International Journal of Computer Science
Oddity...Probe...Reviste...

http://www.ijcsjournal.com Volume 2, Issue 2, No 4, 2014. ISSN: 2348-6600
Reference ID: IJCS-070 PAGE NO: 403-406.

Cluster Based Feature Subsection Algorithm for High-Dimensional Data

Hemanth Kalthireddy[#1], P Muni Sekhar[*2]

[#]*Computer Science & Engineering, JNTU Anantapur*
[1]khemanth1310@gmail.com
[*]Shree Institute Of Technology
[2] munisekhar.prudhvi@gmail.com

***Abstract*— The feature subsection is an effective way for removing irrelevant data, reducing dimensionality, improving result comprehensibility, and increasing learning accuracy. Feature Subsection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature Subsection algorithm may be evaluated from both the efficiency and effectiveness points of view. A feature selection algorithm can be evaluated from both the efficiency and effectiveness points of view. While the efficiency apprehensions the time required to find a subsection of features, the effectiveness is related to the quality of the subsection of features. Based on these criteria, a cluster based feature selection algorithm (CFSA) is proposed and experimentally evaluated in this paper. The CFSA algorithm works in two phases. In the first phase, features are divided into clusters by using graph-theoretic clustering methods. In the second phase, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in dissimilar clusters are relatively independent; the cluster based strategy of CFSA has a high probability of producing a subset of useful and independent features. To ensure the efficiency of CFSA, we implement the efficient minimum-spanning tree (MST) clustering method. Efficiency and effectiveness of the CFSA algorithm are evaluated through an experimental study.**

***Index Terms*—Feature Selection, high dimensional data, Clustering methods, Minimal Spanning Tree.**

I. INTRODUCTION

The performance, robustness, and usefulness of classification algorithms are improved when relatively few features are involved in the classification. Hence, selecting relevant features for the construction of classifiers has received a great deal of attention. From the aim of choosing a subset of good features with respect to target concepts, the feature subsection is an effective way for removing irrelevant data, reducing dimensionality, improving result comprehensibility, and increasing learning accuracy. So-many feature subsection methods had been proposed and studied for the machine learning applications. They can be divided into four broad categories: Wrapper, the Embedded, Hybrid, and Filter approaches. Traditional machine learning algorithms such as artificial neural networks or decision trees are examples of embedded approaches. The embedded methods incorporate feature subset selection as a portion of the training process and are usually specific to given learning algorithms, therefore may be more efficient than the other three categories. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, accuracy of the learning algorithms is usually more. However, the generality of the selected features are limited and the computational complexity is large. The filter methods were independent of learning algorithms, with good overview. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of wrapper and filter methods by using a filter method to reduce search space, which will be considered by the subsequent wrapper. They mainly focus on combining filters and wrapper methods to achieve best possible performance with a particular learning algorithm with alike time complexity of the filter methods. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. The wrapper methods are tending to over fit on small training sets and computationally expensive. Thus, we will focus on the filter method in this paper.

II. LITERATURE SURVEY

A. Statistical Comparisons of Classifiers over Multiple Data Sets

In this method introduce some new pre- or post-processing step has been proposed, and implicit hypothesis is made that such an enhancement yields an improved performance over the existing classification algorithm. Alternatively, various solutions to a problem are proposed and the goal is to tell the successful from the failed. A number of test data sets are selected for testing, the algorithms run and the quality of the resulting models is evaluated using an appropriate measure, most common classification accuracy. The remaining step and the topic of this paper is to statistically verify the hypothesis of improved performance. Various re-searchers have addressed the problem of comparing two classifiers on a single data set and proposed several solutions. The main core of this paper is the study of the statistical tests that could be (or already are) used for comparing two or more classifiers on multiple data sets. Learning algorithms are used for the Classification purpose. The main disadvantage of this process is the problems with the multiple data set tests are quite different, even in a sense complementary.

B. Feature Clustering and Mutual Information for the Selection of Variables In Spectral Data

It face many problems in spectrometry require predicting a quantitative value from measured spectra. The major issue with spectrometric data is their functional nature; they are functions discredited with a high resolution. This leads to a large number of highly-correlated features; many of which are irrelevant for the prediction. The approach for the features is to describe the spectra in a functional basis whose basis functions are local in the sense that they correspond to well-defined portions of the spectra. This process has clustering algorithm that algorithm recursively merges at each step the two most similar consecutive clusters. This algorithm return the output value associated with each cluster, it is representative, chosen to be the mean of the spectra over the range of features defined by the cluster. Main disadvantage of the problem is low number of clusters identified by the method allows the interpretation of the selected variables: several of the selected clusters include the spectral variables identified on these benchmarks as meaningful in the literature.

