Feature Subset Selection Algorithm for High-Dimensional Data by using FAST Clustering Approach

Kumaravel.V
Department of Computer Science and Engineering.
Tagore Institute of Engineering *and Technology, Attur-636112, Tamil Nadu, India.*

Raja.K
Department of Computer Science and Engineering.
Tagore Institute of Engineering and Technology, Attur-636112, Tamil Nadu, India.

ABSTRACT
Feature selection involves the process of identifying the most useful feature's subset which produces compatible results similar to original set of feature. Efficiency and effectiveness are the two measures to evaluate feature selection algorithm. The time to find the cluster concerns to efficiency, while effectiveness is concerned to quality of subset feature. With these criteria, fast clustering algorithm was proposed and experimented in two steps. Features are divided into cluster in first step and followed by selection representative feature related to the target class from each cluster. Fast algorithm has the probability of producing a useful and independent feature subset. Performance of this algorithm is evaluated against several selection algorithms (FCBF, Relief, and CFs) and it outperforms the other algorithm. The result analyzed from 35 real world dataset (image, microarray, text data) proves not only that FAST produces smaller subset but also improves the performance.

General Terms
Feature selection, irrelevant, redundant,

Keywords
Feature selection, filter method, Feature clustering, graph-based clustering.

1. INTRODUCTION
In machine learning feature selection is known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. Feature subset selection is a useful way for reducing dimensionality, removing irrelevant data, increasing learning accuracy [1]. Numerous feature subset selection methods have been planned and studied for machine learning applications. They can be classified into four broad categories: Embedded, Wrapper, Filter, and Hybrid approaches.

The embedded strategies includes feature selection as a part of the training process and are sometimes more specific to given learning algorithms, and so could also be additionally efficient than the other three classes [2]. Ancient machine learning algorithms like decision trees or artificial neural networks are samples of embedded approaches [3]. The wrapper strategies use the prophetical accuracy of a predetermined learning algorithm to work out the goodness of the chosen subsets, the accuracy of the learning algorithms is sometimes high and computational complexness is large.

The filter is a preprocessing step, which is independent of learning algorithms with sensible generality. Their computing complexity is low, however the accuracy of the learning algorithms isn't secured [4], [5], [6]. The hybrid method effectively merges of filter and wrapper method [7], [8], [9]. Regarding to the filter feature selection methods, the application of cluster analysis has been established to be more effective than the existing traditional feature selection algorithms. Pereira et al. [10], Baker and McCallum [11], and Dhillon et al. [12] used the distributional clustering of words to reduce the dimensionality among text data. In cluster analysis, graph-theoretic methods are well studied and utilized in several applications. Their results have, sometimes, the simplest agreement with human performance [13]. The result is a forest and every tree within the forest represents a cluster. In this study, a graph theoretic clustering method was proposed and tends to adopt MST based clustering algorithm by assuming that the points are classified around centers or separated by a regular geometric curve and are widely utilized in practice.

Based on the MST method, Fast clustering bAsed feature Selection algoriThm (FAST) is proposed. The FAST algorithm works in 2 steps. First, features are divided into clusters by using graph-theoretic clustering methods. Second, the foremost representative feature that is powerfully associated with target classes from every cluster to create the final set of features. Features in different clusters are comparatively independent, the clustering based strategy of FAST features a high probability of producing a set of helpful and independent features. The projected feature subset selection algorithm FAST was tested upon 35 in public available image, microarray, and text data sets. By comparing the experimental results of five different kind of feature subset selection algorithm, the proposed algorithm not only reduces the amount of features but also improves the performance.

2. RELATED WORK
The feature subset selection is the process of identifying and removing as many irrelevant and redundant features as possible. This can be as a result of irrelevant features do not assign to the predictive accuracy [14], and redundant features do not add to getting a better predictor for that they provide mostly information which is already present in other feature(s) of the numerous feature set selection algorithms, some will effectively eliminate irrelevant features however fail to handle redundant options [15], [16], [17]. However some number of others will eliminate the impertinent whereas taking care of the redundant options [18], [19].

The proposed FAST algorithm comes under the second type. Historically, feature subset selection analysis has centered on sorting out relevant options. A widely known example is Relief, which measure every feature according to its ability to discriminate instances under different targets supported distance-based criteria.

