# Feature Selection Algorithm Using Fast Clustering and Correlation Measure

**Swapnil Sutar**

*Research scholar, Dept. of Computer Engineering., JSPM's ICOER, Pune, Maharashtra, India*

---

**Abstract -** ***Nowadays most of the application uses tens of thousands of data for their processes. It's very difficult task to make partition of objects having similar features or properties into a group. These groups are classified and features are selected among them. This feature selection should be done such a way that it gives effective and accurate result. Feature selection has been an active research area in pattern recognition, statistics and data mining community. Idea behind feature selection is to choose a subset of input variables by eliminating features with little or no predictive information The Fast Clustering selection method produces highly effective and efficient results. In this paper, The FAST clustering method is explained including correlation techniques for removing irrelevancy among datasets. The FAST works in two steps, making clusters of datasets using Minimum Spanning Tree and then obtain a effective feature among all forest cluster. Various correlations techniques can be added for increasing effectiveness of the feature selection. All clusters resulted in this selection algorithm are relatively independent of each other.***

***Key Words: Feature Selection, FAST, Clustering, Correlation, Minimum Spanning Tree.***

## 1. INTRODUCTION

Nowadays most of the application uses tens of thousands of data for their processes. It's very difficult task to make partition of objects having similar features or properties into a group. These groups are classified and features are selected among them. This feature selection should be done such a way that it gives effective and accurate result. DATA mining is the one of the essential process of extraction of previously unknown and nontrivial information from large databases, which is very useful for finding the useful patterns, plays an important role in different data mining tasks. In data mining, partitioning and clustering of the datasets is common and important task. Clustering is the process consist of grouping a set of objects into classes of similar objects. The performance, usefulness, and robustness of clustering algorithms are depends on Finding similarities between data according to the properties found in the data and grouping similar data objects into clusters. The quality of a clustering result mainly based on both the similarity measure used by the method along with its implementation. So for increasing quality and accuracy FAST clustering subset selection algorithm is used.

In data mining, these patterns play an important role. These patterns give the knowledge of particular search. The same key approach is used in feature selection method.

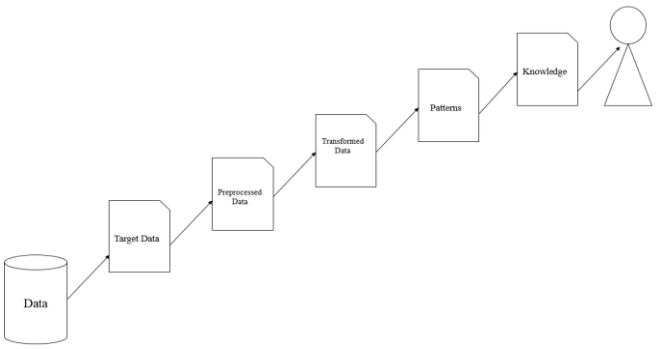


Fig1 data mining Patterns.

Feature selection method is general form of feature extractions. In feature extraction technique, new feature set is generated from already preset data features. Unlike in feature selection technique, it gives the useful subsets of required search feature. Various algorithms such as best search, genetic algorithm, greedy forward selection algorithm, greedy backward elimination algorithm are used for the feature selection process.

Many feature subset selection algorithms have been proposed including above mentioned algorithms for machine learning applications. They are Embedded, Filter, Wrapper, and Hybrid approaches. The wrapper method used to determine the goodness of the selected subsets, the accuracy of this algorithm is ordinarily high. But the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the algorithms is not guaranteed [1], [2], [12].

The wrapper methods are expensive to compute and hence they are not useful on small training sets [1], [2]. When the number of data sets is very large i.e. high

dimensional data, the filter methods are usually a good choice

FAST algorithm is an effective way to reduce dimensions, removing irrelevant data and to produce result in high effective manner.

To achieve goal of FAST clustering algorithm, it works in two steps. In the first step, it divides data into clusters using graph methods. Minimum Spanning Tree technique is used for that purpose. In second step of FAST, the subsets that are more accurate or relative with the targeted search are selected from cluster and then it forms a feature subset. Feature subset identifies and removes as many irrelevant and redundant features as possible. As we are clustering all the features with their relations, all the related feature set and unrelated feature set are clustered separately as per clustering property. This clustering helps to predict the relevant and irrelevant feature. Although some of existing system or algorithm has capability to remove the unrelated feature set, some do not involves efficiency and effectiveness to achieve goal [1].

