

An Improved Feature Subset Selection Method Using Clusters

S. Kowsalya¹, A. Nagajothi²

Author: PG Scholar, Dept. of CSE, Karpagam University, Coimbatore, India

Co-Author: Asst. Professor, Dept. of CSE, Karpagam University, Coimbatore, India

Abstract: Feature extraction is the process of eliminating the irrelevant information and redundant features during Data Mining. Of all the existing feature subset selection algorithms, most of them can effectively eliminate irrelevant features but fail to handle redundant features. The Improved FAST eliminates irrelevant features first and from the result set it removes the redundant features. The Improved FAST method accomplishes four tasks. During the first step the irrelevant features are removed. In the second task features are divided into cluster that posses features with redundant features. The third task accomplishes the selection of most representative feature that is closely related to the target classes. It is selected from each cluster to form the features subset which is the fourth task. The efficiency of the algorithm is improved by applying the search method and their relevant search time is minimized. The Improved FAST algorithm is evaluated using various types of data like text data, micro-array data and image data to represent its performance.

Keywords: Data Mining, Feature subset selection, Filter method, featured clustering, Data search

I. Introduction

As a consequence of technology step up there is breakneck accumulation of data in large volumes with high dimensions. Data mining is a process that converts these data clusters into nuggets. It is a method used to discover new patterns and knowledge from an extended hefty volume of data. Feature selection is a process that selects a subset among the original large volume of features. Always an optimal feature subset is to be evaluated based on some criteria. Due to the technology enhancements the dimensionality of the domain inflates. The number of features to describe those domains also increases on hand. It becomes obstinate to find an optimal feature subset. The main objectives of feature selection are to avoid inappropriateness and improve the performance of decision making models to provide faster and more cost-effective solutions [1]. The selection of optimal features adds an increased burden in the modeling.

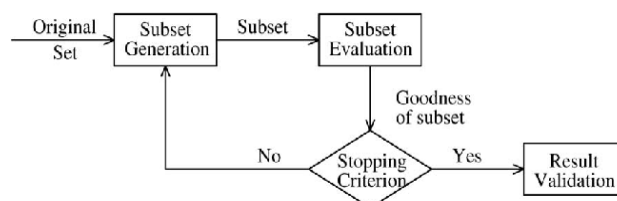


Figure 1: Feature Selection Process

The Feature Selection Process [2] shown in Figure 1 involves four steps to select the relevant subset data. This subset is to be derived from the given sample training data that is of high dimensionality. The feature selection process is evaluated using vast extent of filter selection methods available. First step called Subset generation process is done in an ad hoc manner using heuristic search techniques. Each state in the search space specifies a candidate subset for evaluation. This process assumes two set of predictions to be done in advance. Initially the user must decide the initial point which in turn influences the search direction. Therefore, different strategies have been explored for this process. They are complete, sequential and random search techniques. The second step is called Subset Evaluation. Each newly generated subset needs to be evaluated by an evaluation criterion. An evaluation criterion can be broadly categorized into two groups based on their dependency [3] on mining algorithms that will finally be applied on the selected feature subset. The evaluation criteria may be either Independent or Dependent Criteria. The third step Stopping Criteria determines when the feature selection process should stop. There are many frequently used stopping criteria. The user may follow the appropriate set of criteria based on the application developed. The final step in subset selection is the Result Validation. A straightforward way for result validation is to directly measure the result using prior knowledge about the data. If the relevant features are predicted in advance using historic data, it can be made use of in validating a resultant subset selected by any of the models.

Attribute selection process is divided into four categories based on the approach used by different methods. They are Embedded, Wrapper, Filter and Hybrid methods [4]. In Embedded methods the process of feature selection is embedded inside the training process itself. Traditional learning algorithms like Decision trees and Artificial Neural Networks use this type of approach. The second method named Wrapper defines a possible feature subset in the target space. The results produced by mining approach are used to obtain the relevance of attributes. The intrinsic properties of training data are only taken into consideration in Filter method. A score providing the feature relevance will be calculated. The attributes with low score of relevance will be removed. The classification algorithm gets the relevant feature subset list and reduces the dimensionality of the databases by removing the redundant features. A mixture of Filter and Wrapper method forms the Hybrid method. The Hybrid method uses both the features of Filter method and the Wrapper method to select the relevant attributes using the predictive accuracy. The Filter method is considered to be the efficient one since a multidimensional dataset is reduced to a simple, fast and independent attributes list. Though Data mining process provides an immense key on useful knowledge extraction our survey focuses on the Filter method of attribute selection.

