ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Implementation of MST Based Feature Subset Selection Process Using FAST Algorithm

Chavan Akshay S.[1], Hambarde Balaprasad S.[2]

Assistant Professor, Dept. of C.S.E, MPGI SOE, S.R.T.M.U, Nanded, Maharashtra, India[1]

Assistant Professor, Dept. of C.S.E, MPGI SOE, S.R.T.M.U, Nanded, Maharashtra, India[2]

ABSTRACT: The Feature selection is very important process for selecting a subset of features from original data set that containing huge amount of data. By eliminating irrelevant features and ignoring redundant features to improve accuracy in projecting information. Most of the data mining task are relay on selection criteria of Feature selection.Because of this reasons feature selection is the most important field of research in data mining in which most innovative achievements have been reported. In this implementation the minimum spanning tree (MST) - based feature selection based algorithms is adopted. A Using Fast clustering based feature Selection algorithm (FAST) is based on MST method, In the FAST algorithm, features are divided into clusters by using graph-theoretic clustering methods and then, the most representative feature that is strongly related to target classes is selected This implementation provides the clear insight to application of FAST algorithm for feature subset selection. The main theme of this implementation of FAST algorithm is to find correlation between features and generating a MST by handling each feature as a cluster and then choosing relevant feature from these clusters. Instead of linear correlation based measure this implementation comprises use of nonlinear correlation measure i.e. symmetric uncertainty equation is used for finding correlation between features. MST constructed using relevance score of feature's as a weight of edge between vertices compared with target relevance score.

KEYWORDS: Feature clustering, Minimum spanning tree, Target relevance Score, Feature correlation score, Symmetric uncertainty.

I. INTRODUCTION

Data mining is a process of knowledge discovery in large data sets by performing its own Task. The overall goal of the data mining process is to gather the information from a data set for providing knowledge base information that relevant to user requirement. In data mining field the Feature selection is one of the important and frequently used techniques in data reduction and preprocessing for data mining.

A. *Feature selection process.*

Feature selection is one of the significant and regularly used technique in data reduction and pre-processing of the data for data mining. The Feature selection is very useful technique to select a subset of features from data set containing huge amount of data by reducing irrelevant features to improve projecting information. Feature selection is commonly used as a pre-processing step to machine learning [9]. It is a process of selecting a subset of original features so that the feature space is optimally reduced according to a certain evaluation measure. Feature selection has been a productive field of research and development since 1970's and proven to be operative in removing irrelevant and redundant features, increasing effectiveness in learning tasks, improving learning performance like predictive accuracy, and augmenting clarity of learned results. Therefore, feature selection becomes very essential for machine learning tasks when facing high dimensional data. However, this trend of flagrancy on both size and dimensionality also postures severe challenges to feature selection algorithms. Certainrecent research determinations in feature selection have remained focused on these challenges from handling a huge number of instances to dealing with high dimensional data [3].

