Feature Subset Selection Algorithms for Irrelevant Removal Using Minimum Spanning Tree Construction

[1] Asifa Akthar Shaik , [2] M.Purushottam

[1] M.Tech 2nd Year, Department of CSE, SEAT, Tirupati, AP, India

[2] Assistant Professor, Department of CSE, SEAT, Tirupati, AP, India

Abstract - Feature selection in clustering is used for extracting the relevant data from a large collection of data by analyzing on various patterns of similar data. Based on the accuracy and efficiency of the data, major issue occurs in clustering. Feature selection may remedy this issue and thus enhance the prediction accuracy and minimize the particular computational overhead associated with classification algorithms. Irrelevant features usually do not contribute to the actual predictive accuracy, and also redundant features usually do not contribute in order to obtaining a better predictor with the they provide mostly information which can be already contained in other feature(s).

A lot of feature subset selection methods happen to be proposed and also studied with regard to machine learning applications. We propose the Swift clustering-based feature Selection algorithm according to MST as well as PCA. Features in various clusters are usually relatively independent; this particular clustering based technique of SWIFT incorporates a high possibility of producing a subset of useful as well as independent features. SWIFT is in comparison with PCA within the work and also overcomes the particular drawbacks of PCA .

***Keywords* - MST, Swift, Symmetric uncertainty, T-Relevance, F-Correlation, PCA.**

1. Introduction

Clustering has been recognized as an important and valuable capability in the data mining field. For high-dimensional data, traditional clustering techniques may suffer from the problem of discovering meaningful clusters due to the curse of dimensionality. A cluster is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations in the Same cluster are similar in some case. Clustering is an unsupervised learning method, and a common technique for statistical data analysis used in many fields such as Medical, Science, and Engineering. This Survey summaries various known feature selection methods to achieve classification accuracy by using subset of relevant feature. Due to the computational complexity the full original feature set cannot be used. Attribute selection is the process of selecting a subset of relevant features. Feature selection technique is based on the data contains many redundant and irrelevant features. Redundant features provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Subset of Feature selection technique is general field of feature extraction.

1.1 Feature subset selection

Feature subset selection is used to improve the accuracy and comprehensibility that has been explored in machine learning. Some features are dependent on other features and there is no need to include the feature or noisy. Some feature will randomly fit the data and hence the probability of over fitting increases. There are 4 basic steps in any feature selection method. They are Generation, Evaluation, Validation and Stopping criterion. In generation process to select the candidate feature subset, In evaluation process to evaluate the generated candidate feature subset and output a relevancy value, where the stopping criteria will determine whether it is the defined optimal feature subset.

If yes, the process end else the generation process will start again to generate the next candidate feature subset. After the required number of features is obtained the process will be terminated.

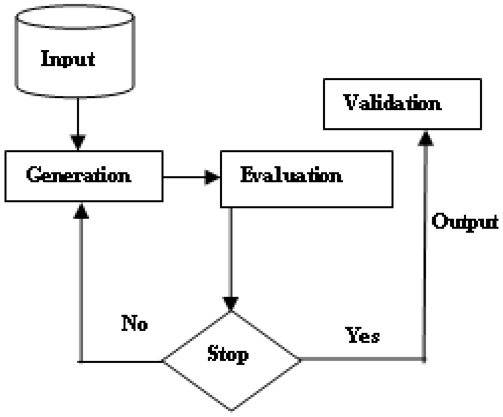

Fig.1 Feature Selection steps

1.2 Feature Selection Benefits

Feature selection method consist of potential benefits are

a. A reduction in the amount of training data needed to achieve learning.

b. The generation of learning models with improved predictive accuracy.

c. Learned knowledge more compact, simpler and easier to understand.

d. Reduced execution time required for learning.
e. Reduced storage requirements.

