7th International Conference on Recent Development in Engineering Science, Humanities and Management
National Institute of Technical Teachers Training & Research, Chandigarh, India
3rd June 2017, www.conferenceworld.in
(ESHM-17)
ISBN: 978-93-86171-26-9

CLUSTERING-BASED FEATURE SUBSET SELECTION ALGORITHM USING FAST

Mr. Akshay S. Agrawal

Assistant Professor, Department of Computer Engineering, SSJCET, Asangaon, (India)

ABSTRACT

Feature selection involves 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. Duplicate detection is the process of identifying multiple representations of same real world entities. Today, duplicate detection methods need to process ever larger datasets in ever shorter time: maintaining the quality of a dataset becomes increasingly difficult. The algorithm maximizes the gain of the overall process within the time available by reporting most results much earlier than traditional approaches.

The FAST 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 that is strongly related to 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. To ensure the efficiency of FAST, the efficient minimum-spanning tree (MST) clustering method is adopted.

Keywords: Clustering, Feature subset selection, Minimum Spanning Tree, T-Relevance, F-Correlation.

I. INTRODUCTION

Data are among the most important assets of a company, but due to data changes and sloppy data entry, errors such as duplicate entries might occur, making data cleansing and in particular duplicate detection indispensable. However, the pure size of today's datasets renders duplicate detection processes expensive. This project identifies most duplicate pairs early in the detection process. Instead of reducing the overall time needed to finish the entire process, progressive approaches try to reduce the average time after which a duplicate is found. Early termination in particular then yields more complete results on a progressive algorithm than on any traditional approach.

With the aim of choosing a subset of good features with reverence to the target concepts, feature subset is an efficacious way for reducing dimensionality, abstracting extraneous data, incrementing learning precision, and

(ESHM-17)
ISBN: 978-93-86171-26-9

amending result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches[1]. The embedded methods incorporate feature selection component of the training process and are conventionally concrete to given learning algorithms, and consequently may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods utilize the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the precision of the learning algorithms is customarily high. However, the generality of the selected features is inhibited and the computational intricacy is immensely colossal. 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 ensured. 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 homogeneous time complexity of the filter methods.

The wrapper methods are computationally extravagant and incline to over fit on small training sets [1]-[9]. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, focus will be on the filter method in this report. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms. In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Their results have, sometimes, the best agreement with human performance. The general graph-theoretic clustering is simple: Compute a neighborhood graph of instances, then delete any edge in the graph that is much longer/shorter than its neighbors. The result is a forest and each tree in the forest represents a cluster.

In the proposed study, a graph theoretic clustering method has been applied to features. In particular, the minimum spanning tree (MST) predicated clustering algorithms is adopted because it is assumed that data points are grouped around centers or dissevered by a customary geometric curve and have been widely utilized in practice.

Based on the MST method, a **F**ast clustering-b**A**sed feature **S**election algori**T**hm (FAST) is proposed. The FAST 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 that is strongly related to target classes is selected from each cluster to form the final 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

ISBN: 978-93-86171-26-9

II REVIEW OF LITERATURE

Table No. 2.1: Comparative Study

Sr. No	Paper Name	Author Name	Technology / Algorithm	Advantages	Disadvantages
01	A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data [1]	Qinbao Song, Jingjie Ni and Guangtao Wang	FAST	1.Improve the performance of The classifiers. 2.Efficiently and effectively deal with both irrelevant and redundant features, and obtains	--
02	Fast Correlation Based Filter (FCBF) with a Different Search Strategy [1]	Baris Senliol, Gokhan Gulgezen1 , Lei Yu2 and Zehra Cataltepe	FCBF	1. FCBF compares only individual features with each other. 2. Select fewer features with higher accuracy.	Cannot detect some features.
03	An efficient k-means clustering algorithm: analysis and implementation [14]	T. Kanungo	K-MEAN	1. K-Means produce tighter clusters. 2. Fast, robust and easier to understand.	1. Fixed number of clusters can make it difficult to predict. 2. Does not work well with non-globular clusters. 3. Different initial partitions can result in different final clusters. 4. It does not work well with clusters (in the original data) of Different size and Different density.

