

FEATURE SELECTION METHOD FOR HIGH DIMENSIONAL DATA

Swati V. Jadhav¹ and Vishwakarma Pinki²

^{1,2}Computer Engineering, Shah & Anchor Kutchi Engineering College Mumbai

Abstract— Feature selection is the process of identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, a Fast clustering-based feature Selection algorithm (FAST) is proposed and experimentally evaluated. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic methods. In the second step, the most representative feature that is strongly related to clustering target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. The Minimum-Spanning Tree (MST) using Prim's algorithm can concentrate on one tree at a time. To ensure the efficiency of FAST, adopt the efficient MST using the Kruskal's Algorithm clustering method. Display the graph with respect to time showing comparison between Prim's and Kruskal's algorithm.

Keywords— Feature subset selection, filter method, feature clustering, graph based clustering, kruskal's algorithm.

I. INTRODUCTION

With the rapid growth of computational biology and e-commerce applications, high dimensional data becomes very common. The mining of high dimensional data is an urgent problem in day today life. Data mining is the extraction of hidden predictive information from large database, is a powerful new technology to help companies focus on the most important information in their data warehouses. Data mining incorporated many techniques such as machine learning, pattern recognition, database and data warehouse systems, visualization, algorithms, high performance computing, and many application domains.

Another name for data mining is the knowledge discover process, it typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation and knowledge representation.

We present a multidimensional view of data mining. The major dimensions are data, knowledge, technologies, and applications.

Data mining functionalities are:

- Characterization and Discrimination
- Mining Frequent Patterns
- Association and Correlations
- Classification and Prediction
- Cluster Analysis
- Outlier Analysis
- Evolution Analysis

1.1 High Dimensional Data

The technologies present investigators with the task of extracting meaningful statistical and biological information from high dimensional data.

A great deal of data from different domains such as medicine, finance, science is high dimensional. Many objects can be represented with high dimensional such as speech signals, images, videos, text documents, hand writing letters and numbers. We often need to analyze large amount of data and process them. For e.g. need to identify person fingerprints, certain hidden patterns and images, to trace objects from videos. To complete these tasks, we develop the systems to process data. However due to high dimension of data, the system directly processing them may be very complicated and unstable so that it is infeasible.

1.1.1 Challenges in High Dimensional

Curse of Dimensionality: It is phenomena that arise when analyzing and organizing data in high dimensional spaces that do not occur in low dimensional such as three-dimensional space in every day. Therefore, to process high dimensional data in the system dimensionality reduction becomes necessary [2].

Effect of High dimensionality on distance measures in Euclidian spaces: For any point in high dimensional space the expected gap between Euclidian distance to the closest neighbor and that to farthest point shrinks as the dimensionality grows.

Visualization: It is difficult to visualize and understand as it is high dimensional data.

1.2 Feature Selection

In machine learning and statistics feature selection also known as variable selection, attribute selection or variable subset selection. It is the process of detecting relevant features and removing irrelevant, redundant or noisy data [1].

1.2.1 Two Approaches for Feature Selection

- Individual Evaluation: The weight of an individual feature is assigned per its degree of relevance.
- Subset Evaluation: candid feature subsets are constructed using search strategy.

1.2.2 Feature Selection process

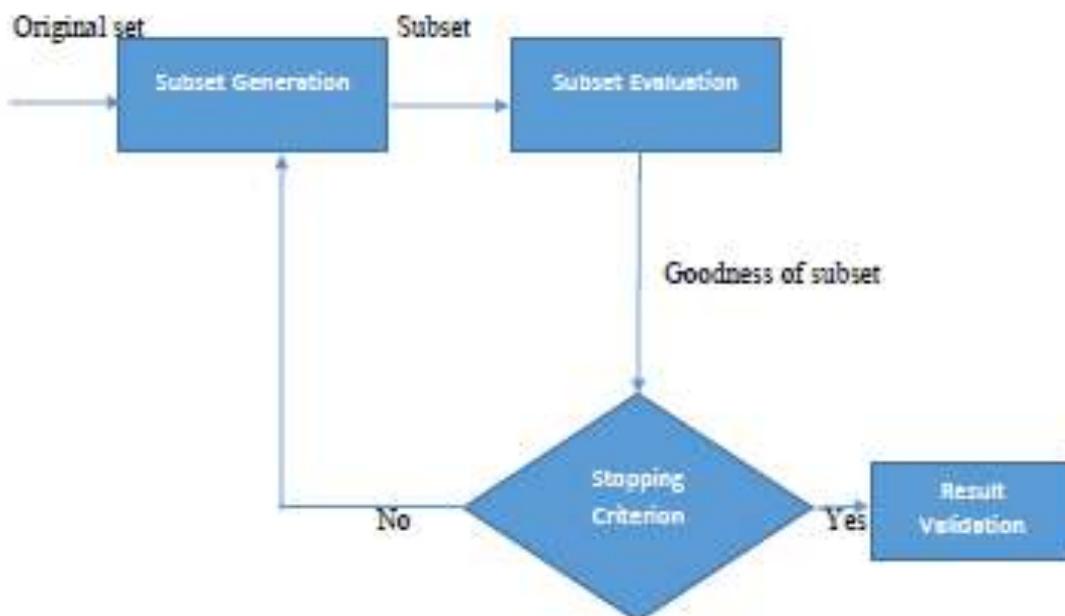


Figure.1 Four key steps for the feature selection process [9]

Subset generation is a heuristic search in which each state specifies a candid subset for evaluation in search space. In this process successor generation decides the search starting point, which influences

the search direction and search organization is responsible for feature selection process with specific strategy such as sequential search, random search.