II. Related Work

In 1992 Kenji Kira and Larry A. Rendell proposed a Sequential and Distance based algorithms called RELIEF [5]. This algorithm selects relevant features using a feature weight selection method. It is an accurate algorithm even if the sample data is accumulated with noise. The RELIEF method is supplied with a training data set S , sample size m and a threshold value. The training data set S is subdivided into positive and negative instances. Each time a random positive and negative instance is picked up and its Near Hit or Near Miss instance is calculated using Euclid distance. An average weight for each instance is calculated and it is compared with the given threshold. If the weighted instance is greater than threshold then it is assumed to have higher relevance. Since RELIEF follows a statistical method it can be used for any number of sample spaces.

To solve the two class problem of RELIEF, in 1994 Igor Kononenko proposed another new method called RELIEF-F [6]. This algorithm is just an extended form of RELIEF that contributes to solve problems with multi-class data. It also can handle samples that hold noise and incomplete data. The RELIEF favors the selection of attributes from one Near miss from different classes. But this RELIEF-F contributes to the selection of one Near miss from each category of classes and averages these to calculate the Weight estimation.

Another enhanced algorithm was proposed by Manoranjan Dash, Huan Liu and Hiroshi Motoda [7] in the year 1999 who worked on the inconsistency measure of the features selected. A feature subset is assumed to be inconsistent if there is occurrence of two instances with same values but with different class labels. In his work the inconsistency measure is applied to different search strategies like exhaustive search, complete search, heuristic search, probabilistic search and hybrid search comprised of complete and probabilistic search combination.

Along the path of enhancements in the Data mining approaches an enactive problem in Machine learning was identified. To solve this problem in 2000, Mark A. Hall proposed a method called Correlation-based Feature Subset Selection (CFS) [8]. His new algorithm was based on Sequential and Dependency method for Machine Learning. The sequential dependency based algorithms selects the subset features in a chronological order and the relevance of the feature is calculated using the correlation measures between the selected features. This algorithm pairs the methods of correlation measure and a heuristic strategy. CFS algorithm reduced the complication involved in selecting the feature subset that paved the way for increase in classification accuracy. This method may be ineffective in some cases of small areas in instance space. The vulnerability problem of heuristic approach is overcome by a probabilistic approach proposed by Huan Liu and Rudy Setiono [9] during the year 2000. It is the Las Vegas approach (LVF) for filtering attributes. They used the Random feature selection method and a consistency criterion threshold is defined. A random subset S is generated from N features in every try of selection process. Maximum number of tries is carried out to select that random subset. If an inconsistency of feature with the data set is less than the minimum threshold then it is assumed to be the best number of features. The final subset obtained is the best dimensionally reduced attribute set.

Another improvement in RELIEF was done by Huan Liu, Hiroshi Motoda and Lei Yu. In 2002, they proposed a Feature Selection algorithm called RELIEF-S [10] with selective sampling concept. In their work it has been brought to the idea that uniform distribution of instances lacks in some cases and the selected features gain more representation than others that are not selected. In their work a sample binary KD tree is chosen where k number of features is taken for the fast nearest neighbour search. In this tree for a given vertex V the left edges represent an associated feature with values less than V and the right edge represents a feature greater than V . Each KD tree constructed divides the sample space into m number of classes out of which representative features can be selected. During 2003 further enhancements on the RELIEF algorithm have been diverted by a new concept proposed by Lei Yu and Huan Li. It was a Sequential and Information based algorithm called Fast Correlation Based Feature Selection method (FCBF) [11]. FCBF was activated as a prevalent filter based algorithm that focuses on correlation analysis techniques to extract subset of features. It is not required to perform pair-wise correlation analysis in FCBF. This algorithm performs the two most significant process of

