensionality and efficiently and effectively deal with both irrelevant and redundant features. It is highly effective for machine learning applications and accuracy is usually high.

A. Architecture

This section explains the architecture diagram of the system to be developed. The feature selection framework (shown in Fig.1) is 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.

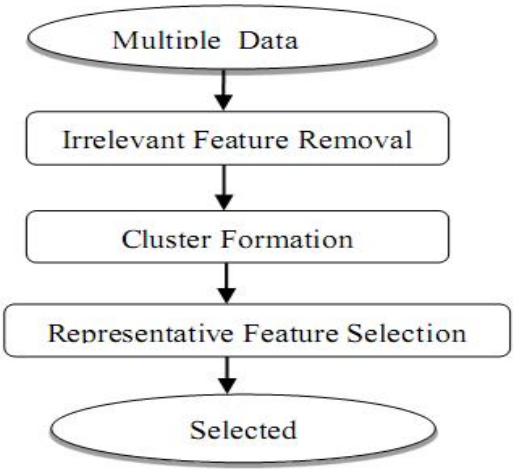

Fig 1: Architecture of the system

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 the FAST algorithm, it involves (i) the construction of clusters from the relevant features; (ii) the selection of representative features.

Mutual information measures how much the distribution of the feature values and target classes differ from statistical independence. This is a nonlinear estimation of correlation between feature values or feature values and target classes. The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes. Symmetric Uncertainty is the measure of correlation between either two features or a feature and the target concept.

The proposed FAST algorithm logically consists of three steps: (i) removing irrelevant features, (ii) constructing a MST from relative ones, and (iii) partitioning the MST and selecting representative features. For each data set D with m features $F = \{F1, F2, ..., Fm\}$ and class C, we compute the T-elevance $SU(Fi, C)$ value for each feature Fi $(1 \leq i \leq m)$ in the first step. The features whose (Fi, C) values are greater than a predefined threshold θ comprise the target-relevant feature subset $F' = \{F'1, F'2, ..., F'k\}$ $(k \leq m)$.

In the second step, we first calculate the F-Correlation$(F'i, F'j)$ value for each pair of features $F'i$ and $F'j$ $(F'i, F'j \in F' \wedge i \neq j)$. Then, viewing features $F'i$ and $F'j$ as vertices and $SU(F'i, F'j)$ $(i \neq j)$ as the weight of the edge between vertices $F'i$ and $F'j$, a weighted complete graph $G = (V,E)$ is constructed where $V = \{F'i \in F' i \in F' \wedge i \in [1, k]\}$ and $E = \{(F'i, F'j) \in (F'i, F'j \in F' \wedge i, j \in [1, k] \wedge i \neq j\}$. As symmetric uncertainty is symmetric further the F - Correlation $(F'i, F'j)$ is symmetric as well, thus G is an undirected graph. The complete graph G reflects the correlations among all the target-relevant

features. Unfortunately, graph G has k vertices and $(k-1)/2$ edges. For high dimensional data, it is heavily dense and the edges with different weights are strongly interweaved Thus for graph G, we build a MST, which connects all vertices such that the sum of the weights of the edges is the minimum, using the well-known Prim algorithm. The weight of edge $(F'i, F'j)$ is F-Correlation $(F'i, F'j)$.

After building the MST, in the third step, we first remove the edges $E = \{(F'i, F'j) \in (F'i, F'j \in F' \wedge i, j \in [1, k] \wedge i \neq j\}$, whose weights are smaller than both of the T-Relevance $SU(F'i, C)$ and $SU(F'j, C)$, from the MST. Each deletion results in two disconnected trees $T1$ and $T2$.

Assuming the set of vertices in any one of the final trees to be $V(T)$, we have the property that for each pair of vertices $(F'i, F'j \in V(T))$, $(F'i, F'j) \geq S(F'i, C) \vee SU(F'i, F'j) \geq SU(F'j, C)$ always holds. The features in $V(T)$ are redundant.

IV. RESULT ANALYSIS AND DISCUSSION

The fast algorithm is implemented on combined multiple datasets. It analyses the irrelevant features as well as the redundant features and remove them. First it removes irrelevant features from the dataset. The features relevance values are greater than a predefined threshold comprise the target-relevant feature subset. Redundant features are selected from the relevant features. This algorithm achieves significant reduction of dimensionality by selecting only a small portion of the original features . FAST is an individual evaluation based feature selection algorithm . This is much faster than the subset evaluation based algorithms.The run time of FAST is less when using multiple combined datasets than using separate individual datasets. Like many other feature selection algorithms, our proposed FAST also requires a parameter θ that is the threshold of feature relevance. Different θ values might end with different classification results. When determining the value of θ, the proportion of the selected features should be taken into account. This is because improper proportion of the selected features results in a large number of features, and further affects the classification efficiency. The default θ values used for FAST in the experiments are often not the optimal.

CONCLUSIONS

In the clustering-based feature subset selection algorithm for high dimensional data involves (i) removing irrelevant features, (ii) cluster formation, and (iii) removing redundant features and selecting representative features. In this, a cluster consists of features. It is an effective way for reducing dimensionality. This FAST algorithm has advantages like efficiency and effectively deal with both irrelevant and redundant features, highly effective for machine learning applications, and accuracy is usually high .In the proposed system multiple dataset can be used as combined one and hence dimensionality can be further reduced. Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime.

ACKNOWLEDGMENT

This work is supported in part by the Department of Computer Science & Engineering, TKMIT, Kollam. I would like to show my gratitude to Prof P. Mohamed Shameem & Asst. Prof. Revathy N for their valuable guidance.

REFERENCES

[1] Arauzo-Azofra A., Benitez J.M. and Castro J.L., " A feature set measure based on relief", In Proceedings of the fifth international conference on Recent Advances in Soft Computing, 2004,pp 104-109.

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

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

[4] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data" , IEEE Transactions On Knowledge And Data Engineering Vol:25 No:1,2013

[5] Bhaskar Adepu,kiran Kumar, "A novel approach for minimum spanning tree based clustering algorithm",2008,1-8.

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

[7] Zheng Chen,Heng Ji , " Graph-based Clustering for Computational Lingustics : A Survey", Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL , 2010,1–9.

[8] Yu L. and Liu H., "Efficiently handling feature redundancy in high dimensional data", in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03). ACM, New York, NY, USA, 2003,pp 685-690..

★★★