Evaluation criteria determine the goodness of the candid subset of features. The criteria can be of two types:

- Independent Criteria: without involving any mining algorithm it evaluates the goodness of features. The criteria namely distance measures, probability of error measures, consistency measures.
- Dependent Criteria: It involve predetermine mining algorithm for feature selection to select features based on the performance of mining algorithm applied to the selected subset of features.

Stopping Criterion is used to stop the selection process. There are some general stopping criteria:

- When the search completes
- Deletion or addition of features to the subset.

Result validation: Feature selection method must be validating by carrying out different tests and comparison with previously established results or comparisons with the result of competing methods using artificial datasets, real world datasets or both.

I. RELATED WORK

With respect to the filter feature selections, the application of cluster analysis has been demonstrated in this paper. They stated the FAST clustering based algorithm is effective and efficient. The algorithm works in two steps in the first step features are divided into clusters by using graph theoretic clustering methods. In the second step the most representative feature is strongly related to target classes is selected from each cluster to form a subset of features. The framework of proposed feature composed of the two connected components of irrelevant feature removal and redundant feature elimination. FAST algorithm uses minimum spanning tree based method to cluster features. They have experimented FCBF, Relief F, CFS, Consist, Focus-SF techniques on 35 different datasets and conclude that the FAST algorithm is effective than all others [1]. A new FR algorithm termed as class dependent density based feature elimination (CDFE) for high dimensional binary data sets. CDFE uses filter wrapper approach to select a final subset. For data set having hundreds of thousands of features. Feature selection with FR algorithm is simple and computationally efficient but redundant information may not be removed. FSS algorithm analyses the data for redundancies but may become computationally impractical on high dimensional datasets. They address these problems by combining FR and FSS methods in the form of two stage feature selection algorithm. CDFE not only presents them with feature subset good in terms of classification but also relieves them from heavy computation. Two FSS algorithms are employed in second stage to test the two stage feature selection idea. Instead of using threshold value CDFE determines the final subset with the help of classifier [2]. The framework developed to perform feature selection for graph embedding in which a category of graph embedding method is cast as least squares regression problem. In contrast to filter methods, wrapper methods are application dependent. The embedded method encapsulates the feature selection into sparse regression method termed as LASSO. In this framework a binary feature selector is introduced to naturally handle the feature cardinality in the least squares formulation. The resultant integral programming problem is then relaxed into a convex quadratic ally constraint quadratic program (QCQP) learning problem which can be efficiently solved via a sequence accelerated proximal gradient (AGP) methods. The proposed framework is applied to several is embedding learning problems including supervised, unsupervised and semi supervised graph embedding. The graph embedding suffers from two weakness that is it is hard to interpret the

resultant features when using all dimensions for embedding and the original data inevitably contains noisy feature could make graph embedding unreliable and noisy [3]. To find nearest neighbor matching, the two algorithms are most efficient the randomized k-d forest and a new algorithm the priority search k-means tree. Also proposed new algorithm for matching binary features by searching multiple hierarchical clustering trees. They show that the optimal nearest neighbor algorithm and its parameter depend on the data set characteristics and describe an automated configuration procedure for finding the best algorithm to search a particular data set. They have been released as an open source library called fast library for approximate nearest neighbors (FLANN) into openCV and is now one of the most popular libraries for nearest neighbor matching [4]. They presented novel concept predominant correlation and propose a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis [5]. They presented filter-wrapper hybrid method (FWHM) to optimize the efficiency of feature selection. FWHM is divided into two phase, which orders these features according to reasonable criterion at first, then selected best features based on final criterion. These experiments on benchmark model and engineering model prove that FWHM has better performance both in accuracy and efficiency more than conventional methods [6]. A new hybrid algorithm that uses boosting and incorporates some of the features of wrapper methods into a fast filter method. For feature selection results are reported on six world datasets and hybrid method is much faster and scales well to datasets with thousands of features [7]. The definitions for irrelevance and for two degrees of relevance incorporated in this paper. The features selected should depend not only on the features and the target concept but also on the induction algorithm. A method is described for feature subset selection using cross validation that is applicable to any induction algorithm and experiments conducted with ID3 and C4.5 on artificial and real datasets [8].

II. FEATURE SUBSET SELECTION

To remove irrelevant features and redundant features, the FAST [1] algorithm has two connected components. Irrelevant feature removal and redundant feature elimination. 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 our proposed FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the clusters.

3.1. Load Data

The data has to be pre-processed for removing missing values, noise and outliers. Then the given dataset must be converted into the arff format. From the arff format, only the attributes and the values are extracted and stored into the database. By considering the last column of the dataset as the class attribute and select the distinct class labels from that and classify the entire dataset with respect to class labels.