feature selection that is removal of irrelevant and redundant features using Symmetrical Uncertainty (SU) as the goodness measure. This algorithm takes in random a feature of a class and calculates its goodness measure. The goodness measure is assumed to be the Symmetric Uncertainty (SU) values. If the SU is greater than a minimum threshold value then it is appended to the list of selected features. After the formation of these selected features, each feature is compared with the subsequent attribute to calculate the correlation between them. If any attribute is found to have less correlation then it is removed from the selected list. The resultant list forms the minimal feature subset from the given high dimensional data set. FCBF increases the accuracy and achieves the highest level of performance in diminishing the dimensionality. Another binary feature selection process is proposed by Francois Fleuret [12] in 2004 that acts as a new method of feature selection using conditional mutual information. He proposed the conditional entropy $H(U/V)$ based intuitive tool to pick features. If the two variables U and V are independent, no information can be gained from each other. So the value of conditional entropy $H(U/V)$ is equal to the entropy itself. If they are dependent and deterministic then conditional entropy is zero as no new information is needed from U if V is known. This approach is applied for two types of datasets one with image data to find edges of face and the other with active molecule of drug design dataset. The pre-processing step of feature selection may lead to confusions thus moving way for false predictions and decision making. So a good feature selection method must be followed. Yuxuan SUN, Xiaojun LOU and Bisai BAO in 2011 proposed another Relief feature selection method based on Mean-Variance model [14]. This model obtains feature weight estimation based on the mean and variance. The most relevant feature $W[F]$ is obtained that is a reasonable weight estimation W of feature F leads to minimal variance value. Using Lagrange Objective function a final weight measure problem is solved. This makes the result more stable and accurate. If the sample data obtained from training set is random then the frequency of instance sampling is uncertain [15]. It leads to the fluctuation of the instances set.

Qinbao Song, Jingjie Ni and Guangtao Wang proposed a new clustering tree based feature subset selection method for high dimensional data. This algorithm FAST [16] was proposed in the year 2013. It evaluates to its result in two steps. In the first step, the relevance information of the features is calculated using Entropy method which is used to construct a tree. The tree is divided into clusters [12] by using graph-theoretic clustering methods. In the second step, from the forest of set of clusters in the form of trees the most relevant feature is selected. This step is done by calculating the Symmetric Uncertainty (SU) value. This SU value represents the correlation between two features or between a feature and a concept. Using this, the irrelevant values are removed. In the next step the redundant features are eliminated. From the first step, the resultant data set without any irrelevant feature is acquired. This is taken as input for the next step. A minimum spanning tree structure is constructed that passes through all the possible edges. From this tree the partition is done to form a forest. Each forest represents the nodes with similar features. All the nodes of the representative forest seem to be similar and they have analogous characteristics. So each representative attribute is selected from every forest to form a resultant subset.

The remainder of this paper is organized as follows. Section III describes the proposed method for Feature subset selection. Section IV discusses the improved FAST method. Experiments are demonstrated in Section V, and this work is concluded in Section VI.

III. Feature Selection

Of all the existing feature subset selection algorithms, most of them can effectively eliminate irrelevant features but fail to handle redundant features. There are also algorithms that can eliminate the irrelevant features also taking care of the redundant features. Our proposed work falls into the second group. The Improved FAST algorithm is proposed to provide an efficient method to feature subset selection for different categories of data. The architecture of proposed system (Figure 2) shows the Functional diagram of our Improved FAST subset selection method.

This architecture takes a search dataset as input and performs the subset selection process. The user must login with their authenticated user id and password before going in to the search process. Four types of search operations are performed here that is text search, image search, news search and sports search. Before the search operation is performed our Improved FAST performs the feature extraction process and eliminates the irrelevant and redundant information from our search space. This reduces the time complexity and increases the performance of the search. The Improved FAST eliminates irrelevant features first and from the result set it removes the redundant features. The Improved FAST method accomplishes two tasks. During the first step the features are divided into clusters using graphical cluster methods removing the irrelevant features. In the second step, the most representative feature that is closely related to the target classes is selected from each cluster to form the features subset. The efficiency of the algorithm is improved using Minimum spanning tree clustering method. The Improved FAST algorithm is evaluated using various types of data like text data, micro-array data and image data to represent its performance.

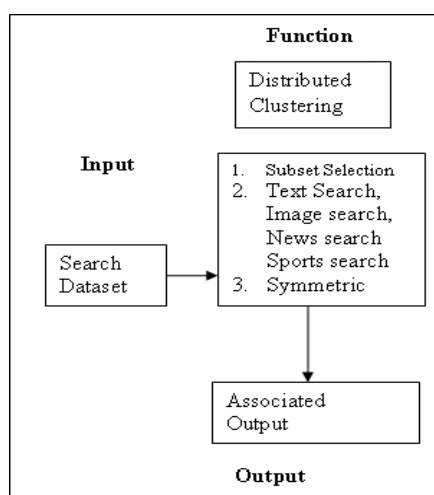


Figure 2: Functional diagram of improved FAST

The proposed method undergoes the Feature selection process by undergoing four phases.