1.3 Feature subset selection Methods

Several feature subset selection methods are actually proposed as well as studied with regard to machine learning applications. They may be separated into four broad categories: the Embedded, Wrapper, Filter, as well as Hybrid approaches. The particular embedded methods include feature selection as part of the training process and are also usually particular in order to given learning algorithms, and thus could possibly be extremely effective compared to the other three categories. Wrapper methods make use of a predictive model to attain feature subsets. Each and every new subset can be used to learn a model, which can be tested on the hold-out set. Filter methods make use of a proxy measure rather than the error rate to score an element subset. This measure is chosen to become fast in order to compute, whilst even now capturing the particular usefulness from the feature set. Filters tend to be less computationally intensive when compared with wrappers; however they develop a set of features which can be not tuned into a specific sort of predictive model. Hybrid methods certainly are a mixture of filter and also wrapper methods simply by using a filter technique to reduce search space which is to be considered through the subsequent wrapper. Thus, we will focus on the Hybrid method in this paper.

2. Related Work

2.1 Clustering Data

Cluster analysis groups data objects based on only information found in the data, that describes the object and their relationship .The goal is that the objects within a group be similar to one another and different from the objects in other groups. The greater the similarity within a group and the greater difference between groups, the better or more distinct the clustering.

2.2 Clustering High Dimensional Data

Clustering high-dimensional data is the cluster analysis of data with anywhere from a few dozen to many thousands of dimensions. Such high-dimensional data spaces are often encountered in areas such as medicine, where DNA,microarray technology can produce a large number of measurements at once, and the clustering of text documents, where, if a word-frequency vector is used, the number of dimensions equals the size of the dictionary.

2.3 Irrelevant Features

Irrelevant features provide no useful information in any context. The process of identifying and removing as many irrelevant data features as possible. This is because irrelevant features do not contribute to the predictive accuracy. By removing irrelevant data you get immediate gains such as increased query performance and reduced storage requirements.

2.4 Redundant Features

Redundant features are the type which usually provides no more information compared to the currently selected features. Data redundancy brings about data anomalies as well as corruption and usually needs to be avoided simply by design. Database normalization prevents redundancy as well as definitely makes the most beneficial use of storage. Redundant features usually do not redound that will get a much better predictor for the they provide mostly information which can be already contained in other feature(s). After that redundancy of data can be a known source of inconsistency, considering that customer might appear with some other values with regard to given attribute. Redundant features also affect the speed as well as accuracy of learning algorithms.

2.5 Feature Selection

Feature selection in supervised learning has been well studied, where the main goal is to find a feature subset that produces higher classification accuracy. Feature selections in unsupervised learning, learning algorithms are designed to find natural grouping of the examples in the feature space. Thus feature selection in unsupervised learning aims to find a good subset of features that forms high quality of clusters for a given number of clusters. In the search process identify three parts: The choice of a starting point, the process of generating the next set to explore, a stopping criterion. In Relief algorithm proposed measure can be better than the wrapper approach to guide a common feature selection search process. The performance of the measure to guide the search is evaluated by using a greedy search method. Greedy search strongly relies on the measure to select the search path, as only one path will be explored with no possible back-track. The algorithm is simple, easy to implement and computationally efficient. The computational cost is low in this algorithm. The Relief Algorithm is ineffective in removing redundant features. A novel concept, predominant correlation, and propose a fast filter method which can identify relevant features as well as redundancy. FCBF (Fast Correlation Based Feature Selection) algorithm is a fast filter method. FCBF algorithm selects the good subset of features by removing both irrelevant and redundant features.

The algorithm finds a set of predominant features, for finding a set of predominant feature the algorithm consists of two major parts. In the first part it calculates the SU value for each feature, selects relevant features list based on the predefined threshold, and orders them in descending order according to their SU values. In the second part it further processes the ordered list to remove redundant features and only keeps predominant ones among all the selected relevant features.FCBF achieves the highest level of dimensionality reduction by selecting the least number of features. FCBF can remove a large number of features that are redundant. FCBF is practical for feature selection for classification of high dimensional data. All features contain numerical values calculations. FCBF algorithm does not provide pair wise correlation among the relevant features. FCBF is usually a fast filter method which may identify relevant features together with redundancy between relevant features without pair wise correlation analysis. CMIM iteratively picks characteristics which usually maximize their particular mutual information with all the class to predict, conditionally towards the response of any kind of feature already picked. Totally different from these algorithms, the actual proposed SWIFT algorithm utilizes clustering based technique to choose features. Recently, hierarchical clustering has become adopted within word selection within the context of text classification. Distributional clustering has been utilized to cluster words into groups based mostly often for their participation specifically grammatical relations for some other words by Pereira et al. or about the distribution regarding class labels connected with each word simply by Baker as well as McCallum.