3.2. Entropy and Conditional Entropy Calculation

Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in terms of feature correlation and feature-target concept correlation.

In existing system to find the relevance of each attribute with the class label, Information gain is computed. This is also said to be Mutual Information measure.

Mutual information measures how much the distribution of the feature values and target classes differ from statistical independence. This is a nonlinear estimation of correlation between feature values or feature values and target classes. The Symmetric Uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of feature values or feature values and target classes, and has been used to evaluate the goodness of features for classification. The SU is defined

as follows:

$$SU(X, Y) = \frac{2 \times \text{Gain}(X|Y)}{H(X) + H(Y)}$$

Where, H(X) is the entropy of a random variable X. Gain(X|Y) is the amount by which the entropy of Y decreases. It reflects the additional information about Y provided by X and is called the information gain which is given by

$$\text{Gain}(X|Y) = H(X) - H(X|Y)$$

$$I_{XY} = -(H_{XY} - H_X - H_Y) = H(Y) - H(Y|X)$$

$$H_X + H_Y - H_{XY}$$

Where H(X|Y) is the conditional entropy which quantifies the remaining entropy (i.e., uncertainty) of a random variable X given that the value of another random variable Y is known. In proposed system, to find the relevance of each attribute with the class label entropy correlation coefficient is used. It can be calculated using following formula.

$$(12) \rho_H = \frac{\sqrt{\frac{I_{XY}}{\frac{1}{2}(H_X + H_Y)}}}{\sqrt{2 \left(1 - \frac{H_{XY}}{H_X + H_Y}\right)}}$$

WHERE, I IS MEAN DEPENDENCE INFORMATION AND IS GIVEN BY,

3.3. T-Relevance and F-Correlation Computation

The relevance between the feature $F_i \in F$ and the target concept C is referred to as the T-Relevance of F_i and C, and denoted by $SU(F_i, C)$. If $SU(F_i, C)$ is greater than a predetermined threshold, then F_i is a strong T-Relevance feature.

After finding the relevance value, the redundant attributes will be removed with respect to the threshold value. The correlation between any pair of features F_i and F_j ($F_i, F_j \in F \wedge i \neq j$) is called the F-Correlation of F_i and F_j , and denoted by $SU(F_i, F_j)$. The equation symmetric uncertainty which is used for finding the relevance between the attribute and the class is again applied to find the similarity between two attributes with respect to each label.

3.4. MST CONSTRUCTION

With the F-Correlation value computed above, the MST is constructed. A MST is a sub-graph of a weighted, connected and undirected graph. It is acyclic, connects all the nodes in the graph, and the sum of all of the weight of all of its edges is minimum. That is, there is no other spanning tree, or sub-graph which connects all the nodes and has a smaller sum. If the weights of all the edges are unique, then the MST is unique. The nodes in the tree will represent the samples, and the axis of the

n-dimensional graph represents the n features.

The complete graph G reflects the correlations among all the target-relevant features. Unfortunately, graph G has k vertices and $k(k-1)/2$ edges. For high-dimensional data, it is heavily dense and the edges with different weights are strongly interwoven. Moreover, the decomposition of complete graph is NP-hard. Thus for graph G , build an MST, which connects all vertices such that the sum of the weights of the edges is the minimum, using the well-known Kruskal's algorithm. The weight of edge (F_i, F_j) is F-Correlation $SU(F_i, F_j)$.

Kruskal's algorithm is a greedy algorithm in graph theory that finds a MST for a connected weighted graph. This means it finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized. If the graph is not connected, then it finds a minimum spanning forest (a MST for each connected component). If the graph is connected, the forest has a single component and forms a MST. In this tree, the vertices represent the relevance value and the edges represent the F-Correlation value.

3.5. PARTITIONING MST AND FEATURE SUBSET SELECTION

After building the MST, in the third step, first remove the edges whose weights are smaller than both of the T-Relevance $SU(F_i, C)$ and $SU(F_j, C)$, from the MST. After removing all the unnecessary edges, a forest F is obtained. Each tree $T_j \in F$ 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 $V(T_j)$ chooses a representative features whose T-Relevance is the greatest. All representative features comprise the final feature subset.

III. EXISTING SYSTEM

The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms. 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 subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and 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 learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods.

Disadvantages

1. The generality of the selected features is limited and the computational complexity is large.
2. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed.

IV. PROPOSED SYSTEM

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant

features. Our proposed FAST algorithm falls into the second group. Traditionally, feature subset selection research has focused on searching for relevant features. A well-known example is Relief which weighs each feature according to its ability to discriminate instances under different targets based on distance-based criteria function.

Advantages:

1. Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with each other.
2. The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

Application scenario: Bank Dataset

5.1 Main Modules:

5.1.1 Authentication Module :

In this module, Admins are having authentication and security to access the detail which is presented in the system. Before accessing or searching the details user should have the account in that otherwise they should register first.

5.1.2 Feature Cluster Analysis:

The various details generated here is analyzed to generate a weighted graph consisting of various values in the system as nodes and their relationships as node weights thereby creating a spanning tree mechanism.

5.1.3 Subset Selection Analysis:

Feature subset selection will identify the unwanted features from the medical dataset and eliminate the unwanted and redundant data to perform subset selection analysis using minimum spanning tree analysis.