ISBN: 978-93-86171-26-9

04	A Simulated Annealing-Based Multiobjective Optimization Algorithm [15]	Sanghamitra Bandyopadhyay	Simulating Annealing	Accuracy, Useful for small datasets	Single feature for Single turn.
05	Clustering Approach to Collaborative Filtering Using Social Networks [16]	Emir Cogo and Dzenana Donko	Filter Approach	Suitable for very large features.	Accuracy is not Guaranteed.

III EXISTING SYSTEM

The embedded method incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected 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.

Limitations:

- Lacks speed
- Cannot detect some features.
- Performance Related Issues

The generality of the selected features is limited and the computational complexity is large.

The accuracy of the learning algorithms is not guaranteed.

So the focus of our new system is to enhance the throughput for any basis to eliminate the data security lacks therein and make a system prominent handler for handling data in an efficient manner.

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

(ESHM-17)
ISBN: 978-93-86171-26-9

features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. Overall the system will be effective in generating more relevant and accurate features which can provide faster results. To remove irrelevant features and redundant features, the FAST [14] algorithm has two connected components. Irrelevant feature removal and redundant feature elimination are the two most important aspect of FAST Algorithm. 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.

Advantages:

Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.

The efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset.

Generally the algorithm achieves significant reduction of dimensionality by selecting only a small portion of the original features.

Generally the algorithm is also used for increasing learning accuracy and improving result comprehensibility.

V. PROPOSED SYSTEM DESIGN

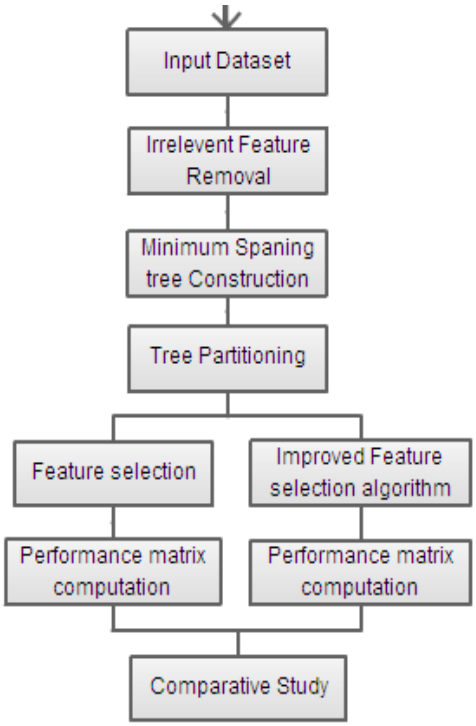

Figure No. 5.1: System Flow

ISBN: 978-93-86171-26-9

Algorithm:

Algorithm 1: FAST

inputs: $D(F1, F2, ..., Fm, C)$ - the given data set

θ- the T-Relevance threshold.

output: S - selected feature subset .

//==== Part 1 : Irrelevant Feature Removal ====

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

2 T-Relevance = **SU** (Fi, C)

3 **if** T-Relevance $> \theta$ **then**

4 $S = S \cup \{F\}$;

//==== Part 2 : Minimum Spanning Tree Construction ====

5 G = NULL; //G is a complete graph

6 **for** *each pair of features* $\{F'_i, F'_j\} \subset S$ **do**

7 F-Correlation = **SU** (F'_i, F'_j)

8 *Add* F'_i *and/or* F'_j *with* F-Correlation as the weight of the corresponding edge;

9 minSpanTree = *Prim/Kruskal*(G); //Using Prim/Kruskal Algorithm to generate the min spanning tree

//==== Part 3 : Tree Partition and Representative Feature Selection ====

10 Forest = minSpanTree

11 **for** *each edge* $E_{i,j} \in$ Forest **do**

12 **if SU** (F'_i, F'_j) **< SU(**F'_i, C**)** $\wedge$ **SU**(F'_i, F'_j)<**SU(**F'_j, C**) then**

