Eminent Feature Selection for High Dimensional Data through the Propositional Fast-Foil Rules

S.Deepika,
Assistant Professor, B.Com (Business Analytics),
PSGR Krishnammal College for Women, Coimbatore, India
Email - deepika@psgrkc.ac.in,

Abstract: *Feature subset selection can be viewed as the process of finding and eradicating many irrelevant and redundant features. Feature subset selection achieves its intended purpose as possible with three major steps. First becomes the construction of minimum spanning tree, after that the partition the data into each tree by clustering the similar features. Finally selected features are represented into clusters. Feature interaction is an important issue in feature subset selection.*

Key Words: *Feature Subset Selection, Filter Method, Clustering, Feature Interaction, Fast-Foil.*

1. INTRODUCTION:

Data mining is a said to be discovery of knowledge from huge data sets. Process of mining apart from data analysis is classification, cluster, association rule mining and forecast. The work attempts to predict efficiently analysis with reduced number of features using classification data mining technique. The objective of clustering is to see the core grouping during a set of unlabelled information.

The filter approach figures the feature evaluation weight but without performing classification of data, eventually finding the ‘good’ subset of features. The principle of filter approaches is to select the subset of features which have high dependence on target class and while have less connection among them and to measure the importance by maximizing the clustering performance.

- Selecting the subsets of variables is termed to be pre-processing step, independently of the used classifier.
- Filter method is said to be Variable Ranking-FS.
- Usually fast.
- Offer a standard selection of features, not tuned by given learner (universal).

Classifier are employed to classify data sets before and after feature selection, Instance-Based - IB1. IB classifier, refers a simple distance measure to determine the training instance contiguous to the given test instance, and predict the same class as this training instance. If multiple instances are the same (minimum) distance to the experiment instance, the initial one found is used.

Decision tree learning are used in statistics, processing and machine learning, employ a decision tree as a predictive model that maps observations concerning an item to conclusions concerning the item's target price. In this tree structure, leaves stands for category labels and branches stand for conjunctions of choices that result in those category labels.

FAST_FOIL rule based algorithm eradicates irrelevant and redundant ones in addition to that feature interaction are measured, is introduced for selecting features subset for high dimensional data. FOIL RULE based subset selection algorithm selects the features based on the rules generated from the FRFAST first merges the features appeared in the antecedents of all FOIL rules, which will achieve a candidate feature subset to avoid redundant features and reserves interactive ones. Finally the experiments are carried out to compare FAST-FOIL, FAST and representative feature selection algorithms, namely, FCBF and Consist with respect to the classifier, namely, Instance-based IB1 before and after feature selection. Accuracy and time expenditure for FAST_FOIL is preferably fast while compared to other algorithm.

Therefore, there is improvement in algorithm. Thus proposed system produces a significant difference in the result. The process was developed in R Studio.

2. MATERIALS:

A filter method on Fast will determine relevant features and redundant features without pair wise correlation analysis. The efficiency and effectiveness of the method has extensive comparisons with other methods using real-world data of high dimensionality. FCBF is applied and evaluated through extensive experiments comparing with related feature selection algorithms.

A feature interaction is to attain efficient feature selection and present wide experimental results of evaluation. To measure the feature relevancy c-contribution is used for feature interaction. The key issues deterring the use of consistency measures and developed INTERACT that handles feature interaction and efficiently selects relevant features and result in a significant speed-up of the algorithm.

Feature selection involves two steps like eliminating redundant and irrelevant data. By constructing minimum spanning tree from a weighted complete graph; partitioning of the MST (minimum spanning tree) into a forest with each tree representing a cluster; selection of representative features from the clusters finally feature interaction is included by using FOIL rule based algorithm. Classifiers are used as a set of predictor variables (or features) that are most relevant to the classification task.

3. METHOD:

FAST-FOIL ALGORITHM:

Step 1: Given the dataset D from 1 to m features and class label C. D = (F1; F2; ...; FM, C) Where F ={F1;F2; ... ;Fm} and Fi (1 ≤ i ≤ m).
Step 2: Determine T-relevance SU (Fi; C) value for each feature Fi (1 ≤ i ≤ m).
Step 3: CRSet is excavated from a given dataset by the restricted-FOIL algorithm.
Step 4: Candidate feature subset is achieved by joining the features whose values appeared in the past history of the classification rules in CRSet.
Step 5: For each feature in a feature set is then combined to produce candidate feature set.
Step 6: If T-relevance is found to be greater than the predefined threshold value θ, encompass Target relevant feature subset, F'={F1',F2',…FK'}(K ≤ M).