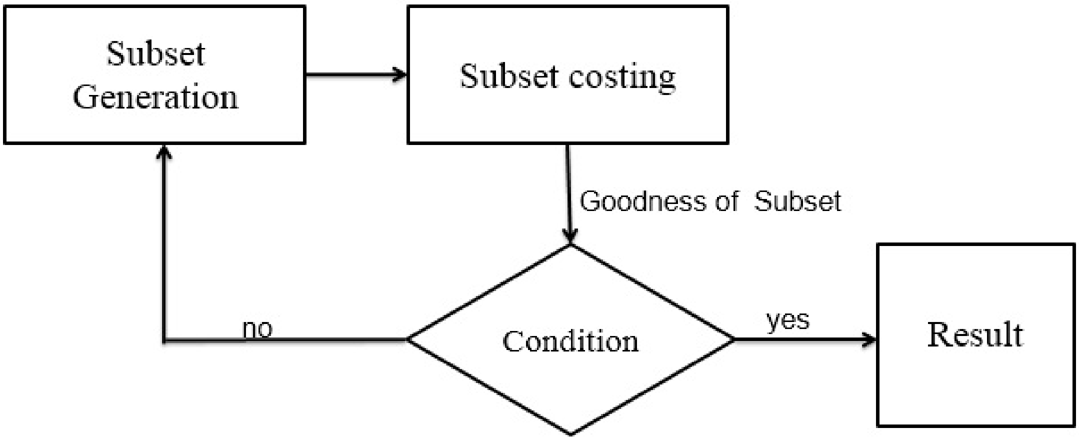

Fig. 1. Typical feature selection process

B. *Feature selection categories.*

A different Feature subset selection processes are classified into three main groups: filter methods, wrapper methods [17] and embedded methods. Filter methods depend on general appearances of the training data to evaluation and selecting subsets of features without comprising a learning algorithm. Divergent to that, wrapper approaches use a classification algorithm as a black container to weigh the prediction accuracy of various subsets. The last group, embedded approaches, does the feature selection process as afundamentalmeasure of the machine learning algorithm [14], [17].

Filter Methods:

Filter methods are classifier sceptical, no-feedback, preselecting methods that are independent of the machine learning algorithm to be applied [10].Filter methods can further be divided into two techniques i.e. univariate and multivariate techniques [14]. univariate methods consider features separately and typically make use of certain scoring function to assign weights to features independently and rank them on the basis of their relevance score to the target concept.In the literature, this procedure is commonly known as feature ranking or feature weighting. A feature will be selected if its weight or relevance greater than threshold value. Members of this second group of filter methods also mentioned to as subset search evaluationby search through the applicant's feature subsets guided by a certain estimation measure which captures the quality of each subset not only the individual predictive power of single features[18].

Wrapper Methods:

Wrapper methods are feedback methods which combine the machine learning algorithm in the feature selection process, i.e., they depend on the performance of a specific classifier. Here, the classification algorithm is used as a black box. Wrapper methods search through the space of feature subsets using a learning algorithm to monitor the search. To search the space of different feature subsets, a search algorithm is "wrapped" around the classification model. In a search procedure the space of possible feature subsets is well-defined and produced various subsets of features. This subset's estimated the classification accuracy of the learning algorithm for each dataset [14], [17].

Embedded Approaches:

Embedded approaches, sometimes also referred to as nested subset methods [14], doing as an integral part of the machine learning algorithm itself. For the duration of the operation of the classification process, the algorithm itself adopts which attributes to use and which to ignore. Just identical to wrapper methods, embedded approaches consequently depend on aexact learning algorithm, but might be further efficient in some characteristics. Moreover, this approach comprises the better use of the available data by not needing to split the data into a training and test/validation set. Decision trees are well-known examples that use implanted feature selection approach throughpicking the attribute that achieves the "best" split in relations of class dissemination at each leave. This procedure is repeatedon the feature subsets till some stopping criterion is fulfilled [1] [14].

II. RELATED WORK

A. *Framework of FAST algorithm.*

Aim of feature selection process is selecting a best subset of features by eliminating irrelevant and redundant features without using predictive information. It is a process of selecting a subset of original features according to specific criteria. Irrelevant features are not applicable for the accuracy of feature selection result. Redundant features ordinarily provide the facts which is already present in other features. There are many feature selection algorithm are existing, some of these are useful for removing irrelevant features but not effective for handling redundant features. However some of other can eradicate irrelevant feature while taking care of redundant features. One of the feature selection algorithms is Relief [18], which evaluates each feature according to its ability to distinguish instances under different objectives based on nearest-near criteria function. Though, Relief is useless for removing redundant features as two prognostic but highly correlated features are expected both to be highly weighted. Relief-F [18] extends Relief, aiding this method to work with noisy and incomplete data sets and to deal with multiclass problems, but still cannot detect redundant features. Redundant features also affect the accuracy and speed of learning algorithm; it is necessary to remove it. CFS, FCBF are examples that taken into consideration for removing redundant features [6], [10]. CFS is achieved by the assumption that a worthy feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other [12]. FCBF is a fast filter method which can identify relevant features as well as redundancy surrounded by relevant features without pair wise correlation analysis [10]. Different from above algorithms, FAST algorithm uses minimum spanning tree-based method to cluster features.