5.1.4 Time Complexity Analysis :

This module will enable the user to analyze the time complexity of the system in graphical format for each attribute analyzed during subset analysis thereby enabling comparison of time taken for the algorithm to analyze different types of attributes during selection.

V. IMPLEMENTATION RESULTS

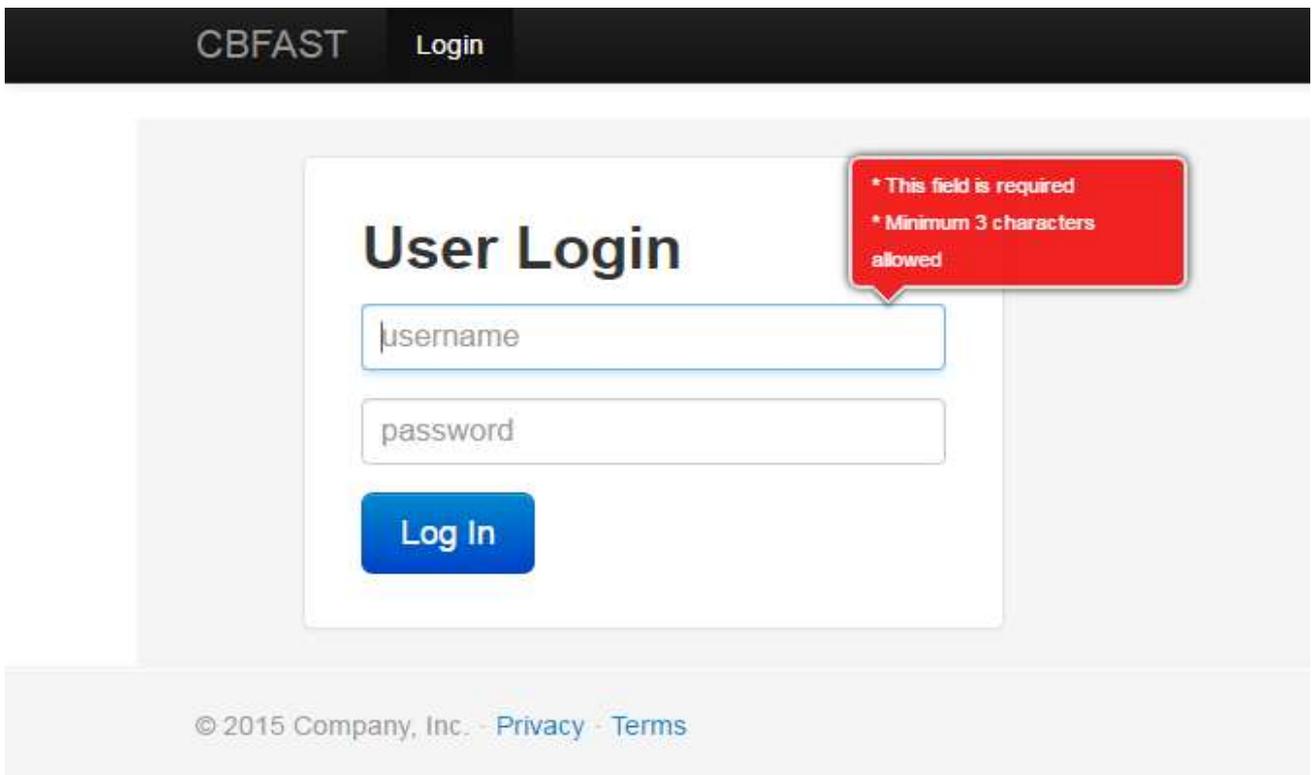


Figure.2 LogIn Screen

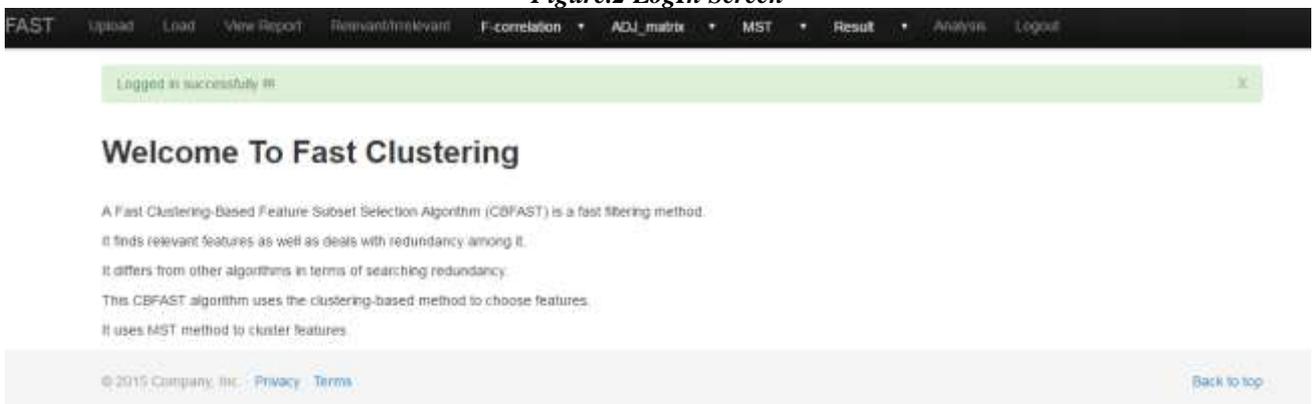


Figure.3 Welcome Screen

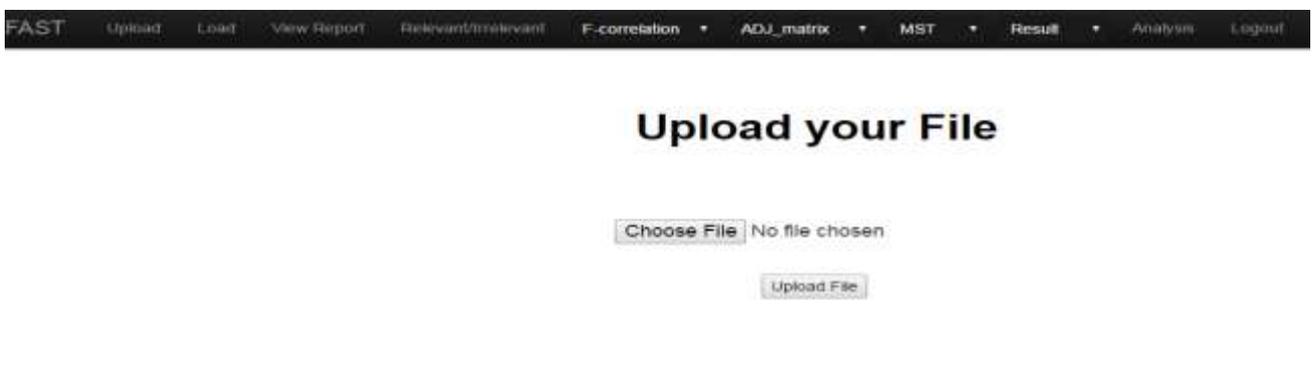


Figure.4 Data file upload screen

FAST Upload Load View Report Relevant/Irrelevant F-correlation ADJ_matrix MST Result Analysis Logout

Data Set is Successfully Loaded into Database

© 2015 Company, Inc. Privacy Terms Back to top

Figure 5 Successful database load

FAST Upload Load View Report Relevant/Irrelevant F-correlation ADJ_matrix MST Result Analysis Logout

Id	contract_no	product_code	prod_long_desc	sum_assured	bill_frequency	premium_cessation_term	status_code	premium_payingstatus	med_nonmed	mode_of_
1	57252	KEP	Kotak Endowment Plan	96000	4	20	SU	SU	NIL	C
2	120178	KSI	Kotak Safe Investment Plan	2272699.1	1	20	SU	SU	NIL	C
3	163552	KEP	Kotak Endowment Plan	90000	2	15	SU	SU	NIL	C
4	165767	KSI	Kotak Safe Investment Plan	110000	1	10	SU	SU	NIL	C
5	335398	KFP	Kotak Flexi Plan	160000	1	15	SU	SU	NME	C
6	357999	KSI	Kotak Safe Investment Plan	105000	1	10	SU	SU	NME	C
7	309498	KEG	Kotak Easy Growth Plan	21509.69	0	89	SU	SU	NME	N
8	361204	KFP	Kotak Flexi Plan	452000	1	25	SU	SU	NME	C

Figure.6 Report

FAST Upload Load View Report Relevant/Irrelevant F-correlation ADJ_matrix MST Result Analysis Logout

Feature Classification After T-Relevance Calculation