C. A Features Set Measure Based On Relief

It used six real world dataset from the UCI repository have been used. Three of them have classification Problem with discrete features, next two classifications with discrete and continuous features, and the last one is approximation problem. Learning algorithm is used to check the quality of feature selected are a classification and regression tree layer with pruning. Hence this process and algorithms is implemented by the orange data mining System. The non-parametric tests, namely the Wilcox on and Friedman test are suitable for our problem. Those are appropriate since they assume some, but limited commensurability. They are safer than parametric tests since they do not assume normal distributions or homogeneity of variance. There is an alternative opinion among statisticians that significance tests should not be per-formed at all since they are often misused, maybe either due to misinterpretation or by putting too much stress on their results The main disadvantage of the system is it measure to low accuracy of the search process.

III. CLUSTER BASED FEATURE SELECTION ALGORITHM

A. Framework

Irrelevant features, along with redundant features, severely affect accuracy of the learning machines. Thus, feature sub selection should be able to identify and clear redundant and irrelevant information as possible. However, good feature subsets contain features (predictive of) highly correlated with the class, yet uncorrelated with (not predictive.

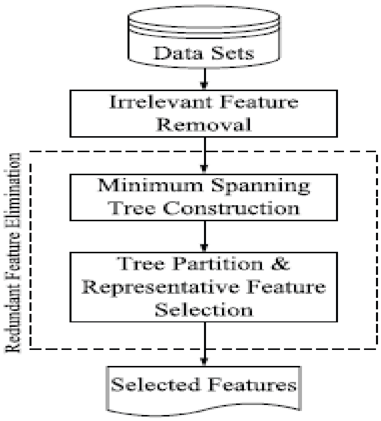

Data Sets
Irrelevant Feature Removal
Minimum Spanning Tree Construction
Tree Partition & Representative Feature Selection
Redundant Feature Elimination
Selected Features

Fig 1: Framework proposed for CFSA Algorithm

International Journal of Computer Science
Oddity...Probe...Reviste...

http://www.ijcsjournal.com Volume 2, Issue 2, No 4, 2014. ISSN: 2348-6600
Reference ID: IJCS-070 PAGE NO: 403-406.

By this, we developed a novel algorithm which will 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 (shown in Fig.1) which composed of the two connected components of redundant feature elimination and irrelevant feature removal. The first component obtains features related to the target concept by eliminating irrelevant ones, and the second component removes redundant features from relevant ones by choosing representatives from different feature clusters, and hence produces the final subset.

Whereas the irrelevant feature removal is straightforward once the right relevance measure is defined/selected, while the redundant feature removal is a bit of difficult. In our proposed CFSA algorithm, it involves (a) the construction of the minimum spanning tree (MST) from a weighted complete graph (b) the partitioning of the Minimum Spanning Tree into forest with each tree representing a cluster; and (c) the selection of representative features from the clusters.

B. Algorithm and Analysis

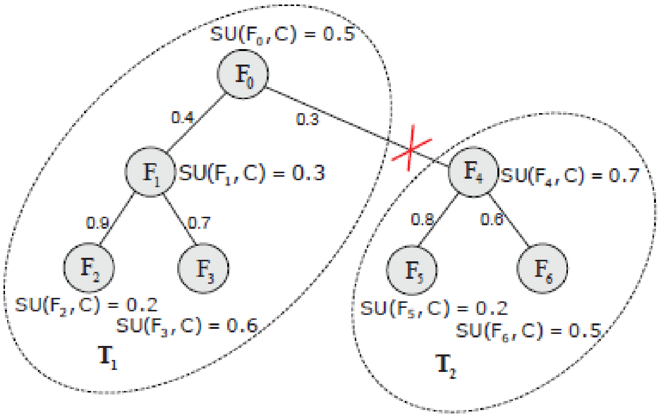
The proposed CFSA algorithm logically consists of three steps: (a) removing irrelevant features, (b) constructing a MST from relative ones, and (c) partitioning the MST and selecting representative features.

For a data set D with m features $F = \{F1, F2, ...,Fm\}$ and class C, we compute the T-Relevance $SU(Fi, C)$ value for each feature Fi ($1 \leq i \leq m$) in the first step. The features whose $SU(Fi, C)$ values are greater than a predefined threshold θ comprise the target-relevant feature subset $F' = \{F'1, F'2, ..., F'n\}$ ($n \leq m$).