A. Irrelevant Feature Removal

This phase is concerned with the removal of irrelevant features that does not match with the target concept. The features are extracted as irrelevant using a match concept that reveals the relevance property between a feature and its target class. If there is no match between the values of the selected feature f and the target class c , it is said to be irrelevant and thus removed from the set of features. If the relevance measure is beyond the threshold then that feature is selected.

B. Clustering

The Distributional clustering has been used to cluster words into groups based either on their participation in particular grammatical relations with other words or on the distribution of class labels associated with each word. As distributional clustering of words is agglomerative in nature, and result is suboptimal word clusters and acquires high computational cost. A new information-theoretic divisive algorithm for word clustering is proposed and applied it to text classification. The proposed algorithm is used to cluster features using a special metric of distance, and then makes use of the resulting cluster hierarchy to choose the most relevant attributes.

C. Redundant Feature Removal

The next phase in FAST method is redundant feature removal. After removing the irrelevant features, there occurs need to remove the redundant features. If a feature is embedded with redundant information, then it may not contribute to the better prediction of target classes. Redundant features completely correlate with each other. So if F is a set of features then it is said to be redundant if it has Markov Blanket within F . Assuming this as the redundant feature is removed. The major amount of work for FAST algorithm involves the computation of Symmetric Uncertainty (SU) values from which the T-Relevance and F-Correlation are calculated. This measure has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity in terms of the number of features m . The Improved FAST algorithm strives to improve the time complexity by reducing the time taken to calculate the SU values thus increasing the overall performance.

D. Subset Selection

Relevant features are grouped into clusters and a representative of each cluster is retrieved to get a required feature without redundancy. The Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, good feature subsets selection methods must be used to obtain features that are highly correlated with the class, yet uncorrelated with each other. A novel method is proposed which can efficiently and effectively deal with both irrelevant and redundant features, and obtains a good feature subset.

IV. Algorithm Analysis

The improved FAST algorithm contains four phases. (i) Irrelevant feature removal (ii) Distributed Clustering (iii) Redundant Feature removal (iv) Subset selection. The modified FAST algorithm is shown that represents the reduction in time to match and retrieve the data search process. This algorithm when used in various classifiers enhances their prediction capability and increases search performance.

Algorithm: Improved FAST

Inputs

$D (F_1, F_2, \dots, F_m)$ - Data set
 C - Target class
 TH - Threshold value

Output

S -selected feature subset.

Start

//Phase 1: Irrelevant Feature Removal

for $i = 1$ to m **do**

$T\text{-Relevance} = SU (F_i, C)$

if $T\text{-Relevance} > TH$ **then**

$S = SU\{F_i\}$

//Phase 2: Clustering

Construct a Graph G such that

for each feature in D calculate

$F\text{-Correlation} = SU (F_i, F_{i+1})$

Construct a tree with clusters

//Phase 3: Redundant feature removal

With each tree construct a forest F

Select a representative feature F_i^R from each cluster

//Phase 4: Subset selection

With edges of Forest select the subset S such that

$S = S \cup \{ F_i^R \}$, where F_i^R is the relevant feature subset selected.

End

V. Experimental Results

The performance of the proposed feature selection method has been evaluated text and image data. This study has analyzed the time complexity of the algorithms FAST and improved FAST. The result of the study is illustrated with a chart. The **table-1** shows the runtime of text data search using FAST and Improved FAST method.

Data \ Algorithm	Runtime (ms)
FAST	6989
Improved FAST	1812

Table-1: Runtime of Fast and Improved Fast - Text Data

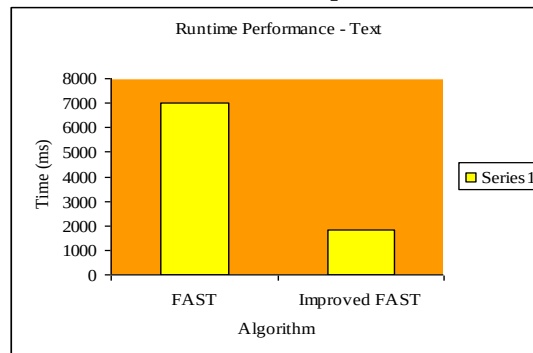


Figure 3: Performance difference between FAST and improved FAST for Text data

This chart reveals the fact that the modified FAST algorithm shows higher performance than the existing FAST algorithm. The **table-2** shows the runtime of image data search.