FOR MINIMUM SPANNING TRESS CONSTRUCTION

Step 7: Graph G is a complete graph and found it has null value then.
Step 8: Calculate the F-Correlation SU (Fi',Fj') value for each pair of features Fi' and Fj'.
Step 9: Add Fi' and / or, Fj' to G as vertices with F-correlation SU (Fi',Fj') as the weight of the corresponding edge.
Step 10: Thus a weighted complete graph G (V, E) is constructed.
Step 11: MST is generated using prism algorithm for graph G. All vertices are connected so that Summation of the weight of each edges is seems to be minimum, using prism algorithm.

> Step 11.1 Select any node to be the first node of T.
> Step 11.2 Consider the arcs which connect nodes in T to nodes outside T. Select minimum weight node. Adjoin the arc and the extra node with T. (If there are two or more arcs of minimum weight, fix on with any one of them.)
> Step 11.3 Repeat Step 2 until T contains every node of the graph.

TREE PARTITIONING AND REPRESENTATIVE FEATURE SELECTION

Step 12: Remove the edges whose weights are smaller than both of the T-Relevance SU (Fi',C) and SU (Fj',C) from the MST.
Step 13: A Forest is obtained, for each tree in the forest represents a cluster that is denoted as V(Tj).
Redundancy Are Eliminated
Step 14: For each cluster V(Tj) choose a representative feature FRj whose T-relevance SU(FR j , C) is the greatest.
Step 15: All FR j (j=1, … [FOREST]) comprise the final feature subset U FR j.

4. DISCUSSION:

FOIL Rule based Fast clustering bAsed feature Selection algoriThm (FAST_FOIL). It initially generates the FOIL classification rules using a modified propositional implementation of the FOIL algorithm. Then, it merge the features that appeared in the past history of all rules are collected, and accomplished a candidate feature subset that excludes redundant features and reserves the interactive ones. Lastly, it measures the relevance of the features in the candidate feature subset by our proposed with threshold to select features in the minimum spanning tree and identifies and removes the irrelevant features. It also can significantly improve the performance of the two classifiers such as IB1.

5. ANALYSIS:

Fast clustering bAsed feature Selection algoriThm (FAST) is built based on the MST method. The FAST algorithm works in two steps. Firstly, features are segregated into clusters by using graph-theoretic clustering technique. Secondly, the most representative feature that is strappingly related to target classes is selected from each cluster to form the ultimate subset of features. Features with no similarities in clusters are relatively independent; the clustering based approach of FAST has a high probability of producing a subset of handy and independent features. The proposed discriminative clustering based feature selection algorithm has,

- The MST construction from a weighted complete graph;
- The partitioning of the MST (minimum spanning tree) into a forest with each tree representing a cluster;
- The representative features are selected from the clusters.

6. FINDINGS:

FOIL Rule based Feature subset Selection algorithm (FRFS). FRFS firstly generates the FOIL classification rules using a modified propositional implementation of the FOIL algorithm. Then, it combines the features that appeared in the antecedents of all rules together, and achieves a candidate feature subset that excludes redundant features and reserves the interactive ones. Lastly, it measures the relevance of the features in the candidate feature subset by our proposed new metric Cover Ratio and identifies and removes the irrelevant features. FRFS the not only produces smaller subsets of features but also improves the performances of the two types of classifiers. Its purpose is to focus a learning algorithm on those aspects of the data most useful for analysis and future prediction. It can reduce the dimensionality of the data and may improve a learner either in terms of learning performance, generalization capacity or model simplicity.

7. RESULT:

Efficiency & effectiveness of an algorithm of FAST-FOIL is achieved by selecting a feature subset of the most useful features. Before that features has to extract throughout the dataset. From a heart dataset, using FAST_FOIL algorithm features are extracted by eliminating the redundant & irrelevant data. Thus class label1 produces a positive result whereas class 2 depicts that negative with the help of classifier such as IB1. 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, FOIL RULE based subset selection algorithm selects the features based on the rules generated from the FRFAST is proposed and experimentally evaluated. The proposed method has been estimated using the following measures:

- Accuracy= (TP+TN) / (TP+TN+FP+FN)
- Precision= TP / (TP+FP)
- Recall= TP / (TP+FN), Where TP, TN, FP and FN are the number of positive cases number of true negative.