The Minimum Spanning Tree generates a neighborhood graph of instances, then delete any edge in the graph that is much longer or shorter than its neighbors. The result is a forest and each tree in the forest represents a cluster. This cluster is used to select feature subset.

## 2. RELATED WORK

The feature selection algorithm concerns with removing of unrelated feature set as well as eliminating redundant feature sets. Many machine learning algorithms used for feature selections can be viewed in two ways. Many of them are responsible only for removing of irrelevant feature sets. i.e. they handles only irrelevancy of the feature sets excluding handling of redundant data sets.

On the basis of this functionality, all machine learning algorithms can be grouped in two classes one which handles only unrelated data and other class handling both unrelated datasets as well as redundant data.

Feature selection algorithms which exist are Wrapper, Relief, filter, embedded methods. All these machine learning algorithm are used for feature selection. The Embedded method includes the traditional algorithms like decision support trees, artificial neural networks for the feature selection and learning approach. The Wrapper method works on the basis classification algorithm. The main objective of this function is pattern classifier. These classifiers evaluate the feature sets by their predictive accuracy. They typically uses cross-validation for finding goodness of the feature sets. As it is tied to bias of the classifier, it lacks in generality. Also the computational complexity is much higher.

The filter method is independent of classification algorithms. They evaluate the features by information contents, statistical dependencies or inter class distance among the datasets [1],[14]. Mostly the wrapper and filter methods are combinely used for feature selection. The filter method has fast execution. By considering the generality of the algorithm, they are suitable for large data sets.

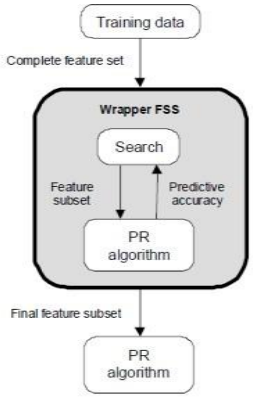


Fig2. Wrapper Method[14].

The Relief method is considered in first group mentioned above. It uses distance based criteria for feature selection. In this distance-based criteria, it gives weights to each set according to their predictive accuracy. But it always gives high weights to both sets which are redundant or highly correlated with each other. Hence this method only removes irrelevant datasets and cannot handle redundant datasets. Relief F is extension of Relief algorithm with ability to solve multiclass problem. But still it does not remove redundant datasets.

As the redundant sets are also affects the efficiency of the result obtained, it is necessary to consider the redundancy of data by the algorithms. CFS, Fast Correlation Based Feature Selection (FCBF) [1],[10],[12] are the algorithms considering this redundancy of data while feature selection.

The CFS works on the basis that good feature set contains highly correlated with required search but uncorrelated with each other [1]. The FCBF is fast filter method which includes both removing of irrelevant data and elimination of redundant datasets.

The FAST algorithm considers both issues of irrelevancy and redundancy of data.

## 3. FEATURE SELECTION

The algorithm mainly focuses on removal of irrelevant feature set and elimination of redundant data sets. Followings are the traditional definitions which are used in this algorithm including other correlation measures. There are many correlation measures which are supposed be performed for comparison are Pearson's correlation measure, spearman's correlation method, Kendall's Tau Coefficient, fast correlation based feature (FCBF)[1].

**Method to calculate entropy of variable X i.e. H(X):**

```
for(int i=0;i<X.length;i++){
sum+=X[i][1]*Math.log(X[i][1]);
}
```

**Method to calculate entropy of variable Y i.e. H(Y):**

```
for(int i=0;i<Y.length;i++){
sum+=Y[i][1]*Math.log(Y[i][1]);
}
```

**Method to calculate information gain:**

```
infoGain = xEntropy-xyEntropy;
```

**Method to calculate symmetrical uncertainity:**

```
su=2*(gain/(xEntropy+yEntropy));
```

Let *F* be a full set of features, *Fi* a feature, and

$Si = F - \{Fi\}$.

The relevance can be formalized as follows.

**Strong relevance:** *A feature Fi is strongly relevant iff*

$\mathbf{P}(C \mid Fi, Si) \neq \mathbf{P}(C \mid Si)$ .

**Weak relevance:** *A feature Fi is weakly relevant iff*

$\mathbf{P}(C \mid Fi, Si) = \mathbf{P}(C \mid Si)$, and

$\exists S0\ i \subset Si$, *such that* $\mathbf{P}(C \mid Fi, S0i) \neq \mathbf{P}(C \mid S0i)$ .