13 Forest = Forest $- E_{ij}$

14 $S = \phi$

15 **for** *each tree* $T_i \in$ Forest **do**

16 F^j_R=argmax$F_k \in T_i$**SU(**F'_k, C**)**

17 $S = S \cup \{F^j_R\}$;

18 return S

The major amount of work for Algorithm 1 involves the computation of SU values for *T-relevance* and *F-Correlation*, which 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 (m) in terms of the number of features m. Assuming ($1 \leq k \leq m$) features are selected as relevant ones in the first part, when k= 1, only one feature is selected. Thus, there is no need to continue the rest parts of the algorithm, and the complexity is (m). When $1 < k \leq m$, the second part of the algorithm firstly constructs a complete graph from relevant features and the complexity is $O(k2)$, and then generates a MST from the graph using Prim algorithm whose time complexity is $O(k2)$. The third part partitions the MST and chooses the representative features with the complexity of (k). Thus when $1 < k \leq m$, the complexity of the algorithm is ($m+k2$).

ISBN: 978-93-86171-26-9

This means when $k \leq \sqrt{m}$, FAST has linear complexity (m), while obtains the worst complexity ($m2$) when $k = m$. However, $k$is heuristically set to be $\lfloor \sqrt{m} \lg m \rfloor$in the implementation of FAST. So the complexity is ($m \lg 2m$), which is typically less than ($m2$) since $\lg 2m < m$. This can be explained as follows. Let (m) = $m - \lg 2m$, so the derivative $f'(m) = 1 - 2 \lg e/m$, which is greater than zero when $m > 1$. So (m) is a increasing function and it is greater than (1) which is equal to 1, i.e., $m > \lg 2m$, when $m > 1$. This means the bigger the $m$is, the farther the time complexity of FAST deviates from ($m2$). Thus, on high dimensional data, the time complexity of FAST is far more less than ($m2$). This makes FAST has a better runtime performance with high dimensional data.

Mathematical Model:

H(X) is the entropy of a discrete random variable X. Let (x) be the prior probabilities for all values of X, then (X) is defined by,

$$H(X) = -\sum_{x \varepsilon X} p(x) \log_2 p(x)$$

Gain (X | Y) determines the amount by which the entropy of Y decreases. It is given by,

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

Where, H(X | Y) is the conditional entropy and is calculated as,

$$H\left(\frac{X}{Y}\right) = -\sum_{y \varepsilon Y} p(y) \sum_{x \varepsilon X} p(x) \log_2 p(x)$$

Where, X is a Feature and Y is a Class.

The **symmetric uncertainty (SU)** is defined as follows,

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

Given that (X, Y)be the symmetric uncertainty of variables X and Y, the relevance T-Relevance between a feature and the target concept C, the correlation F- Correlation between a pair of features, the feature redundancy F-Redundancy and the representative feature R- feature of a feature cluster can be defined as follows. T-Relevance - The relevance between the feature Fi ∈ F and the target concept is referred to as the T-Relevance of Fi and C, and denoted by SU (Fi,C). If SU (Fi,C) is greater than a predetermined threshold θ,Symmetric Uncertainty of each Feature is greater than the T-Relevance threshold is checked.