Threshold1 - 0.04532222478218756 Threshold2 - 0.5134914074042993 X

Selected Features (Relevant Features)

Feature Name	T-Relevance
status_code	1.0
premium_payingstatus	1.0
med_nonmed	0.3213807773273637
prod_long_desc	0.09187329941534309
link_unlinked_flag	0.30537460462177946
Premium_Paid	0.048327309702517925
first_issue_date	0.049083175259360935
agent_full_name	0.055250196833786416
premium_payer_date_of_birth	0.045505133596614934
combined_fund_value	0.536998848103398
status_change_date	0.07640884015456543

Selected Features (Relevant Features)

Feature Name	T-Relevance
status_code	1.4142135623730951
premium_payingstatus	1.4142135623730951
med_nonmed	0.9757045661965185
premium_payer_gender	0.6519679498148003
top_up_value	1.2456456115632857
link_unlinked_flag	0.9661107545070673
description	1.4142135623730951
agent_designation	1.1226830575294455
contract_owner_married_flag	0.6366796678903084
mode_of_payment	0.6489484230419696
la_gender	0.631000977648455

Figure.7 Relevant Data

Feature 1	Feature 2	Correlation
status_code	premium_payingstatus	1.0
status_code	description	1.0
status_code	combined_fund_value	0.7757598039442698
status_code	med_nonmed	0.6445019211552279
status_code	link_unlinked_flag	0.5716061368734199
status_code	prod_long_desc	0.31430801065075775
status_code	product_code	0.23485491758514582
status_code	contract_owner_education	0.1924351695350728
status_code	status_change_date	0.15860775446363295
status_code	contract_owner_occupation_code	0.1473186414894321
status_code	contract_owner_married_flag	0.14730562481988244
status_code	agent_full_name	0.12459923748504682
status_code	agent_client_id	0.110855600769436
status_code	agent_number	0.10165530326597794