The relevant and redundant features additionally have an effect on the speed and accuracy of learning algorithms. CFS [20] is achieved by the hypothesis that an honest feature set is one that contains features extremely related to with the target, however unrelated with one another. FCBF ([21], [22]) is a quick filter method which may determine relevant features yet as redundancy among relevant features exist. CMIM [23] iteratively picks features that increase their mutual information with the category to predict, not absolutely to the response of any feature already picked. The planned FAST algorithm employs the clustering-based methodology to choose features.

Hierarchal clustering is adopted in word selection within the context of text classification (e.g., [10],[11], and [12]). Distributed cluster does not cluster words into group based either on their participation in particular grammatical relations with alternative words by Pereira et al. [10] or on the distribution of sophistication labels associated with every word by Baker and McCallum [11]. As distributional cluster of words area unit agglomerated in nature, and lead to suboptimal word clusters and high computational value, Dhillon et al. [12] planned a replacement information-theoretic discordant algorithmic rule for word cluster and applied it to text classification.

Quite totally different from these hierarchal clustering-based algorithms, the planned FAST algorithm uses minimum spanning tree-based methodology to cluster features. The planned FAST doesn't limit to some specific types of data.

3. FEATURE SUBSET SELECTION ALGORITHMS

3.1 Framework and Definitions

Irrelevant features with redundant features severely have an effect on the accuracy of the learning machines [16], [23]. Keeping these in mind, a unique algorithm is proposed which may efficiently and effectively contend with each irrelevant and redundant features and procure an good feature set.

A brand new feature selection framework (shown in Fig. 1) that composed of the 2 connected components of irrelevant feature removal and redundant feature elimination. The previous obtains features relevant to the target concept by eliminating irrelevant ones, and therefore the latter removes redundant features from relevant ones via selecting representatives from completely different feature clusters, and thus produces the ultimate set.

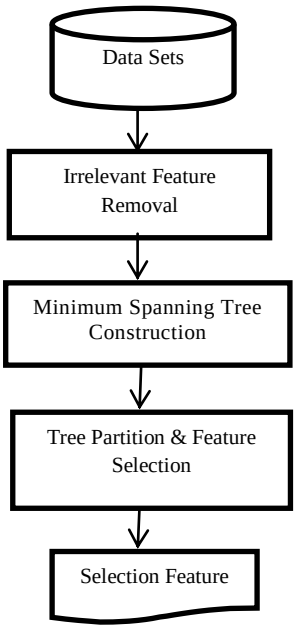

Fig 1: Proposed feature subset algorithm

In the proposed algorithm, it involves 1) the development of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with every tree representing a cluster; and 3) the selection of representative features from the clusters. John et al. [14] given a definition of relevant features.

Definition 1 (Relevant Feature). F_i has relevancy to the target concept C if and as long as there exists some s'_i, f_i, and c, such that, for likelihood $p(S'_i=s'_i,F_i=f_i)>0$,

$$P(C=c \mid S'_i=s'_i,F_i=f_i) \neq p(C=c \text{ I } S'_i=s'_i) \quad (1)$$

Otherwise, feature Fi is associate irrelevant feature.

Definition 2 (Markov Blanket). Given a feature F_i two F, let $M_i \subset F(F_i \varepsilon M_i), M_i$ is claimed to be a Markov blanket for F_i if and only if

$$P(F\text{-}M_j\text{-}\{F\},C|F_j, M_j) = P(F\text{-}M_j\text{-}\{F\},C|M_j) \quad (2)$$

Definition 3 (Redundant Feature). Let S be a collection of features, a feature in S is redundant if and only if as it's a Markov Blanket within S.

The symmetric uncertainty (SU) [25] comes from the mutual information by normalizing it to the entropies of feature values or feature values and target categories, and has been useful to appraise the goodness of features for classification by variety of researchers. Therefore, symmetric uncertainty is considered as the measure of correlation between either 2 features or a feature and therefore the target concept. The symmetric uncertainty is outlined as follows:

$$SU(X,Y) = 2xGain(X|Y) / H(X)+H(Y) \quad (3)$$

Where, H(X) is the entropy of a discrete random variable X. Gain(X/Y) is the information gain and H(X/Y) is the conditional entropy which measure the remaining entropy of a random variable X given that the value another random variable Y is known.

Definition 4 (T-Relevance). The relevance between the feature $F_i \in F$ and therefore the target concept C is mentioned because the T-Relevance of F_i and C, and denoted by $SU(F_i,C)$.
If $SU(F_i,C)$. is bigger than a planned threshold , we say that F_i could be a strong T-Relevance feature.