**Irrelevance:** *A feature Fi is irrelevant iff*

$\forall S0i \subseteq Si, \mathbf{P}(C \mid Fi, S0i) = \mathbf{P}(C \mid S0i)$ .

Strong relevance suggests that it gives accurate result with target search. Weak relevance indicates the little relationship among class and Irrelevance indicates no relation between particular datasets. For elimination of redundant data following definitions are used.

**Markov blanket:** *Given a feature Fi, let Mi ⊄ F (Fi /2Mi), Mi is said to be a Markov blanket for Fi iff*
$\mathbf{P}(F - Mi - \{Fi\}, C \mid Fi, Mi) = \mathbf{P}(F - Mi - \{Fi\}, C \mid Mi)$ .

**Redundant feature:** *Let G be the current set of features, a feature is redundant and hence should be removed from G iff it is weakly relevant and has a Markov blanket Mi within G .*

**Correlation Measure:** Non-linear correlation measures, many measures are based on the information-theoretical concept of

*entropy*, a measure of the uncertainty of a random variable.

$$r_{XY} = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{nS_X S_Y}$$

(Pearson's correlation measure.)

**F-Correlation:** The correlation between any pair of features

Fi and Fj (Fi,Fj Є FΛI ≠ j) is called the F-Correlation of Fi and Fj, and denoted by SU(Fi, Fj).

### 3.1. Algorithm:

Input: D datasets with required search, T- relevance threshold value.

Output: feature set with different clusters.

Steps:

→ for i (first node) to last node (n) do

{

Check T-relevance value,

If (Threshold value >0)

Then it is relevant to target search

}

→G=null; (construction of minimum spanning tree)

→for each pair of node or dataset

{

Apply correlation technique;

Assign relevance value as edge;

}

→MST algorithm

→forest gereration

→for each edge

{

If value of edge is greater,

Delete the edge.

}

→for each forest

Check the condition and feature set is obtained.

→end.

The Fast algorithm with the help of correlation measure is achieved for high accuracy unlike previous algorithms.

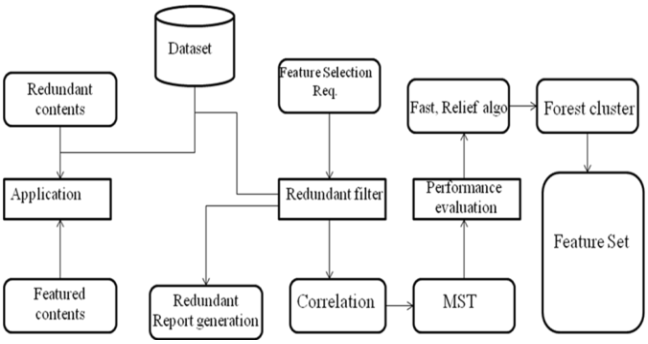


Fig 3. System Architecture.

In this proposed work, different types of correlation techniques are going to be used. The results obtained from all these correlation measure applied on datasets is then compared with each other.

The targeted datasets are taken as first input to the system. The front end application then communicates with the datasets for targeted requirements. These datasets will be then filtered for checking of redundancy. The redundant filter removes the redundant data and generates the reports of the same according to relevance value which will be calculated at this phase. Then the datasets will transferred for the relevancy check. The different correlation techniques are used for determining relevancy among datasets and featured sets. This phase will be helpful in providing accurate and efficient result for the targeted search. After this the main phase of tree generation will be takes place. The minimum spanning tree will be generated and clustered will be obtained. All these clusters will be independent of each other. Hence the required feature set can be easily and accurately obtained from the large number of datasets.

The correlation measures are going to be implemented in FAST algorithm and all results are compared with each other to find out which measure is useful with FAST algorithm for high dimensional data

## 4. MATHEMATICAL MODEL

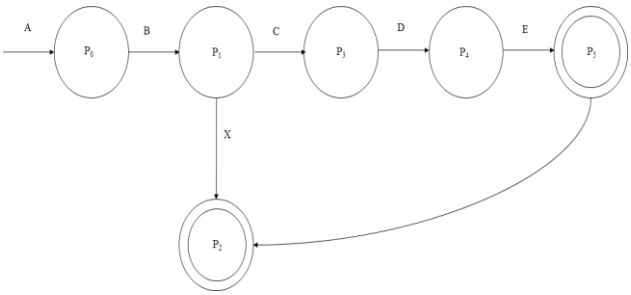


Fig 4. DFA for this system.