Figure.8 Existing correlation

Feature 1	Feature 2	Correlation
status_code	premium_payingstatus	1.4142135623730951
status_code	description	1.4142135623730951
status_code	top_up_value	1.2737667123589145
status_code	agent_designation	1.1917689885629226
status_code	combined_fund_value	2.428628199080988
status_code	link_unlinked_flag	2.227763719341789
status_code	med_nonmed	2.084445430377563
status_code	value_of_pact_withdrawal	1.2242827241325913
status_code	premium_payer_gender	1.230596364904291
status_code	contract_owner_gender	1.2347915793207944
status_code	mode_of_payment	1.2405299618678347
status_code	contract_owner_married_flag	1.6214681888925222
status_code	la_gender	1.8653057054900683
premium_payingstatus	description	1.8480529092126832

Figure. 9 Proposed correlation

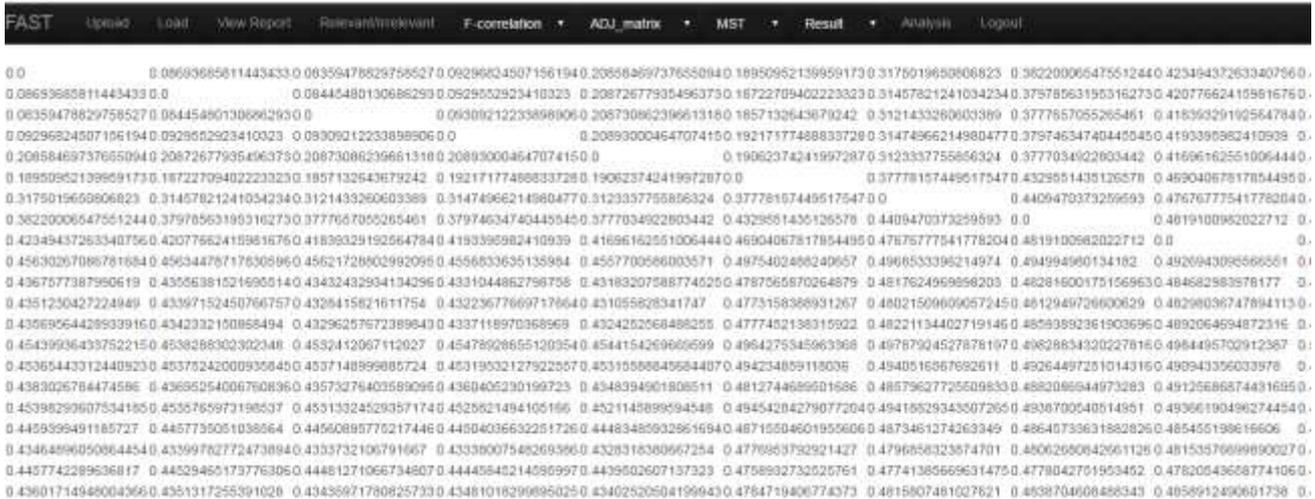


Figure.10 Existing ADJ Matrix

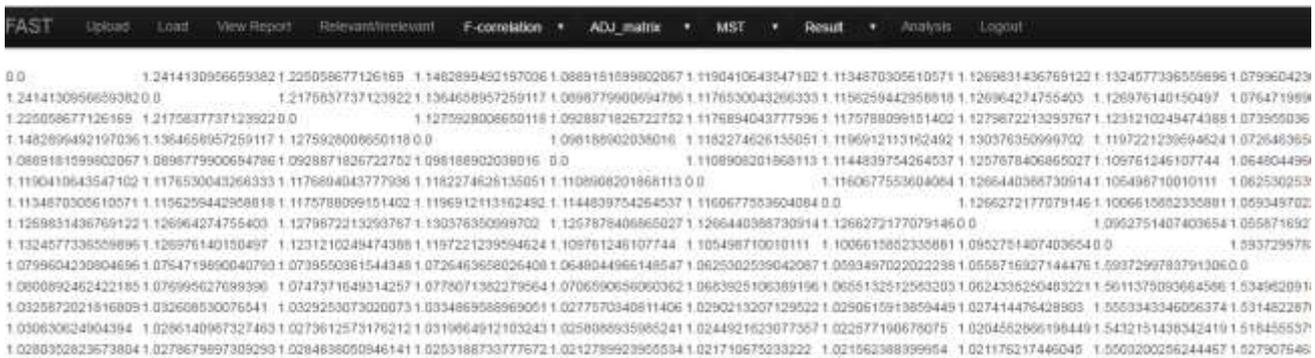


Figure.11 proposed ADJ Matrix

SR NO	Parent	Child	Value
1	16 (som_assured)	19 (contract_owner_occupation_code)	0.4363839920640604
2	5 (Premium_Paid)	4 (link_unlinked_flag)	0.19062374241997287
3	2 (med_nonmed)	6 (first_issue_date)	0.3121433260603389
4	2 (med_notimed)	5 (Premium_Paid)	0.1857132643679242
5	4 (link_unlinked_flag)	15 (risk_commencement_date)	0.4346394901808511
6	4 (link_unlinked_flag)	12 (agent_client_id)	0.4324252568489255
7	4 (link_unlinked_flag)	10 (status_change_date)	0.4318320758974525
8	4 (link_unlinked_flag)	11 (contract_owner_annual_income)	0.431055828341747
9	3 (prod_long_desc)	9 (combined_fund_value)	0.4556833635135984
10	4 (link_unlinked_flag)	20 (agent_number)	0.43402520504199943
11	2 (med_nonmed)	1 (premium_payingstatus)	0.0844580130686293
12	4 (link_unlinked_flag)	7 (agent_full_name)	0.3777034922803442
13	4 (link_unlinked_flag)	8 (premium_payer_date_of_birth)	0.41696162551006444