As distributional clustering associated with words are agglomerative in nature, and result in sub-optimal word clusters as well as high computational cost, Dhillon et al. proposed a new information-theoretic divisive algorithm with regard to word clustering as well as applied this to text classification. Butterworth et al. proposed in order to cluster features utilizing a special metric of Barthelemy-Montjardet distance, after which works by using the dendrogram on the resulting cluster hierarchy to find the most recent attributes. Unfortunately, the actual cluster evaluation measure depending on Barthelemy-Montjardet distance isn't going to identify a feature subset which allows the classifiers to enhance their original performance accuracy. Furthermore, even in comparison with other feature selection methods, the obtained accuracy is lower. Hierarchical clustering also offers been utilized to select features upon spectral data.

3. Proposed Work

3.1 Feature Selection Algorithms and Flow Diagram

The Irrelevant features, together with redundant features, severely affect the particular accuracy of the learning machines. Hence, feature subset selection must be able to identify and remove because the irrelevant as well as redundant information as is possible. Moreover, “good feature subsets include features highly correlated together with (predictive of) the class, yet uncorrelated along with (not predictive of) one another. Keeping these types of in mind, we created a novel algorithm which may efficiently and effectively deal with equally irrelevant as well as redundant features, and find a good feature subset.

Algorithm: PCA

Input: Data Matrix

Output: Reduced set of features

Step-1: X ← Create N x d data matrix, with one row vector xn per data point.
Step-2: X subtract mean *x* from each row vector *xn* in X.
Step-3: Σ ← covariance matrix of X.
Step-4: Find eigenvectors and eigen values of Σ.
Step-5: PC's ← the M eigenvectors with largest eigen values.
Step-6: Output PCs.

ALGORITHM : SWIFT

inputs: D(F1, F2, ..., Fm, C) - the given data set
ɵ - the T-Relevance threshold.
output: S - selected feature subset .

Step 1: Irrelevant Feature Removal

```
for i = 1 to m do
    T-Relevance = SU (Fi,C)
    if T-Relevance > ɵ then
        S = S ∪ {Fi};
```

//== Step 2: Minimum Spanning Tree Construction ====

```
G = NULL; //G is a complete graph
for each pair of features {F′i, F′ } ⊂ S do
    F-Correlation = SU (F′ , F′j )
    Add F′i and/or F′j to G with F-Correlation as the weight of The corresponding edge;
minSpanTree = Prim (G); //Using Prim Algorithm to generate the minimum spanning tree

Forest = minSpanTree
for each edge Eij ∈ Forest do
    if SU(F′i, F′j ) < SU(F′i, C) ∧ SU(F′i, F′j ) < SU(F′j,C) then
        Forest = Forest − Eji
S = Φ
for each tree Ti ∈ Forest do
    FjR = argmax F′K∈Ti SU(F′k,C)
    S = S ∪ {Fj };
return  S
```

Step 2: Minimum Spanning Tree Construction

Given SU(X, *Y*) the symmetric uncertainty of variables X and Y the correlation *F*-Correlation between a pair of features can be defined as follows.

Definition 2*:* (F-Correlation*)* The correlation between any pair of features Fi and Fj (Fi, Fj ∈ F ∧ i ≠ j) is called the F-Correlation of Fi and Fj , and denoted by SU(Fi, Fj).

PRIMS ALGORITHM:

A Minimum Spanning Tree in an undirected connected weighted graph of a spanning tree of minimum weight among all spanning trees.