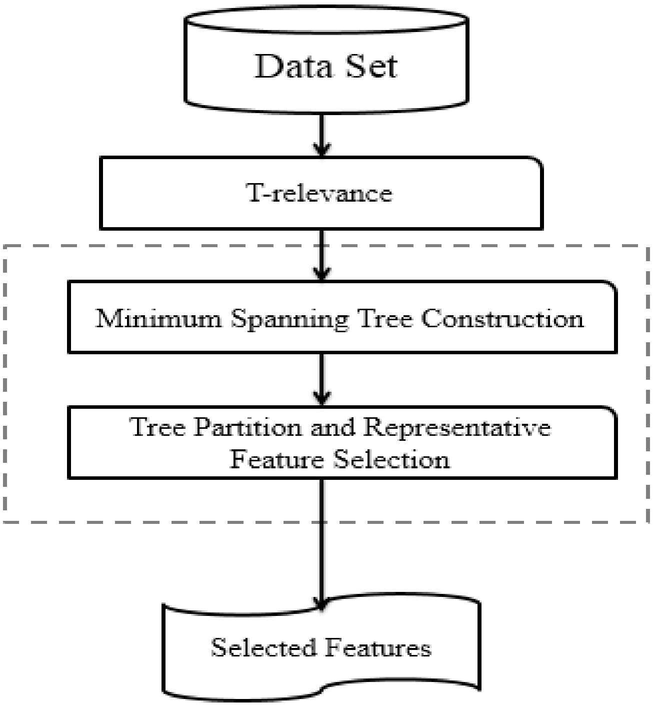
Data Set
T-relevance
Minimum Spanning Tree Construction
Tree Partition and Representative Feature Selection
Selected Features

Fig. 2. Framework of FAST algorithm.

FAST algorithm, it involves 1) the construction of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the minimum spanning tree into a forest such that each tree representing a cluster; and 3) and then the selection of representative features from the clusters. Relevant features have strong correlation with target concept hence they are always needed for a best subset, while redundant features are not needed because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally defined in terms of feature correlation and feature-target concept correlation.

B. *Correlation based measures.*

If correlation between two variables is adopted as a goodness measure, then feature is good if it is highly correlated to the class but not highly correlated to other features. To evaluate the goodness of features, a feature is good if it is relevant to the class concept but it should not redundant to any of the other relevant features [18]. This section comprise problem statement i.e., how to evaluate the goodness of features for classification. If we adopt the correlation among two features as a goodness measure, then above definition becomes that a feature is good if it is extremely correlated to the class but not extremely correlated to several of the other features. In other words, if the correlation between a feature and the class is high enough to make it relevant to the class and the correlation between it and any other relevant features does not reach at that level so that it can be predicted by any of the other relevant features, it will be considered as a good feature for the classification task. In this sense, the problem of feature selection boils down to find a suitable measure of correlations between features and a sound procedure to select Features based on this measure [18].

There exist broadly two approaches to measure the correlation:

1) Linear correlation coefficient.

2) Information-theoretical concept of entropy. (Symmetrical uncertainty).

Linear correlation coefficient.

For a pair of variables (X; Y), the linear correlation coefficient r is given by the formula:

$$r = \frac{\sum_i \left(x_i - \overline{x}_i\right)\left(y_i - \overline{y}_i\right)}{\sqrt{\sum_i \left(x_i - \overline{x}_i\right)^2}\sqrt{\sum_i \left(y_i - \overline{y}_i\right)^2}}$$

Where $\overline{x_i}$ is the mean of X, and $\overline{y_i}$ is mean of Y. the value of r lies between -1 and 1, inclusive. If X and Y completely correlated, r takes the value of 1 or -1; if X and Y are totally independent, r is Zero. It is a symmetrical measure for two features. Other measures in this type are basically variations of the above formula, such as least square regression error and maximal information compression index. There are several benefits of choosing linear correlation as a feature goodness measure for classification. First, it helps remove features with near zero linear correlation to the class. Second thing, it helps to reduce redundancy between selected features. It is known that if data is linearly distinguishable in the original depiction, it is still linearly separable if all but one of a group of linearly dependent features are removed [18].Linear correlation measures might not be able to capture correlations that are not linear in nature. Another limitation is that the calculation needs all features that encompass numerical values.