In the second step, we first calculate the F-Correlation $SU(F'x, F'y)$ value for each pair of features $F'i$ and $F'y$ ($F'x, F'y \in F' \wedge x \neq y$). Later, viewing features $F'x$ and $F'y$ as vertices and $SU(F'x, F'y)$ ($x \neq y$) as weight of the edge between vertices $F'x$ and $F'y$,one weighted complete graph $G = (V,E)$ is constructed where $V = \{F'x| F'x \in F' \wedge x \in [1, n]\}$ & $E = \{(F'x, F'y) \mid (F'x, F'y \in F' \wedge x, y \in [1, n] \wedge x \neq y\}$. Symmetric uncertainty is symmetric further the F-Correlation $SU(F'x, F'y)$ is symmetric also, thus G is an undirected graph.

The complete graph G reflects correlations among all target-relevant features. Inappropriately, graph G has n vertices and $n(n-1)/2$ edges. For multi-dimensional data, it is heavily dense and the edges with different weights are strongly interweaved. However, the decomposition of complete graph is NP-hard. Hence for graph G, we build a Minimum Spanning Tree, which connects all vertices in such a way that the sum of the weights of the edges is the minimum, using the Prim algorithm. The weight of edge ($F'x$, $F'y$) is F-Correlation $SU(F'x, F'y)$.After building the MST, in the third step, we first remove the edges $E = \{(F'x, F'y)|(F'x, F'y \in F' \wedge x, y \in [1, n] \wedge x \neq y\}$, whose weights are smaller than both of the T-Relevance $SU(F'x, C)$ and $SU(F'y, C)$, from the MST.

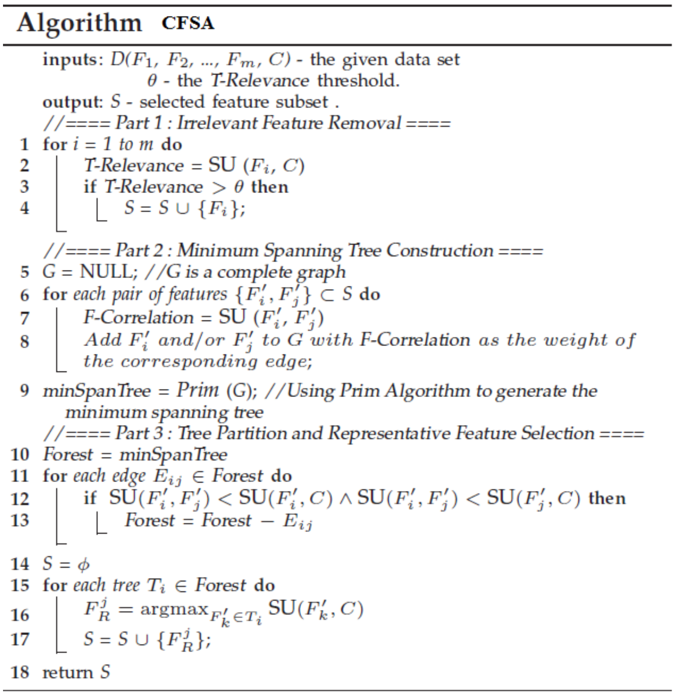
Each deletion results into two dis-connected trees $T1$ and $T2$. Assume the set of vertices in any one of the final trees to be V (T), we would have the property that for each pair of vertices ($F'x, F'y \in V$ (T)), $SU(F'x, F'y) \geq SU(F'y, C) \vee SU(F'x, F'y) \geq SU(F'y, C)$ always holds. As this property guarantees the features in V (T) are redundant. This can be illustrated with one example. Assume the MST shown in Fig.2 is generated from a complete graph G. In order to cluster the features, first we traverse all the six edges, and then decide to remove the edge ($F0$, $F4$) because its weight $SU(F0, F4) = 0.4$ was smaller than $SU(F0, C) = 0.6$ and $SU(F4, C) = 0.8$. This makes the MST is clustered into two clusters denoted by V ($T1$) and V ($T2$). Each cluster is a MST. Take V ($T1$) as example.