Grow a MST:

- Start by picking any vertex to be the root of the tree.
- While the tree does not contain all vertices in the graph
- find shortest edge leaving the tree and add it to the Tree

3.2 Flow Diagram

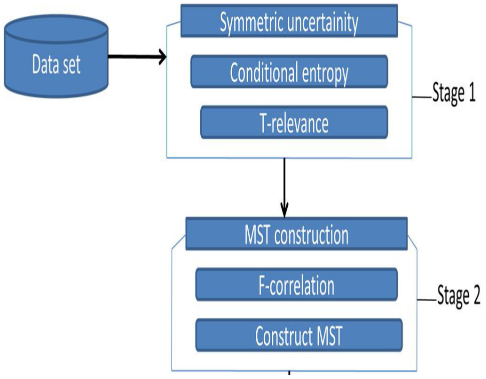

Figure 1 Flow Diagram

4. Dataset Description

We have used two dataset from the UCI repository of which first is LUNG CANCER dataset with 32 instances and 57 attributes(1 class attribute, 56 predictive) with attribute class label in which all predictive attributes are nominal, taking on integer values 0-3 and Class Distribution as class 1 with 9 observation and class 2 with 13 observations and class 3 with 10 observations and second is LIBRAS Movement Database with 360 (24 in each of fifteen classes) instances and 90 numeric (double) and 1 for the class (integer) attributes with class distribution of 6.66% for each of 15 classes.

5. Results and Discussion

As proposed the swift clustering algorithm and PCA is implemented in Net beans IDE. To evaluate this two datasets have been used and the outputs are tabulated in the below results.

DATASET	LUNG CANCER	LIBRAS MOVEMENT
ATTRIBUTES	56	90
INSTANCES	32	360
SWIFT OUTPUT	7	2
PCA OUTPUT	21	9

6. Conclusion

Feature selection method is an efficient way to improve the accuracy of classifiers, dimensionality reduction, removing both irrelevant and redundant data. Thus SWIFT algorithm selects only fewer and relevant features which adds to the classifier accuracy when compared with PCA as shown in table 1. For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space.

References

[1] Bell D.A. and Wang, H., A formalism for relevance and its application in feature subset selection, Machine Learning, 41(2), pp 175-195, 2000.

[2] Biesiada J. and Duch W., Features election for highdimensionaldatała Pearson redundancy based filter,AdvancesinSoftComputing, 45, pp 242C249,2008.

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

[4] Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009.

[5] Cohen W., Fast Effective Rule Induction, In Proc. 12th international Conf. Machine Learning (ICML'95), pp 115-123, 1995.

[6] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.

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

[8] Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.

[9] Dash M. and Liu H., Consistency-based search in feature selection.Artificial Intelligence, 151(1-2), pp 155-176, 2003.

[10] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach.Learn. Res., 3, pp 1265-1287, 2003.

[11] Dougherty, E. R., Small sample issues for microarray-based classification. Comparative and Functional Genomics, 2(1), pp 28-34, 2001.

[12] Fayyad U. and Irani K., Multi-interval discretization of continuousvalued attributes for classification learning, In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence,pp 1022-1027, 1993.

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

[14] Forman G., An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 3, pp 1289-1305, 2003.

[15] Guyon I. and Elisseeff A., An introduction to variable and feature selection, Journal of Machine Learning Research, 3, pp 1157-1182,2003.

[16] Hall M.A., Correlation-Based Feature Subset Selection for MachineLearning, Ph.D. dissertation Waikato, New Zealand: Univ. Waikato,1999.

[17] 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 Twelth nternational Florida Artificialintelligence Research Society Conference, pp 235-239, 1999.

[18] Hall M.A., Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, In Proceedings of 17th International Conference on Machine Learning, pp 359-366, 2000

[19] Jaromczyk J.W. and Toussaint G.T., Relative Neighborhood Graphs and Their Relatives, In Proceedings of the IEEE, 80, pp 1502-1517,1992.