Information-theoretical concept of entropy. (Symmetrical uncertainty):

The symmetric uncertainty is consequential from the mutual information by normalizing it to the entropies of feature values or feature values and target classes.Symmetrical uncertaintyhas been used to evaluate the goodness of features [18] [15]. Therefore, we take symmetric uncertainty as the measure of correlation between either two features or a feature and the target concept.

The symmetric uncertaintyis well-defined in following equations:

$$SU(A, B) = \frac{2 * Gain(A|B)}{H(A) + H(B)}$$

Where,

H (A) is the entropy of discrete random variable A. Assumep (a) is the prior probabilities for all Values of A.

H (A) is defined by:

$$H(A) = -\sum_{a \in A} p(a)\log_2 p(a)$$

Gain (A|B) is the amount by which the entropy of Bdeclines. It returns the additional information about B provided by A and is called the information gain which is given by:

$$\begin{aligned} gain(A|B) &= H(A) - H(A|B) \\ &= H(B) - H(B|A) \end{aligned}$$

Where H (A|B) is the conditional entropy which quantifies the remaining entropy (i.e. uncertainty) of a random variable A given that the value of another random variable Bis known. Suppose (a) is the prior probabilities for all values of A and (a|b) is the posterior probabilities of A given the Values of H (A|B), is defined by:

$$H(A|B) = -\sum_{b \in B} p(b)\sum_{a \in A} p(a|b)\log_2 p(a|b)$$

Information gain is a symmetrical measure. That is the aggregate of information is gained about after observing is equal to the amount of information gained about after observing. This confirms that the order of two variables (e.g. (A, B) or (B, A)) will not affect the value of measure.Symmetric uncertainty treats a pair of variables symmetrically, it compensates for information gain'spreference toward variables with more values and normalizes its value to the range [0, 1] A value 1 of SU (A, B) indicates that information of the value of either one fully predicts the value of the other and the value 0 reveals that and are independent. Though the entropy based measure handles nominal or discrete variables, they can deal with continuous features as well, if the values are discretized properly in advance [15].

III. Implementation of System

Implementation of FAST algorithm that can produce relevant information comprises calculation of T-relevance between features and target concept and F-correlation between features. T-relevance and F-correlation score is being calculated by using correlation based measure. Here in this implementation nonlinear correlation measure is used i.e. symmetric uncertainty. Feature subset selection can be observed as the process of identifying and removing as many irrelevant and redundant features as possible. Irrelevant features, along with redundant features, severely affect the accuracy of the learning machines. Thus, feature subset selection should be able to recognize relevant features and remove as much of the irrelevant and redundant features as possible. Moreover, "good feature subsets contain features highly correlated with each other, however irrelevant features are uncorrelated with each other. In our proposed FAST algorithm, it involves the construction of the minimum spanning tree (MST) from a weighted complete graph and the partitioning of the MST into a forest. Each tree representing a cluster, then selecting the representative features from the clusters.

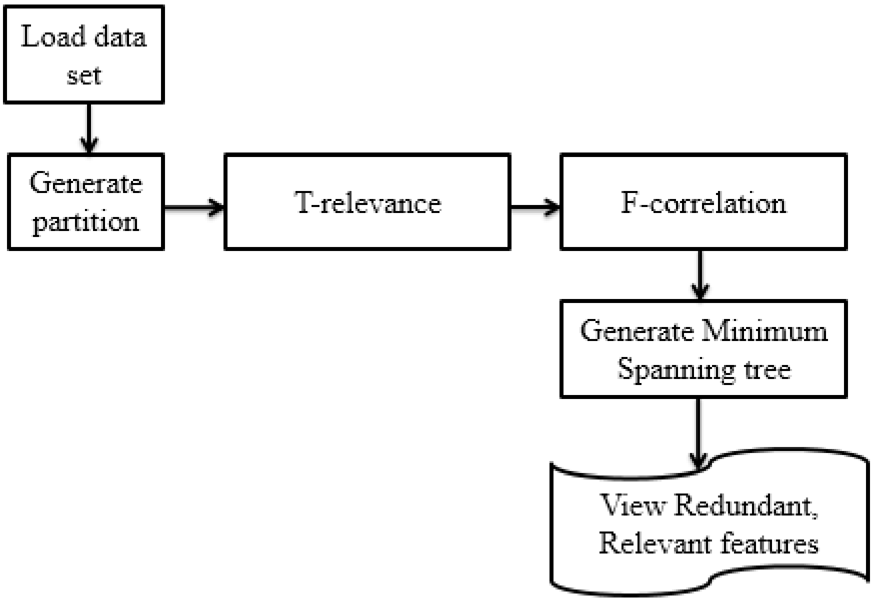
Load data set
Generate partition
T-relevance
F-correlation
Generate Minimum Spanning tree
View Redundant, Relevant features