ISBN: 978-93-86171-26-9

VI. RESULT

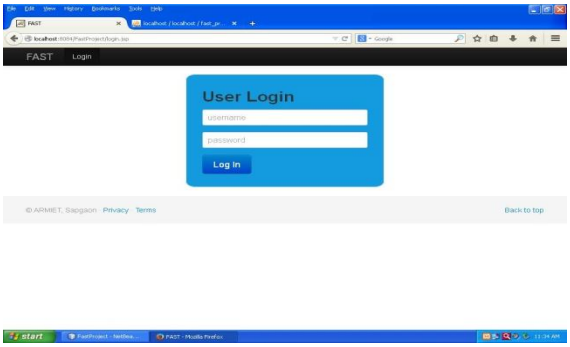

Figure No. 6.1: Home Page

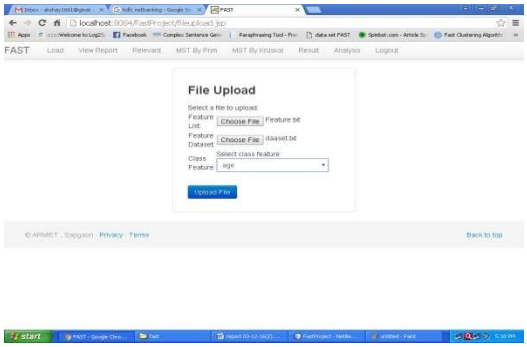

Figure No. 6.2: Database Uploading

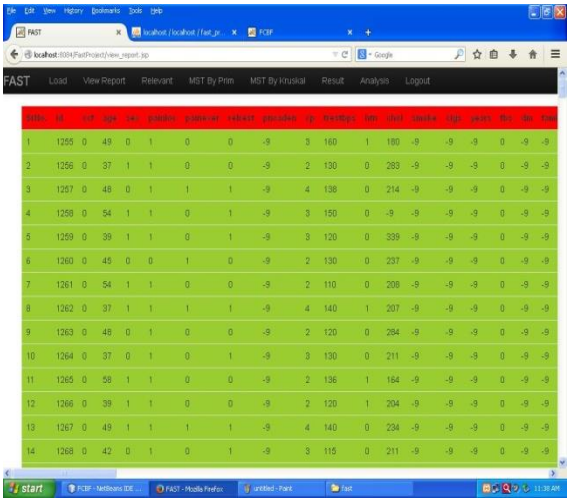

Figure No. 6.3: Report

Figure No. 6.4: Relevant Features

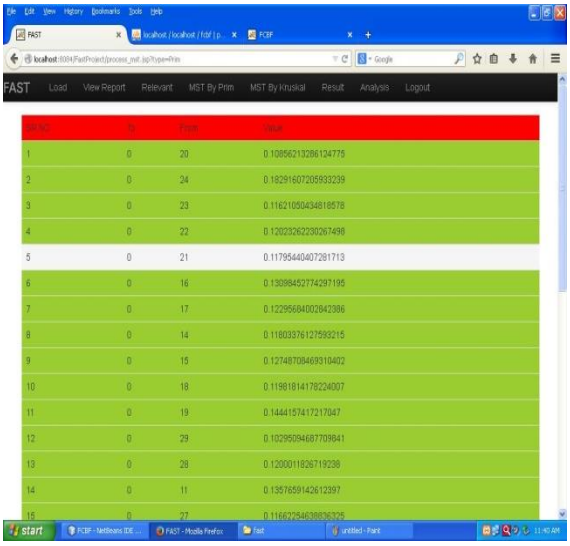

Figure No. 6.5: MST by Prim's

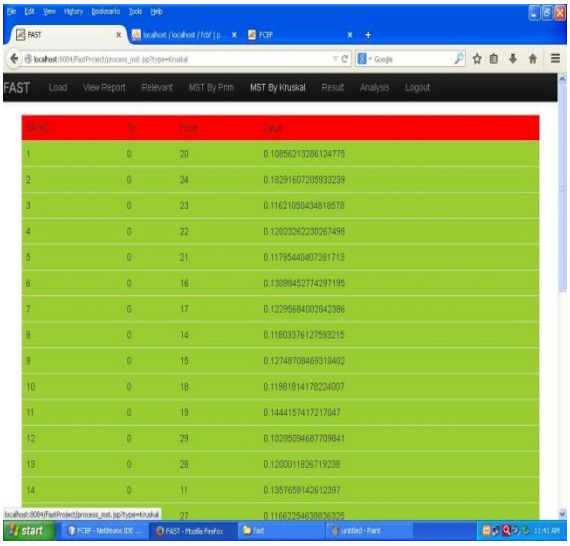

Figure No. 6.6: MST by Kruskal's

ISBN: 978-93-86171-26-9

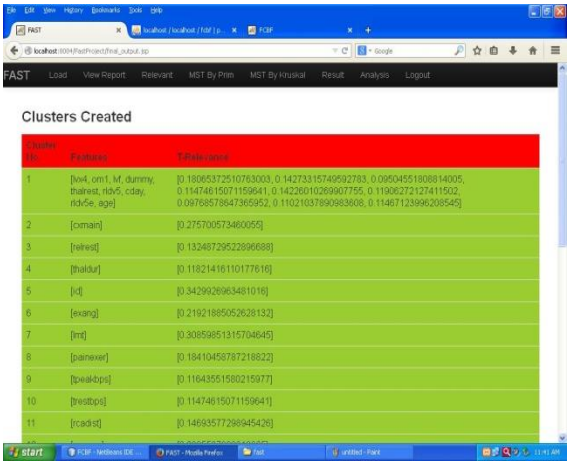

Figure No. 6.7: Cluster's

VII. ANALYSIS

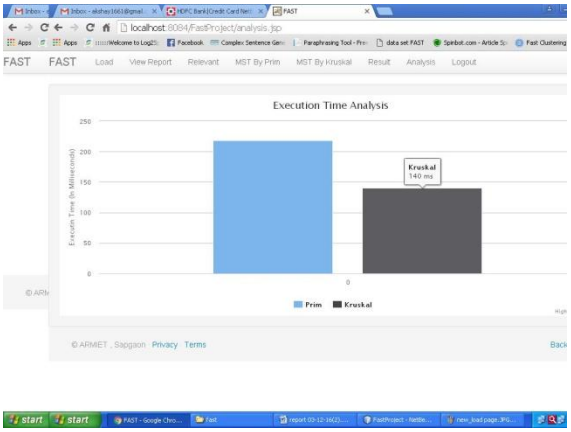

Figure No. 7.1: Analysis

VIII FUTURE SCOPE

For the future it is planned to explore different types of correlation measures, and study some formal properties of feature space.

Different correlation measure along with fuzzy logic can be included in the present algorithm. Symmetric Uncertainty can be extended for extracting the feature subset selection.

IX. CONCLUSION

This paper explains about the feature subset selection. In the proposed algorithm, a cluster consists of features. . Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

ISBN: 978-93-86171-26-9

Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy.

A novel clustering-based feature subset selection algorithm for high dimensional data is presented. The algorithm involves removing irrelevant features, constructing a minimum spanning tree from relative ones, and partitioning the MST and selecting representative features.

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 High Dimensional 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] Ding Cheng Feng, Feng Chen_, and Wenli Xu "Efficient Leave-One-Out Strategy for Supervised Feature Selection" TSINGHUA SCIENCE AND TECHNOLOGY ISSNll1007- 0214ll09/10llpp629 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" Springer Plus 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] Jesna Jose,"Fast for Feature Subset Selection Over Dataset" International Journal of Science and Research (IJSR), Volume 3 Issue 6, June 2014.