14	16 (contract_owner_education)	17 (contract_owner_married_flag)	0.4446762244123627
15	19 (contract_owner_occupation_code)	16 (contract_owner_education)	0.4428249015328281

Figure.12 Existing MST

SR.NO	Parent	Child	Value
1	13 (value_of_part_withdrawal)	1 (premium_payingstatus)	1.0278679697308293
2	13 (value_of_part_withdrawal)	0 (status_code)	1.0280352823673804
3	4 (top_up_value)	12 (contract_owner_gender)	1.0258068935985241
4	4 (top_up_value)	13 (value_of_part_withdrawal)	1.0212789623956534
5	4 (top_up_value)	11 (combined_fund_value)	1.0277570340811406
6	5 (link_unlinked_flag)	12 (contract_owner_gender)	1.0244921623077357
7	13 (value_of_part_withdrawal)	7 (agent_designation)	1.021176217446045
8	5 (link_unlinked_flag)	13 (value_of_part_withdrawal)	1.021710675233222
9	13 (value_of_part_withdrawal)	5 (link_unlinked_flag)	1.025710675233222
10	13 (value_of_part_withdrawal)	6 (description)	1.021962388399954
11	13 (value_of_part_withdrawal)	3 (premium_payer_gender)	1.0253188733777672
12	13 (value_of_part_withdrawal)	4 (top_up_value)	1.0212789623956534
13	7 (agent_designation)	13 (value_of_part_withdrawal)	1.021176217446045
14	11 (combined_fund_value)	7 (agent_designation)	1.027414476428903
15	12 (contract_owner_gender)	2 (med_nonmed)	1.0273612573176212

Figure.13 Proposed MST

FAST Upload Load View Report Relevant/Irrelevant F-correlation F-correlation MST Result Analysis Logout

Clusters Created

Cluster No.	Features	T-Relevance
1	[prod_long_desc, combined_fund_value]	[0.09187329941534339, 0.5369996848103398]
2	[contract_owner_occupation_code, product_code]	[0.07104883251666883, 0.08989858742287572]
3	[med_nonmed]	[0.3213807773273637]
4	[premium_payingstatus]	[1.0]
5	[status_code]	[1.0]

Final Subset of Features

[combined_fund_value, product_code, med_nonmed, premium_payingstatus, status_code]