Definition 5 (F-Correlation). The correlation between any pair of features F_i and F_j $(F_i,F_j \in F \wedge i \neq j)$ is known as FCorrelation of F_i and F_j, and denoted by $SU(F_i, F_j)$.
According to the on top of definitions, feature set selection will be the method that identifies and retains the strong T-Relevance features and selects R-Features from feature clusters. The behind heuristics are that

1. Irrelevant features have no/weak correlation with target concept;
2. Redundant features are assembled in an exceedingly cluster and a representative feature will be taken out of the cluster.

3.2 Algorithm and Analysis

The proposed FAST algorithm mainly consists of 3 steps: 1) Removing irrelevant features, 2) constructing an MST and 3) Partitioning the MST and select exchanging features. In first step, we calculate the T-Relevance. In second step, we compute F-Correlation value for each pair of feature. In the third step, we remove the edges, whose weight is less than both of the T-Relevance and $SU(F_j,C)$ from the MST.

After removing all the surplus edges, a forest is obtained. Every tree $T_j \in$ Forest represents a cluster that's denoted as $V(T_j)$, that is that the vertex set of T_j moreover. As illustrated on top, the features in every cluster are redundant.

3.2.1 Time complexity Analysis

The computation of SU values for T-Relevance and F-Correlation that has linear complexity. The first part of the algorithm a linear time complexity is O(m),the second part of the program initial constructs a complete graph and complexity is $O(k^2)$,third part divide the MST and select the representative features with the complexity of $O(m^2)$.The FAST has linear complexity of O(m) and worst complexity is $O(m^2)$.

4. EMPIRICAL STUDY

4.1 Data Source

For the needs of measuring the performance and effectiveness of the projected FAST algorithm, 35 publically available data set were used. The numbers of features of the 35 data sets vary from 37 to 50 with a mean of 7,874. The dimension of the 54.3 % data sets exceed 5,000, of that 28.6 percent data sets have over 10,000 features. The 35 data sets cowl a variety of application domains such as text, image and bio microarray data classification.

4.2 Experiment Setup

The performance of the projected FAST algorithm is compared with alternative feature selection algorithms in a truthful and affordable method and figured out an experimental study as follows:

1. The projected algorithm is compared with 5 different kinds of feature selection algorithms. They are 1) FCBF [21], 2) ReliefF [26], 3) CFS [19], 4) Consist [27], and 5) FOCUS-SF [25]. FCBF and ReliefF assess features separately. For FCBF, within the experiments, set the relevancy threshold to be the SU worth of the $[m/\log m]_{th}$ ranked feature for every data set. ReliefF searches for nearest neighbors of instances of various classes and weights features according to however well they differentiate instances of different classes. CFS make use of best initial search supported the analysis of a set that contains features extremely correlate with the target concept, nevertheless unrelated with one another. FOCUS-SF may be a variation of FOCUS [26]. FOCUS has constant evaluation strategy as Consist, however it examines all subsets of features. Considering the time potency, FOUCS-SF replaces complete search in FOCUS with successive forward selection.
2. Four differing kinds of classification algorithms are 1) the probability-based Naive Bayes (NB), 2) the tree-based C4.5, 3) the instance-based lazy learning algorithm IB1, and 4) the rule-based RIPPER. Naive Bayes a probabilistic methodology for classification by multiplying the individual chances of every feature-value set. This algorithm assumes independence among the features and even then provides good classification results. Decision tree learning algorithm C4.5 is an extension of ID3 that accounts for inaccessible values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on. Instance-based learner IB1 may be a single-nearest neighbor algorithm, and it classifies entities using distance metrics taking the class of the immediate associated vectors within the training set. It's the best among the algorithms utilized in our study. Inductive rule learner Repeated Incremental Pruning to Produce Error Reduction (IRIPPER) [27] is defined as a rule-based detection method and look to enhance it iteratively by using completely different heuristic techniques.
3. In order to evaluate the performance of the feature subset selection algorithms, the following four metrics are used 1) the proportion of selected features 2) the time to get the feature subset, 3) the classification accuracy, and 4) the Win/Draw/Loss record [28].

The Win/Draw/Loss record presents 3 values on a given measure they are better, equal, and worse performance than alternative 5 feature selection algorithms.

4.3 Results and Analysis

4.3.1 Run time

The analysis shows the runtime of the six feature selection algorithms.