[11] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96- 103, 1998.

[12] Fayyad U. and Irani K., Multi-interval discretization of continuous-valued attributes for classification learning, In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, pp 1022-1027, 1993.

ISBN: 978-93-86171-26-9

[13] Senliol, Baris, et al. "Fast Correlation Based Filter (FCBF) with a different search strategy." *International Symposium on Computer and Information Sciences (ISCIS 2008)*. Istanbul Technical University, Suleyman Demirel Cultural Center, Istanbul, Turkey, 2008.

[14] Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE transactions on pattern analysis and machine intelligence 24.7 (2002): 881-892.

[15]Bandyopadhyay, Sanghamitra, et al. "A simulated annealing-based multiobjective optimization algorithm: AMOSA." IEEE transactions on evolutionary computation 12.3 (2008): 269-283.

[16] Cogo, Emir, and Dzenana Donko. "Clustering approach to collaborative filtering using social networks." 2013 IEEE 4th International Conference on Electronics Information and Emergency Communication. 2013.

[17] Agrawal Akshay S., and S. Bojewar. "**Comparative study of various clustering techniques**" *International Journal of Computer Science and Mobile Computing* 3.10 (2014): 497-504.

[18] Agrawal Akshay S., and S. Bojewar. "**A FAST clustering based Feature Subset Selection Algorithm"** *International Research Conference on Recent Innovations in Science, Engineering and Management in International Journal of Advance.*