Fig. 3. Design for implementation of system.

A. *T-Relevance calculation.*

This step comprises calculation of relevance between the feature and the target concept is referred to as the T-Relevance of Feature. By Using Mutual information measure i.e.; symmetric uncertainty relevance between features and target concept will be calculated, this is a nonlinear estimation of correlation between feature values and target concept. The relevance between the feature $f_i \in f$ and the target concept C is referred as the T-Relevance of feature f_i and target concept C, it is denoted by$SU(f_i$, C). If$SU(f_i$, C) 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 [15]. Once user provide the input value as a class label index, then T-relevance score being calculated on the basis of input value. For example if class label is 4 so correlation of feature and target concept C is calculated using symmetric uncertainty equation. If T-relevance score of cluster id 9= 0.11 then it is denoted by SU (xi, C) = SU (9, 4) =0.11.

B. *F-correlation calculation.*

This step is important comprises calculation of *F*-Correlation between Features. The correlation between any pair of features is called the F-Correlation of Feature with other feature. Equation of symmetric uncertainty is used for finding the relevance between the features is again applied to find the correlation between the features. In this step we need to calculate $\mathrm{SU}(\mathrm{f_i}, \mathrm{f_j})$ value for each pair of features $\mathrm{f_i}, \mathrm{f_j}$. Here for example if f-correlation score of feature id 0 and 19 =0.17 then it is denoted by $\mathrm{SU}(\mathrm{f_i}, \mathrm{f_j})$ = SU (0, 19) =0.17.

C. *MST construction.*

MST that generated on the basis of F-correlation of features and T-relevance. The correlation between any pair of features is called the F-Correlation of Feature viewing features $\mathsf{f}_\mathsf{i}, \mathsf{f}_\mathsf{j}$ as vertices. $\mathsf{SU}(\mathsf{f}_\mathsf{i}, \mathsf{f}_\mathsf{j})(\mathsf{i} \neq \mathsf{j})$ value considering as the weight of the edge between vertices$\mathsf{f}_\mathsf{i}, \mathsf{f}_\mathsf{j}$. Edge between vertices $(\mathsf{f}_\mathsf{i}, \mathsf{f}_\mathsf{j})$ will be removed if weight of edge between vertices $(\mathsf{f}_\mathsf{i}, \mathsf{f}_\mathsf{j})$ is smaller than T-relevance of both $\mathsf{f}_\mathsf{i}, \mathsf{f}_\mathsf{j}$.i.e. IF$SU(f_i, f_j)<SU(f_{i,} C)$ *and* $SU(f_{j,} C)$then $\mathrm{E}(\mathsf{f}_\mathsf{i}, \mathsf{f}_\mathsf{j})$=0. Then. Otherwise it holds edge between vertices. E.g. if SU $(\mathsf{f}_\mathsf{i}, \mathsf{f}_\mathsf{j})$= SU (0, 19) =0.17, SU (f_i, C)= SU (0, 4) =0.15, SU $(\mathsf{f}_\mathsf{j}, C)$ = SU (19, 4) = 0.14 so there will exist edge between 0 and 19. Because 0.17 > 0.15 and 0.14.