1. Usually the individual evaluation-based feature selection algorithms of FAST, FCBF, and ReliefF area unit much faster than the set analysis primarily based algorithms of FOCUS-SF CFS, and Consist. FAST is consistently faster than all alternative algorithms. The runtime of FAST is barely 0.1 % of that of CFS, 2.4 % of that of Consist, 2.8 % of that of FOCUS-SF, 7.8 % of that of ReliefF, and 76.5 % of that of FCBF. The Win/ Draw/Loss records show that FAST outperforms other algorithms.

2. For microarray data, FAST ranks a pair of. Its runtime is only 0.12 % of that of CFS, 15.30 % of that of Consist, 18.21 % of that of ReliefF, 27.25 % of that of FOCUS-SF, and 1.25 % of that of FCBF.

3. For image data, FAST obtains the rank of one. Its runtime is only 0.02 % of that of CFS, 18.50 % of that of ReliefF, 25.27 % of that of Consist, 37.16 % of that of FCBF, and 54.42 % of that of FOCUS-SF. This reveals that FAST is a lot of economical than others once selecting features for image data.

4. For text data, FAST ranks one. Its runtime is one.83 percent of that of Consist, 2.13 % of that of FOCUS-SF, 5.09 % of that of CFS, 6.50 % of that of ReliefF, and 79.34 % of that of FCBF, respectively. This means that FAST is a most efficient than others.

4.3.2 Classification Accuracy

Table 1 shows the classification accuracies of every classifier with the six feature selection algorithms are totally different. It is observed that

1. For image data, CFS acquires rank of one, and FAST ranks 3.
2. For microarray data, FAST acquires ranks one.
3. For text knowledge, CFS acquires the rank of one, and FAST and FCBF square measure alternatives.
4. For all data, FAST get first rank

From the analysis higher than we are able to understand that FAST performs good on the microarray data.

The planned FAST effectively filters out a mass of irrelevant features within the opening move.

In the second step, FAST removes an oversized variety of redundant features by selecting a single representative feature from every cluster of redundant features.

TABLE 4.1 Rank of the six feature selection algorithms under different types

All Data(image, Microarray, text)						
	Fast	**Fcbf**	**Cfs**	**Reliff**	**Consist**	**Focs-Sf**
Nb	1	3	2	4	6	5
C4.5	1	2	4	6	5	3
Ib1	2	3	1	6	4	5
Ripper	1	4	5	6	2	3
Sum	5	12	12	22	17	16
Rank	1	2	2	6	5	4

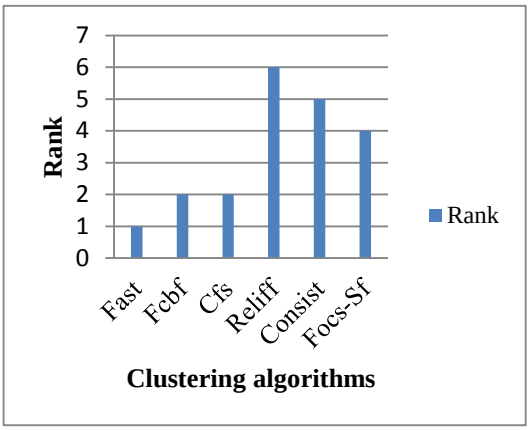

Fig 2: Ranking differences between FAST and the comparing algorithms.

5. CONCLUSION

In this paper, completely unique clustering-based feature set selection algorithm for high dimensional data was presented. The algorithm involves 1) removing irrelevant features from the dataset, 2) constructing a minimum spanning tree from relative features, and 3) partitioning the standard time and choosing representative features. Within the projected algorithm, a cluster consists of features. Every cluster is treated as one feature and therefore spatiality is greatly reduced.

The performance of the projected algorithm is compared with 5 known feature selection algorithms FCBF, ReliefF, CFS, Consist, and FOCUS-SF on the 35 publically available image, text, and microarray data from the four totally different aspects of the runtime, classification accuracy of a given classifier, and also the Win/Draw/Loss record.

Additionally it was found that FAST obtains the rank of one for microarray data, the rank of two for text data, and also the rank of three for image data in terms of classification accuracy of the 4 different types of classifiers, and CFS may be a smart various. At an equivalent time, FCBF may be a smart various for image and text data. Moreover, Consist, and FOCUS-SF area unit alternatives for text data. The prospect is to analyze different kinds of correlation measures, and analyze some formal properties of feature space. This can be achieved by using Dominant Correlation Filter (DCF).

6. REFERENCES

[1] Liu, H., Motoda, H. and Yu, L. 2004 "Selective Sampling Approach to Active Feature Selection," Artificial Intelligence, vol. 159, nos. 1/2, pp. 49-74.