Data \ Algorithm	Runtime (ms)
FAST	1520
Improved FAST	1812

Table-2. Runtime of Fast and Improved Fast - Image Data

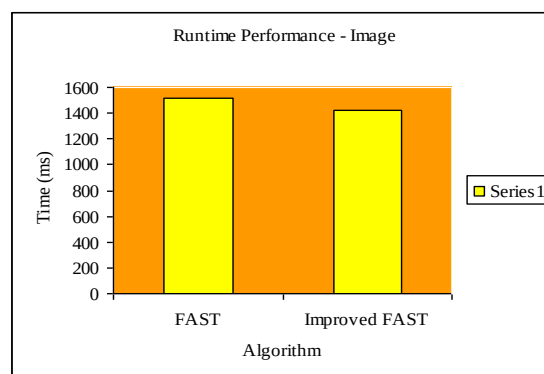


Figure 4: Performance difference between FAST and improved FAST for Image data

This chart reveals the fact that both FAST and Improved FAST provides nearly similar results for image data.

VI. Conclusion and Future Work

The Improved FAST algorithm is proposed to provide an efficient method for feature subset selection for different categories of data. This work is done in four phases to remove irrelevant feature, clustering similar features, removing redundant features and subset selection. These phases are applied to Data mining algorithms to reduce the number of features selected for mining. Performance based on time complexity is measured and results are shown. Thus our Improved FAST algorithm works efficiently and shows higher performance than FAST in terms of search time. In the near future this work may also be enhanced in search of video and audio files that are not performed in this work. The same set of procedures may be applied for video and audio files with large set of features.

References

- [1] Zhao Z. and Liu H. (2009), 'Searching for Interacting Features in Subset Selection', Journal of Intelligent Data Analysis, 13(2), pp 207-228, 2009.
- [2] Huan Liu and Lei Yu, 'Towards Integrating Feature Selection Algorithms for Classification and Clustering', IEEE Transactions on Knowledge and Data Engineering Vol. 17 No. 4, 2005.
- [3] Sha. C, Qiu X. and Zhou A (2008), 'Feature Selection Based on a New Dependency Measure', Fifth International Conference on Fuzzy
- [4] Sunita Beniwal and Jitendar Arora (2012), 'Classification and Feature Selection Techniques in Data Mining', International Journal of Engineering Research and Technology, Volume 1 Issue 6, August 2012.
- [5] K. Kira and L.A. Rendell (1992), 'The Feature Selection Problem: Traditional Methods and a New Algorithm', Proceedings of 10th National Conference on Artificial Intelligence, pp. 129-134, 1992.
- [6] I. Kononenko (1994), 'Estimating Attributes: Analysis and Extension of RELIEF', Proceedings of Sixth European Conference on Machine Learning, pp. 171-182, 1994.
- [7] Huan Liu, Manoranjan Dash and Hiroshi Motoda, 'Feature Selection Using Consistency Measure', Second International Conference, DS'99 Tokyo, Japan, December 6-8, 1999 Proceedings, pp 319-320
- [8] M.A. Hall (2000), 'Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning', Proceedings of 17th International Conference on Machine Learning, pp. 359-366, 2000.
- [9] Huan Liu and Rudy Setiono, 'A Probabilistic approach to Feature Selection – A Filter approach', 2000.
- [10] Huan Liu, Hiroshi Motoda (2002), and Lei Yu, 'Feature Selection with Selective Sampling', Proceedings of 19th International Conference on Machine Learning, pp. 395-402, 2002.
- [11] Huan Liu and Lei Yu (2003), 'Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution', Proceedings of 20th International Conference on Machine Learning, pp. 856-863, 2003.
- [12] Fleuret. F (2004), 'Fast Binary Feature Selection with Conditional Mutual Information', Journal of Machine Learning Research 5 (2004) 1531-1555.
- [13] Huan Liu and Lei Yu (2005), 'Toward Integrating Feature Selection Algorithms for Classification and Clustering', IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 4, April 2005
- [14] Yuxuan SUN, Xiaojun LOU and Bisai BAO., 'A Novel Relief Feature Selection Algorithm based on Mean-Variance Model', Journal of Information and Computational Science, pp 3921-3929, 2011
- [15] Park H. and Kwon H., 'Extended Relief Algorithms in Instance-Based Feature Filtering', In Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology, pp 123-128, 2007.
- [16] Qinqin Song, Jingjie Ni and Guangtao Wang (2013), 'A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data', IEEE Transactions on Knowledge and Data Engineering Vol: 25 No: 1 Year 2013.