GenerateCSV

Figure.14 Existing Cluster

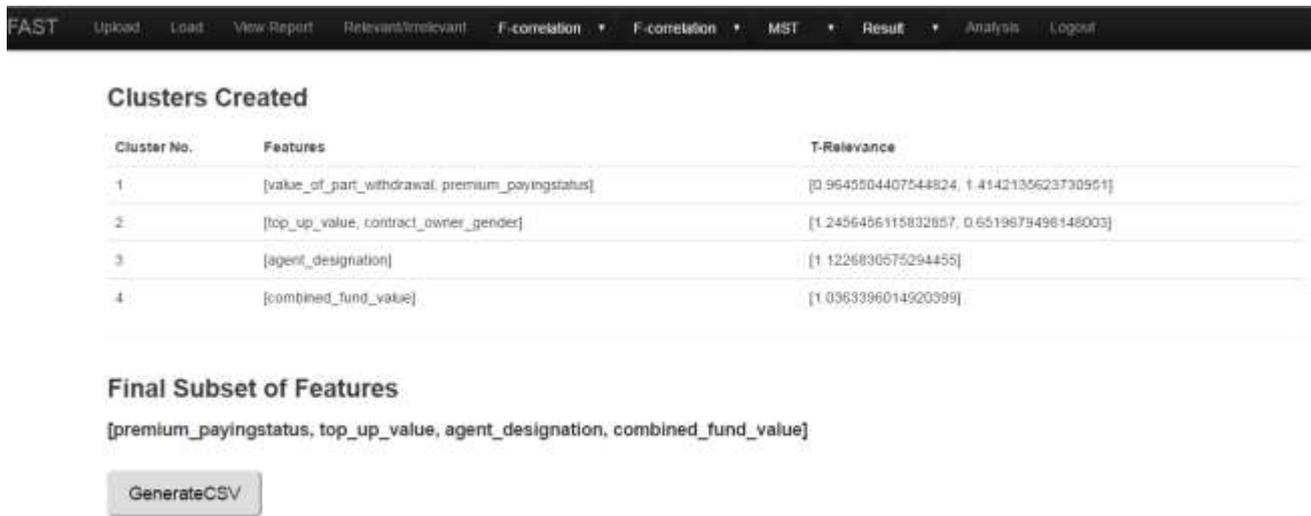


Figure.15 Proposed Cluster

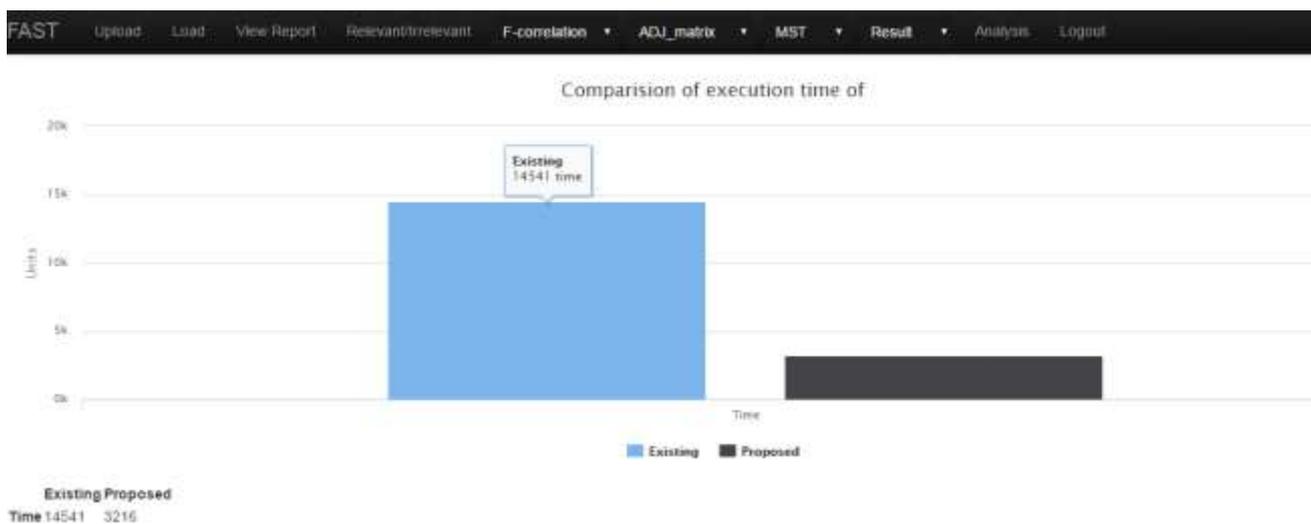


Figure.16 Time Analysis

VI. CONCLUSION

In data mining data from different domains is high dimensional. It can contain high degree of irrelevant and redundant information which may degrade the performance of learning algorithms. The objective is to get subset of good features from high dimensional dataset, to develop feature correlation technique by using entropy correlation coefficient and graph with respect to time showing comparison between prim's and kruskal's algorithm. For future scope it may include dynamic dataset.

VII. REFERENCES

- [1] Qinbao Song, Jingjie Ni, Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge & Data Engineering, vol.25, no. 1, pp. 1-14, Jan. 2013
- [2] Kashif Javed, Haroon A. Babri, Maureen Saeed, "Feature Selection Based on Class-Dependent Densities for High-Dimensional Binary Data", IEEE Transactions on Knowledge & Data Engineering, vol.24, no. 3, pp. 465-477, March 2012

- [3] Marcus Chen, Ivor W. Tsang, Mingkui Tan, Tat Jen Cham, "A Unified Feature Selection Framework for Graph Embedding on High Dimensional Data", IEEE Transactions on Knowledge & Data Engineering, vol.27, no. 6, pp. 1465-1477, June 2015
- [4] Marius Muja and David G. Lowe: "Scalable Nearest Neighbor Algorithms for High Dimensional Data". Pattern Analysis and Machine Intelligence (PAMI), Vol. 36, 2014
- [5] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Learning, 2003.
- [6] Hu Min, Wu Fangfang, "Filter-Wrapper Hybrid Method on Feature Selection", GCIS, 2010, 2010 Second WRI Global Congress on Intelligent Systems, 2010 Second WRI Global Congress on Intelligent Systems 2010, pp. 98-101, doi:10.1109/GCIS.2010.235
- [7] Das S, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection," Proc. 18th Int'l Conf. Machine Learning, pp. 74-81, 2001.
- [8] G.H. John, R. Kohavi, and K. Pflieger, "Irrelevant Features and the Subset Selection Problem," Proc. 11th Int'l Conf. Machine Learning, pp. 121-129, 1994.
- [9] Vipin Kumar, Sonajharia Minz: Feature Selection: A literature Review. Smart CR 4(3): 211-229 (2014)
- [10] Molina L.C., Belanche L. and Nebot A., Feature selection algorithms: A survey and experimental evaluation, in Proc. IEEE Int. Conf. Data Mining, pp 306-313, 2002
- [11] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufman, 1993
- [12] J.W. Jaromczyk and G.T. Toussaint, "Relative Neighborhood Graphs and Their Relatives," Proc. IEEE, vol. 80, no. 9, pp. 1502-1517, Sept. 1992