M= {Q, Є,□, ƒ, $q_0$}

Where,

Q= {P0, P1, P2, P3, P4, P5}

∑= {A, B, C, D, X}

□ = $P_0$:$P_1$: R, $P_1$:$P_2$: R, $P_1$:$P_3$: R, $P_3$:$P_4$: R, $P_4$:$P_5$: R, $P_5$:$P_2$: R.

$q_0$ = {$P_0$}

F= {$P_5$}

$P_0$ =Initial Phase

$P_1$=Redundant Filter

$P_2$=Redundant Report Generated

$P_3$=Correlation Phase

$P_4$=MST Construction

$P_5$=Feature Set

A= Datasets with targeted value.

B= Filtered datasets.

C=Relevant Datasets.

D=Forest Tree.

E= Feature set with redundant data report
X=Redundant Report information.

The modules initiates with taking targeted values as a input for state $P_0$.Then at state P1, redundancy is checked and redundant report will be generated. The phase or state P3 is main phase for using of different types of correlation measures, which will gives results for comparisons. After this measures, minimum spanning tree will be constructed which will generate forests of feature clusters. And finally the feature set related with target class or search will be obtained.

## 5. RESULTS

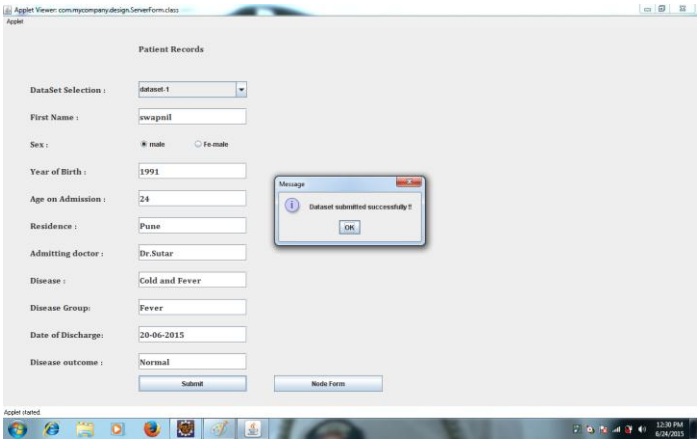


Fig.5 Data insertion

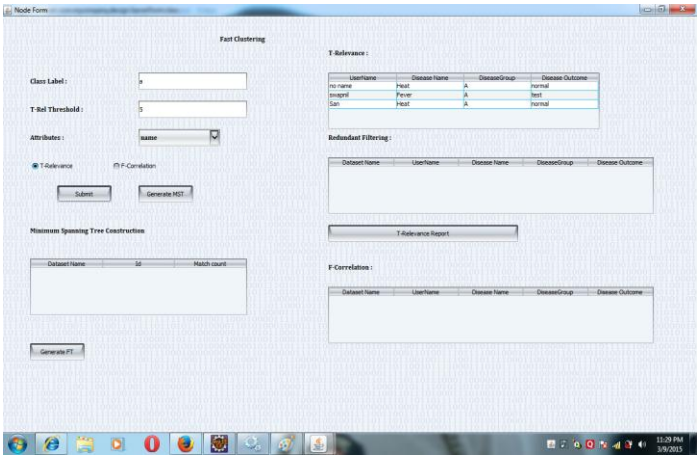


Fig.6.Redundancy Check

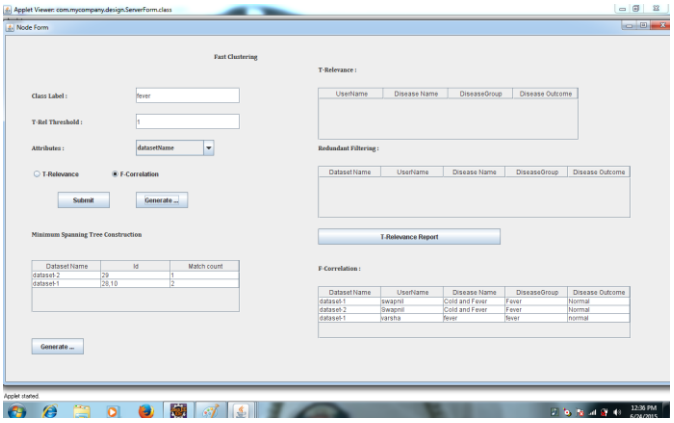


Fig.7.Correlation and MST generation

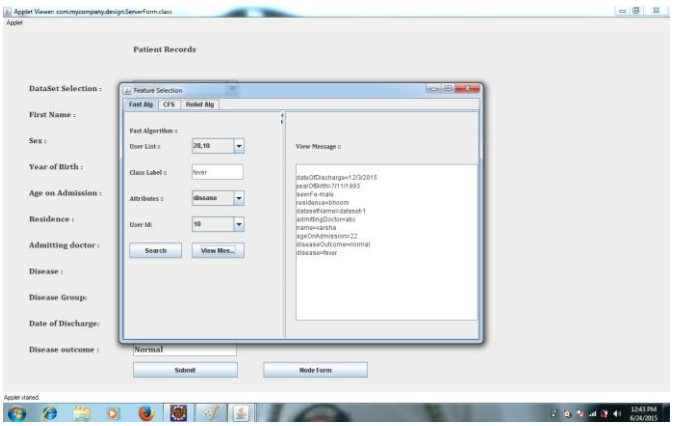


Fig.8.Feature-Output

## 6. CONCLUSION

After surveying literature survey, the existing machine learning algorithms are understood. The Wrapper, Filter, Embedded, Relief algorithms are concerns only with removing of unrelated data sets while learning the feature. The FAST algorithm is used which will consider both the cases of removing unrelated datasets as well as elimination of redundant datasets. The Pearson correlation measure will help to improve the efficiency of result as it considers negative or positive relation.

## ACKNOWLEDGEMENT

This is a great pleasure & immense satisfaction to express my deepest sense of gratitude & thanks to everyone who has directly or indirectly helped me in research paper. I express my gratitude towards project guide Prof. Devendra Gadekar , PG Coordinator Prof. Rajesh Phursule and Prof. S.R. Todmal (Head, Department of Computer Engineering, Wagholi, Pune )who guided & encouraged me the research work. I also thank all friends for being a constant source of my support.