We know $SU(F0, F1) > SU(F1, C)$, $SU(F1, F2) > SU(F1, C) \wedge SU(F1, F2) > SU(F2, C)$, $SU(F1, F3) > SU(F1, C) \wedge SU(F1, F3) > SU(F3, C)$. We also observed that there is no edge exists between $F2$ and $F0$, $F3$ and $F0$, and $F3$ and $F2$. Considering that $T1$ is a MST, so the $SU(F0, F2)$ is greater than $SU(F0, F1)$ and $SU(F2, F1)$, $SU(F3, F0)$ was greater than $SU(F1, F0)$ and $SU(F3, F1)$, and $SU(F3, F2)$ is greater than $SU(F1, F2)$ and $SU(F2, F3)$. Hence, $SU(F0, F2) > SU(F0, C) \wedge SU(F0, F2) > SU(F2, C)$, $SU(F0, F3) > SU(F0, C) \wedge SU(F3, F0) > SU(F3, C)$,and $SU(F2, F3) > SU(F2, C) \wedge SU(F2, F3) > SU(F3, C)$ also hold. As the mutual information between any pair $(Fi, Fj)(i, j = 0, 1, 2, 3 \wedge i \neq j)$ of $F3$, $F2$, $F1$, and $F0$ was greater than mutual information between class C and Fi or Fj,features $F3$, $F2$, $F1$ and $F0$ are redundant.

After removing all the unnecessary edges, a forest *Forest* is obtained. Each tree $T_j \in Forest$ represents a cluster that is denoted as $V(T_j)$, which is the vertex set of T_j as well. As illustrated above, the features in each cluster are redundant, so for each cluster $V(T_j)$ we choose a representative feature F_{jR} whose *T-Relevance* $SU(F_{jR}, C)$ is the greatest. All F_{jR} (j = 1... /*Forest*/) comprise the final feature subset $\cup F_{jR}$.

The details of the CFSA algorithm is shown below



```
Algorithm CFSA
  inputs: $D(F_1, F_2, ..., F_m, C)$ - the given data set
          $\theta$ - the T-Relevance threshold.
  output: $S$ - selected feature subset .
  //==== Part 1 : Irrelevant Feature Removal ====
  for $i$ = 1 to $m$ do
      T-Relevance = SU $(F_i, C)$
      if T-Relevance $> \theta$ then
          $S = S \cup \{F_i\}$;
  //==== Part 2 : Minimum Spanning Tree Construction ====
  $G$ = NULL; //G is a complete graph
  for each pair of features $\{F'_i, F'_j\} \subset S$ do
      F-Correlation = SU $(F'_i, F'_j)$
      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
  //==== Part 3 : Tree Partition and Representative Feature Selection ====
  Forest = minSpanTree
  for each edge $E_{ij} \in$ Forest do
      if $\mathrm{SU}(F'_i, F'_j) < \mathrm{SU}(F'_i, C) \wedge \mathrm{SU}(F'_i, F'_j) < \mathrm{SU}(F'_j, C)$ then
          Forest = Forest $- E_{ij}$
  $S = \phi$
  for each tree $T_i \in$ Forest do
      $F^j_R = \mathrm{argmax}_{F'_k \in T_i} \mathrm{SU}(F'_k, C)$
      $S = S \cup \{F^j_R\}$;
  return $S$
```

IV. Conclusions

In this paper, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. Algorithm involves (a) removing irrelevant features (b) constructing a minimum spanning tree from relative ones and (c) partitioning the MST and selecting representative features. In this proposed algorithm, a cluster consists of feature. Every cluster is treated as a single feature and thus dimensionality is drastically resized. Generally, proposed algorithm obtained the best proportion of selected features, best runtime, and the best classification accuracy of C4.5, Naive Bayes and RIPPER, and the second best classification accuracy for IB1. For the future work, we plan to explore different types of correlation measures and study some formal properties of feature space

References

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data Qinbao Song, Jingjie Ni and Guangtao Wang,2013.

[3] Pereira F., Tishby N. and Lee L., Distributional clustering of English words, In Proceedings of the 31st Annual Meeting on Association For Computational Linguistics, pp 183-190, 1993.

[4] Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T., Numerical recipes in C. Cambridge University Press, Cambridge, 1988.Prim R.C., Shortest connection networks and some generalizations, Bell System Technical Journal, 36, pp 1389-1401.

[5] Quinlan J.R., C4.5: Programs for Machine Learning. San Mateo, Calif: Morgan Kaufman, 1993.

[6] Raman B. and Ioerger T.R., Instance-Based Filter for Feature Selection.

[7] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.

[8] Almuallim H. and Dietterich T.G., Learning boolean concepts in the Presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279- 305, 1994.

[9] 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, pp 104-109, 2004.

[10] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96-103, 1998.

[11] Battiti R., Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537- 550, 1994.

[12] 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.

[13] Biesiada J. and Duch W., Features election for high-dimensionaldata൞ Pearson redundancy based filter, AdvancesinSoftComputing, 45, pp 242C249, 2008.

[14] Fisher D.H., Xu L. and Zard N., Ordering Effects in Clustering, In Proceedings of the Ninth international Workshop on Machine Learning, pp 162-168, 1992.

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