Precision value is calculated is based on the retrieval of information at true positive prediction, false positive. In healthcare data precision is calculated the percentage of positive results returned that are relevant. The precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, are termed as data that are incorrectly labeled as belonging to the class.

FAST_FOIL achieves high precision, recall and less time consumption in the task of support decision making system and with classifier IB1, where IB1 performs well with better values of precision, recall, accuracy & time consumption.

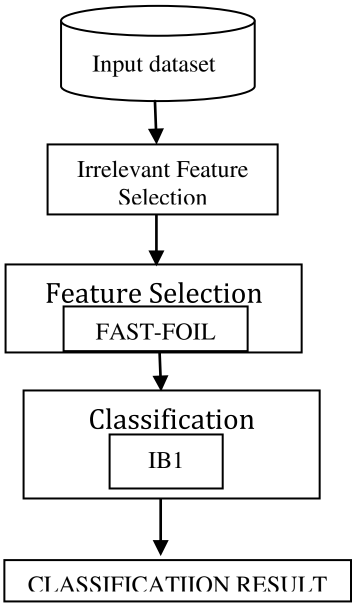

Fig. 1 Flow of Process

TABLE I
FAST AND FOIL KEYWORD DESCRIPTORS

Keyword	Description
T-Relevance	The Relevance between a feature and a target concept C.
F-Correlation	The relationship between any pair of features like F and F .
F-Redundancy	Feature Redundancy is measured from a cluster of features whose value of feature relevance and features correlation should be higher than the other features T-Relevance.
R-Feature	Representative Feature is selected from cluster of features. The feature which has strongest T-Relevance can act as R-Feature for all feature in a cluster.
CRset	Classification Rule set is a predefined rule set.
Candidate Rule Set	The features which are currently selected should satisfy the CRset.
Fset	Feature set which are associated from data set D.
Cover Ratio	Cover Ratio is used to measure relevance of the feature and to find the strength of the Rule set which will detect goodness of the feature.

8. CONCLUSION:

Existing clustering primarily based feature subset selection rule for high dimensional data, the method involves 1) removing extraneous features, 2) building a minimum spanning tree from relative ones, and 3) detaching the MST and selecting representative features, R-Features. In our proposed method, a cluster consists of features. Every cluster is treated as a single feature and thus dimensionality is drastically reduced. Proposed first defined relevant, redundant and interactive features based on classification rules. Then based on these definitions, the feature selection algorithm, which involves two steps (i) redundant feature exclusion and interactive feature reservation and (ii) the irrelevant feature identification. Also explained why these two steps are able to exclude redundant as well as irrelevant features and reserve interactive features with the help of the propositional FOIL rules generated by the restricted FOIL algorithm as FAST_FOIL.

REFERENCES:

1. Almuallim.H and Dietterich .T.G (1992), “Algorithms for Identifying Relevant Features”. Proc. Ninth Canadian Conf. Artificial Intelligence, Pp. 38-45,.
2. Aqueel Ahmed et.al (September 2012), “Data Mining Techniques to Find Out Heart Diseases: An Overview”. International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol.1, pp. 18 – 23,.
3. Bell .D.A and Wang .H (2000), “A Formalism for Relevance and Its Application in Feature Subset Selection”. Machine Learning, vol. 41, no. 2, pp. 175-195,.
4. Butterworth et.al (2005), “On Feature Selection through Clustering”. Proc. IEEE Fifth Int’l Conf. Data Mining, pp. 581-584,.
5. Dash .M and Liu .H and Motoda .H (2000), “Consistency Based Feature Selection”, Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109.
6. Goebel .M and Le Gruenwald (June 1999), “A Survey of Data Mining and Knowledge Discovery Software Tools”, ACM SIGKDD Explorations, vol 1, pp. 20 – 33.
7. Jihoon et.al (May 1997), “Feature Subset Selection Using a Genetic Algorithm”. In Feature extraction, construction and selection, Springer US, pp. 117-136.