## REFERENCES

[1] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast clustering based feature subset selection algorithm for high dimensional data", In proceedings of the IEEE Transactions n Knowledge and data engineering, 2013.

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

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

[4] [A New Clustering Based Algorithm for Feature Subset Selection (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5272-5275.

[5] Dingcheng Feng, Feng Chen_, and Wenli Xu "Efficient Leave-One-Out Strategy for Supervised Feature
Selection" TSINGHUA SCIENCE AND TECHNOLOGY ISSN 1007-0214 09/10 pp629 635 Volume 18, Number 6, December 2013.

[6] Houtao Deng, George Runger "Feature Selection via Regularized Trees" The 2012 International Joint Conference on Neural Networks (IJCNN), IEEE, 2012.

[7] Yijun Sun, Sinisa Todorovic "Local Learning Based Feature Selection for High Dimensional Data Analysis" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 32, NO. 9 SEPT.2010.

[8] Sriparna Saha "Feature selection and semi-supervised clustering using multiobjective optimization" *SpringerPlus* 2014, 10.1186/2193-1801-3-465.

[9] R.Munieswari,"A Survey on Feature Selection Using FAST Approach to Reduce High Dimensional Data" ,IJETT, Volume 8 Number 5- Feb 2014.

[10] [10] Jesna Jose,"Fast for Feature Subset Selection Over Dataset" International Journal of Science and Research (IJSR), Volume 3 Issue 6, June 2014.

[11] S. Chikhi and S. Benhammada, "ReliefMSS: A Variation on a Feature Ranking Relieff Algorithm," Int'l J. Business Intelligence and Data Mining, vol. 4, nos. 3/4, pp. 375390, 2009.

[12] L.C. Molina, L. Belanche, and A. Nebot, "Feature Selection Algorithms: A Survey and Experimental Evaluation," Proc. IEEE Int'l Conf. Data Mining, pp. 306-313, 2002.

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

[14] Jain, A.K.; Duin, P.W.; Jianchang Mao, "Statistical pattern recognition: a review", IEEE Transactions on Pattern Analysis and Machine Intelligence, Jan. 2000.

[15] Isabelle Guyon" An Introduction to Variable and Feature Selection",Journal of Machine Learning Research 3, 2003.