[2] Guyon, I. and Elisseeff, A. 2003 "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol 3, pp. 1157- 1182, 2003.

[3] Mitchell, T.M. 1982 "Generalization as Search," Artificial Intelligence, vol. 18, no. 2, pp. 203-226, 1982.

[4] Dash, M. and Liu, H. 1997 "Feature Selection for Classification," Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156.

[5] Souza, J. 2004 "Feature Selection with a General Hybrid Algorithm, "PhD dissertation, Univ. of Ottawa.

[6] Langley, P. 1994 "Selection of Relevant Features in Machine Learning," Proc. AAAI Fall Symp. Relevance, pp. 1-5.

[7] Ng, A.Y. 1998 “On Feature Selection: Learning with Exponentially Many Irrelevant Features as Training Examples,” Proc. 15th Int’l Conf. Machine Learning, pp. 404-412.

[8] Das, S. 2001 “Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection,” Proc. 18th Int’l Conf. Machine Learning, pp. 74-81.

[9] Xing, E., Jordan, M., and Karp, R. 2001 “Feature Selection for High- Dimensional Genomic Microarray Data,” Proc. 18th Int’l Conf.Machin e Learning, pp. 601-608.

[10] Pereira, F., Tishby, N., and Lee, L., 1993 “Distributional Clustering of English Words,” Proc. 31st Ann. Meeting on Assoc. for Computational Linguistics, pp. 183-190.

[11] Baker, L.D. and McCallum, A.K. 1998 “Distributional Clustering of Words for Text Classification,” Proc. 21st Ann. Int’l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103,.

[12] I.S. Dhillon, and R. Kumar., “A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification,” J. Machine Learning Research, vol. 3, pp. 1265-1287,2003.

[13] Jaromczyk, J.W.. and Toussaint, G.T., “Relative Neighborhood Graphs and Their Relatives,” Proc. IEEE, vol. 80, no. 9, (Sept. 1992), pp. 1502- 1517.

[14] John, G.H.., Kohavi, R., and Pfleger, K., 1994. “Irrelevant Features and the Subset Selection Problem,” Proc. 11th Int’l Conf. Machine Learning, pp. 121-129.

[15] Forman, G., 2003. “An Extensive Empirical Study of Feature Selection Metrics for Text Classification,” J. Machine Learning Research, vol. 3, pp. 1289-1305.

[16] Hall, M.A., 2000. “Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning,” Proc.17th Int’l Conf, Machine Learning pp. 359-366.

[17] Kononenko, I., 1994. “Estimating Attributes: Analysis and Extensions of RELIEF,” Proc. European Conf. Machine Learning, pp. 171-182.

[18] Battiti, R., 1994. “Using Mutual Information for Selecting Features in Supervised Neural Net Learning,” IEEE Trans. Neural Networks, vol. 5, no. 4 , (July 1994), pp. 537-550.

[19] Hall, M.A. 1999. “Correlation-Based Feature Subset Selection for Machine Learning,” PhD dissertation, Univ. of Waikato.

[20] Yu, L. and Liu, H. 2003. “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution,” Proc. 20th Int’l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863.

[21] Yu, L. and Liu, H. 2004. “Efficient Feature Selection via Analysis of Relevance and Redundancy,” J. Machine Learning Research, vol. 10, no. 5, pp. 1205-1224.

[22] Fleuret, F., 2004. “Fast Binary Feature Selection with Conditional Mutual Information,” J. Machine Learning Research, vol. 5, pp. 1531-1555, 2004.

[23] Kohav, R. and John, G.H. 1997., “Wrappers for Feature Subset Selection,” Artificial Intelligence, vol. 97, nos. 1/2, pp. 273-324.

[24] Press, W.H. ,Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T., 1988. Numerical Recipes in C. Cambridge Univ. Press.

[25] Almuallim, H. and Dietterich, T.G., 1994. “Learning Boolean Concepts in the Presence of Many Irrelevant Features,” Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305.

[26] Robnik-Sikonja, M. and Kononenko, I., 2003. “Theoretical and Empirical Analysis of Relief and ReliefF,” Machine Learning, vol. 53, pp. 23- 69.

[27] Dash, M. Liu., H. and Motoda, H., 2000. “Consistency Based Feature Selection,” Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109.

[28] Cohen, W., 1995. “Fast Effective Rule Induction,” Proc. 12th Int’l Conf. Machine Learning (ICML ’95), pp. 115-123,1995.