International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 10 Issue 1 –OCTOBER 2014.

A Survey of Different Feature Selection Algorithms with Fast

N.Shirisha[1] and D.L.Padmaja[2]

[1]*M.Tech (IT), C.V.S.R College of Engineering, Hyderabad, Telangana, India*

[2]*Associate Professor, Department of IT, C.V.S.R College of Engineering, Hyderabad, Telangana, India*

Abstract— Feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving results comprehensibility. This process improved by cluster based FAST Algorithm and FAST Algorithm can be used to identify and removing the irrelevant data set. This algorithm process implements using two different steps that is graph theoretic clustering methods and representative feature cluster is selected. Feature subset selection research has focused on searching for relevant features. The Features in unusual clusters are relatively self governing the clustering-based approach of FAST has elevated possibility of producing a subset of useful features, in the direction of guarantee to the efficiency of FAST using minimum spanning tree clustering technique. general experiments are approved to contrast FAST and some delegate features selection algorithms, namely, FCBF, Relief F, CFS, Consist, and FOCUS-SF, by admiration to four types of famous classifiers, specifically, the probability-based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER.

Index Terms - Feature subset selection, Clustering

I. INTRODUCTION

Out of so many feature subset selection algorithms, a few can successfully remove immaterial features but not succeed to hold unnecessary features, however a few of others can remove the immaterial while taking concern of the unnecessary features. FAST algorithm cascade into the subsequent group. The performance, robustness, and usefulness of classification algorithms are improved when relatively few features are involved in the classification. Thus, selecting relevant features for the construction of classifiers has received a great deal of attention. With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected sub-sets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited.

About Feature Selection:

The high dimensionality of data poses many challenges to learning tasks due to the curse of dimensionality. To reduce the many irrelevant features from a feature space Feature Selection is one of the effective procedures to identify relevant features for dimensionality reduction. Various studies show that features can be removed without performance deterioration.

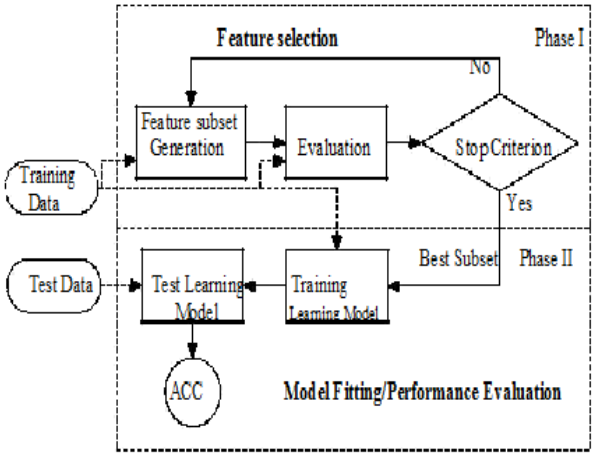

Figure: A unified view of feature selection process

II. ALOGORITHM AND ANALYSIS

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 FAST algorithm it involves, by removing the irrelevant feature from the dataset. For a data set, features and class, we compute the T-Relevance value for each feature in the first step. The features whose values are greater than a predefined threshold comprise the target-relevant feature subset.

Minimum Spanning Tree Construction

A spanning tree of that graph is a sub graph that is a tree and connects all the vertices together. A single graph can have many different spanning trees. We can also assign a weight to each edge, which is a number representing how unfavourable it is, and use this to assign a weight to a spanning tree by computing the sum of the weights of the edges in that spanning tree. A minimum spanning tree (MST) or minimum weight spanning tree is then a spanning tree with weight less than or equal to the weight of every other spanning tree. More generally, any undirected graph (not necessarily connected) has a minimum spanning forest, which is a union of minimum spanning trees for its connected components. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. In this module we construct the minimum spanning tree. Calculate the F-Correlation value for each pair of features. Then, viewing features and as vertices and as the weight of the edge between vertices, a weighted complete graph is constructed. As symmetric uncertainty is symmetric further the F-Correlation is symmetric as well, thus it forms an undirected graph. This undirected graph reflects the correlations among all the target-relevant features. 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 is F-Correlation.

Tree Partition and Representative Feature Selection:

In this module remove the edges, whose weights are smaller than both of the T-Relevance and from the MST. Each deletion results in two disconnected trees $T1$ and $T2$. After removing all the unnecessary edges, a forest is obtained. Each tree Forest represents a cluster that is denoted as $V(Tj)$, which is the vertex set of Tj as well. As illustrated above, the features in each cluster are redundant, so for each cluster we choose a representative feature whose T-Relevance is the greatest. All Forest comprises the final feature subset 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 clustering-based feature subset selection algorithm for high dimensional data.