D. *Relevant feature calculation.*

After removing all the unnecessary edges, a forest is obtained. Each tree $T_j \in forest$represents a cluster that is denoted as$V(T_j)$, which is the vertex set of T_j as well. The features in each cluster are redundant, so for each cluster $V(T_j)$ we choose a representative feature F_R^j whose T-relevance $SU(F_R^j, C)$ is the greatest [15].

IV. CONCLUSION

This implementation comprises FAST clustering based feature subset selection algorithm involves three important steps: 1.Removal of irrelevant features. 2. Elimination of Redundant features using minimum spanning tree. 3. Partitioning the MST and collect the selected features. The clustering-based strategy of FAST produces a subset of useful and independent features. The FAST algorithm can efficiently and effectively deal with both irrelevant and redundant features, and obtain a good feature subset. This Implementation comprises mainly three task T-relevance calculation that produces target relevance score with respect target class, F- correlation calculation that produces correlation score of the features and MST will be constructed according to F-correlation and T-relevance score. Here in this MST generation each features can act as a vertices and correlation between features is the weight between them.

REFERENCES

1. Almuallim H. and Dietterich T.G., ‘Algorithms for Identifying Relevant Features’, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.
2. Arauzo-Azofra A., Benitez J.M. and Castro J.L., ‘A feature set measure based on relief’, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109, 2004.
3. Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., ‘On Feature Selection through Clustering’, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.
4. Biesiada J. and Duch W., ‘Features election for high-dimensional data’ła Pearson redundancy based filter, Advances in Soft Computing, 45, pp 242C249, 2008.
5. Cardie, C., ‘Using decision trees to improve case-based learning’, In Proceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.
6. Chanda P., Cho Y., Zhang A. and Ramanathan M., ‘Mining of Attribute Interactions Using Information Theoretic Metrics’, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355, 2009.
7. Chikhi S. and Benhammada S., ‘ReliefMSS: a variation on a feature ranking ReliefF algorithm.’ Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009.
8. Dash M. and Liu H., ‘Feature Selection for Classification’, Intelligent Data Analysis, 1(3), pp 131-156, 1997.
9. Demsar J., ‘Statistical comparison of classifiers over multiple data sets’, J. Mach. Learn. Res., 7, pp 1-30, 2006.
10. Dash M., Liu H. and Motoda H., ‘Consistency based feature Selection’, In Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109, 2000.
11. Hall M.A., ‘Correlation-Based Feature Subset Selection for Machine Learning’, Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.
12. Hall M.A. and Smith L.A., ‘Feature Selection for Machine Learning: Comparing a Correlation Based Filter Approach to the Wrapper’, In Proceedings of the Twelfth international Florida Artificial intelligence Research Society Conference, pp 235-239, 1999.
13. Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T., ‘Numerical recipes in C’. Cambridge University Press, Cambridge, 1988.
14. Pawan Gupta, Susheel Jain, Anurag Jain, ‘A Review of Fast Clustering-Based Feature Subset Selection Algorithm’. IJSTR volume 3, November 2014.
15. Qinbao Song, Jingjie Ni, and Guangtao Wang, ‘A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data’ IEEE Transactions on knowledge and data engineering, Vol. 25,No.1, January2013.
16. Robnik-Sikonja M. and Kononenko I., ‘Theoretical and empirical analysis of Relief and ReliefF’, Machine Learning, 53, pp 23-69, 2003.
17. Ron Kohavi and George H. John, ‘Wrappers for feature subset selection’, Artificial Intelligence 97 (1997) 273-324.

18. Yu L. and Liu H., 'Feature selection for high-dimensional data: a fast correlation-based filter solution', in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.

BIOGRAPHY

[1]**Chavan Akshay sudhakarrao,** He is aassistant professor, currently doing Master of engineering inthe computer science and engineering Department, MPGI's School of engineering, S.R.T.M.U. He received B.E. in 2013 from S.R.T.M.U, Nanded, MS, India.

[2]**Hambarde BalaprasadShankarrao,** He is a assistant professor in the computer science and engineering department, MPGI's School of engineering**.** He received M.Tech. In computer science from J.N.T.U, Hydrabad, AP, India.