Fast Algorithm

1. Inputs: D($F1$, $F2$, ..., Fm, C) - the given data set
2. θ - the T-Relevance threshold.
3. Output: S - selected feature subset
4. // Irrelevant Feature Removal
5. For i = 1 to m do {
6. T-Relevance = SU (Fi, C)
7. If T-Relevance > θ then
8. S = S $\cup$ $\{Fi\}$;}
9. //Minimum Spanning Tree Construction
10. G = NULL; //G is a complete graph
11. For each pair of features $\{Fi, Fj\} \subset$ S do
12. F-Correlation = SU (Fi , F' j)
13. $Add\ F'\ i\ and/or\ F\ j\ to\ G\ with$ F-Correlation as $the\ weight\ of\ the\ corresponding\ edg$e;
14. minSpanTree = Prim (G); //Using Prim Algorithm to Generate Minimum Spanning Tree
15. //Tree Partition and Representative Feature Selection
16. Forest = minSpanTree
17. For each edge $Eij \in$ Forest do
18. If SU(Fi, Fj) < SU(Fi, C) $\wedge$ SU(Fi, Fj) < SU(Fj,C) then
19. Forest = Forest – Eij
20. For each tree $Ti \in$ Forest do Fj
21. R = argmax$Fk \in Ti$ SU (F k,C)
22. S = S $\cup$ $\{FjR\}$;
23. Return S;

Prims Algorithm

1. Input: A non-empty connected weighted graph with vertices V and edges E (the weights can be negative).
2. Initialize: Vnew = {x}, where x is an arbitrary node (starting point) from V, Enew = {}
3. Repeat until Vnew = V:

a. Choose an edge {u, v} with minimal weight such that ui sin Vnew ad v is not (if there are multiple edges with the same weight, any of them may be picked)

b. Add v to Vnew, and {u, v} to Enew

4. Output: Vnew and Enew describe a minimal spanning tree.

Different Feature Selection Algorithms:

The FOCUS algorithm

The FOCUS algorithm originally defined for noise-free Boolean domains, exhaustively examines all subsets of features, selecting the minimal subset of features that is

sufficient to determine the label value for all instances in the training set. This preference for a small set of features is referred to as the MIN-FEATURES bias. This bias has severe implications when applied blindly without regard for the resulting induced concept. For example, in a medical diagnosis task, a set of features describing a patient might include the patient's social security number (SSN). (We assume that features other than SSN are sufficient to determine the correct diagnosis.) When FOCUS searches for the minimum set of features, it will pick the SSN as the only feature needed to uniquely determine the label. 3 Given only the SSN, any induction algorithm is expected to generalize very poorly.

The RELIEF algorithm

The Relief algorithm assigns a "relevance" weight to each feature, which is meant to denote the relevance of the feature to the target concept. Relief is a randomized algorithm. It samples instances randomly from the training set and updates the relevance values based on the difference between the selected instance and the two nearest instances of the same and opposite class (the "near-hit" and "near-miss"). The Relief algorithm attempts to find all relevant features:

Relief does not help with redundant features. If most of the given features are relevant to the concept, it would select most of them even though only a fraction are necessary for concept description In real domains, many features have high correlations with the label, and thus many

are weakly relevant, and will not be removed by Relief. In the simple parity example used in , there were only strongly relevant and irrelevant features, so Relief found the strongly relevant features most of the time. The Relief algorithm was motivated by nearest-neighbors and it is good specifically for similar types of induction algorithms. In preliminary experiments, we found significant variance in the relevance rankings given by Relief. Since Relief randomly samples instances and their neighbors from the training set, the answers it gives are unreliable without a large number of samples. In our experiments, the required number of samples was on the order of two to three times the number of cases in the training set.

Relief as originally described can only run on binary classification problems, so we used the Relief-F method described by Kononenko which generalizes Relief to multiple classes. We combined Relief-F with our deterministic enhancement to yield the final algorithm Relieved-F. In our experiments, features with relevance rankings below 0 were removed. ' This is true even if SSN is encoded in 30 binary features as long as more than 30 other binary features are required to determine the diagnosis. Specifically, two real-valued attributes, each one with 16 bits of precision,will be inferior under this scheme.

```
Relief(S, m, τ)
    Separate S into S+ = {positive instances} and
        S-= {negative instances}
    W = (0, 0, . . . , 0)
    For i = 1 to m
        Pick at random an instance X ∈ S
        Pick at random one of the positive instances
            closest to X, Z+ ∈ S+
        Pick at random one of the negative instances
            closest to X, Z- ∈ S-
        if (X is a positive instance)
            then    Near-hit = Z+; Near-miss = Z-
            else    Near-hit = Z-; Near-miss = Z+
        update-weight(W, X, Near-hit, Near-miss)
    Relevance = (1/m)W
    For i = 1 to p
        if (relevancei ≥ τ)
            then    fi is a relevant feature
            else    fi is an irrelevant feature

update-weight(W, X, Near-hit, Near-miss)
    For i = 1 to p
        Wi = Wi - diff(xi, near-hiti)2 + diff(xi, near-missi)2
```

Figure 1 Relief Algorithm

Relief Algorithm

Correlation-Based Feature Selection (CFS):

CFS searches feature subsets according to the degree of redundancy among the features. The evaluator aims to find the subsets of features that are individually highly correlated with the class but have low inter-correlation. The subset evaluators use a numeric measure, such as conditional entropy, to guide the search iteratively and add features that have the highest correlation with the class. The downside of univariate filters for eg information gain is,it does not account for interactions between features, which is overcome by multivariate filters for eg CFS. CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Correlation coefficients are used to estimate correlation between subset of attributes and class, as well as inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation. CFS is used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best-first search and genetic search.

Equation for CFS is given.

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}}$$

where rzc is the correlation between the summed feature subsets and the class variable, k is the number of subset features.

rzi is the average of the correlations between the subset features an the class variable, and rii is the average inter-correlation between subset features.

Fast Correlation Based Fs (FCBF):

FCBF (Yu and Liu, ICML 2003) uses also the symmetrical uncertainty measure. But the search algorithm is very different. It is based on the "predominance" idea. The correlation between an attribute X* and the target Y is predominant if and only if $\rho y,x^* \geqslant \delta$ et $\nexists X (X \neq X^*)$, $\rho x,x^* < \rho y,x^*$

Concretely, a predictor is interesting if its correlation with the target attribute is significant (delta is the parameter which allows to assess this one); there is no other predictor which is more strongly correlated to it.

Based on this idea, the authors propose a search algorithm which runs in quasilinear time.

Algorithm for Fcbf:

1. S is the set of candidate predictors, M = ∅ is the set of selected predictors
2. Searching X* (among S) which maximizes its correlation with $Y \rightarrow \rho y,x^*$
3. If $\rho y,x^* \geqslant \delta$ add X* into M and remove X* from S
4. Remove also from S all the variables X such $\rho x,x^* \geqslant \rho y,x^*$ (Very important !)
5. If S ≠∅ then GOTO (2), else END of the algorithm

III. Conclusion

The task of a feature selection algorithm is to provide with a computational solution to the feature selection problem motivated by a certain definition of relevance. This algorithm should be reliable and efficient. The algorithm includes (i) eliminating non-relevant features, (ii) constructing a minimum spanning tree from comparative ones, and (iii) MST is portioned and then choosing representative features. The cluster consists of features. Each cluster is considering as a single feature and thus dimensionality is reduced. The performance of the proposed algorithm can be compared with five famous feature selection algorithms. They are considering as FCBF, ReliefF, CFS, Consist, and the FOCUS-SF on various data sets like image, microarray and text data from the four different aspects of the section of the selected features, the proposed algorithm having the best selected features, best runtime, and also having the best classification correctness for Naive Bayes,C4.5, and RIPPER. With the FAST algorithm it is easy to originate the rank of 1 for microarray data, the rank of 2 for text data, and the rank of 3 for image data in terms of classification correctness of the four different types of classifiers.

IV. FUTURE WORK

For the future work, we plan to explore different types of correlation measures, and study of some formal properties of feature space.

V. REFERENCES

[1] The International Journal Of Engineering And Science (IJES) ||Volume||2 ||Issue|| 10||Pages|| 27-38||2013|| ISSN(e): 2319 – 1813 ISSN(p): 2319 – 1805 www.theijes.com The IJES Page 27

[2] A Survey On Feature Selection Algorithm For High Dimensional Data Using Fuzzy Logic T.Jaga Priya Vathana, 1 C.Saravanabhavan 2, Dr.J.Vellingiri3.

[3] L.Ladha et al. / International Journal on Computer Science and Engineering (IJCSE), Feature Selection Methods and Algorithms Ladha, Research Scholar, T.Deepa, Lecturer.

[4] Department of Computer ScienceHamilton, NewZealand Correlation-based Feature Selection for Machine Learning Mark A. Hall.

[5] IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 2, Apr-May, 2014 ISSN: 2320 – 8791 www.ijreat.org Cluster Based Speed and Effective Feature Extraction for Efficient Search Engine Manjuparkavi A1, Arokiamuthu M2 Feature Subset Selection Algorithm for Elevated Dimensional Data By using Fast Cluster

[6] B.Swarna Kumari M., M.Doorvasulu International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 7 13 Feature Subset Selection Problem Wrapper Approach in Supervised Learning al of Artificial Intelligence Research 47 (2013) 1-34 Submitted 10/2012; published 05/2013 A Feature Subset Selection Algorithm Automatic Recommendation Method,Guangtao Wang

[7] Wrappers for feature subset selection Ron Kohavi a,*, George H. John b,l a Data Mining and Visualization, Silicon Graphics, Inc., 2011 N. Shoreline Boulevard, Mountain view, CA 94043, US b Epiphany Marketing Sofhyare, 2141 Landings Drive, Mountain View, CA 94043, USA.

[8] IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL: 25 NO:1 YEAR 2013, A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data Qinbao Song,

[9] Jingjie Ni and Guangtao Wang.A Practical Approach to feature selection,Kenji kira,Larry A.RendellZheng Zhao, Huan Liu et al. Advanced Feature Selection research.

[10] www.wikipedia.org