Improvised: A Fast Clustering-Based Feature Subset Selection Algorithm

[1]Varsha.S.Sonwane, [2]Prof. Pratap Mohite

[1]ME(C.S.E), [2]ME (Software Engineering)
[1]ME Scholar, [2]Assistant Professor
Computer science and Engineering
Shreeyash college of Engineering & Technology,
Satara parisar,Beed by pass Road,Aurangabad-431010,
Maharashtra , India.

***Abstract*—Feature selection involves identifying a subset of the most useful features that produces well-matched 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 excellence of the subset of features. Based on these criteria, a fast clustering-based feature selection algorithm, FAST, is proposed and experimentally evaluated in this paper. The FAST algorithm works in two steps. In the first step, it involves (I) identify irrelevant features with help of four mehods 1)using Direct method 2)using cosine methos 3)using polynomial method 4)using linear method.(II)create a set of features are to be excluded (III)construct a MST (IV)obtain representative features and their weights (V)create a confusion matrix and obtain TPR and FPR.. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. Extensive experiments are carried out to compare FAST and several representative feature selection algorithms, namely, FCBF, ReliefF, CFS, Consist, and FOCUS-SF, with respect to four types of well-known classifiers, namely, the probability-based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER before and after feature selection. The results, on 35 publicly available real-world high dimensional microarray, and text data, demonstrate that FAST not only produces smaller subsets of features but also improves the performances of the four types of classifiers.**

Index Terms—Feature subset selection, feature clustering, filter method, MST, confusion matrix

INTRODUCTION

With the aim of choosing a subset of good features with respect to the goal concepts, feature subset se-lection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result clarity . 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. The embedded methods include feature selection as a part of the training process and are usually spe- cific to given learning algorithms, and therefore may be more efficient than the other three categories[11]. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of em-bedded approaches. The wrapper methods use the predictive accuracy of a prearranged learning algorithm to determine the integrity of the selected sub- sets, 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 [5].The hybrid methods are a combination of filter and wrapper methods [7] by using a filter method to shrink 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 com- plexity of the filter methods. The wrapper methods are computationally expensive and tend to overfit on small training sets [5], [7]. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, we will focus on the filter method in this paper.With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algo- rithms. Pereira et al Baker et al. [4], and Dhillon et al. [8] employed the distributional clustering of words to reduce the dimensionality of text data.

We propose a **Fast clustering-bAsed feature Selection algoriThm (FAST).**The FAST algorithm works in two steps. it involves (I) identify irrelevant features with help of four mehods 1)using Direct method 2)using cosine methos 3)using polynomial method 4)using linear method.(II)create a set of features are to be excluded (III)construct a MST (IV)obtain representative features and their weights (V)create a confusion matrix and obtain TPR and FPR. the most representative feature that is strongly related to objective 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. The proposed feature subset se- lection algorithm FAST was tested upon 35 publicly available image, microarray, and text data sets. The experimental

results show that, compared with other five different types of feature subset selection algorithms, the proposed algorithm not only reduces the number of features, but also improves the performances of the four well-known different types of classifiers. The rest of the article is organized as follows: In Section 2, we describe the Literature survey. In Section 3, we describe related works, Finally, in Section 4 , we summarize the present study and draw some conclusions.

2.LITERATURE SURVEY

Existing system:

1) RELIEF Algorithm

The Relief algorithm was first described by Kira and Rendell [KIRA92] as a simple, fast, and effective approach to attribute weighing. The output of the Relief algorithm is a weight between −1 and 1 for each attribute, with more positive weights indicating more predictive attributes. The pseudo code for Relief is shown below. The weight of an attribute is updated iteratively as follows. A sample is selected from the data, and the nearest neighboring sample that belongs to the same class (nearest hit) and the nearest neighboring sample that belongs to the contradictory class (nearest miss) are identified. A change in attribute value accompanied by a change in class leads up to weighting of the attribute based on the intuition that the attribute change could be responsible for the class change. On the other hand, a change in attribute value accompanied by no change in class leads to down weighting of the attribute based on the observation that the attribute change had no effect on the class. This procedure of updating the weight of the attribute is performed for a random set of samples in the data or for every sample in the data. The weight updates are then averaged so that the final weight is in the range [−1, 1]. The attribute weight estimated by Relief has a probabilistic interpretation. It is proportional to the difference between two conditional probabilities, namely, the probability of the attribute's value being different conditioned on the given nearest miss and nearest hit respectively .

Disadvantages of Relief:

I. Relief does not help with unnecessary features.
II. Relief is applicable to only two class classification problem.
III. Insufficient instances fools Relief.
IV. Relief requires retention of data in incremental use.

COMPUTATIONAL COMPLEXITY

For n training instances and a attributes Relief (Figure 1) makes O(m · n · a) operations. The most complex operation is selection of the nearest hi and miss as we have to compute the distances between R and all the other instances which takes O(n · a) comparisons.

2) Relief-F

It is a feature selection strategy that chooses instances arbitrarily. It cannot identify unnecessary features. ReliefF is a simple yet efficient procedure to estimate the quality of attributes in problems with strong dependencies between attributes. In practice, ReliefF is usually applied in data pre-processing as a feature subset selection method. The ReliefF (Relief-F) algorithm (Kononenko, 1994) is not limited to two class problems, is more robust and can deal with incomplete and noisy data. Similarly to Relief, ReliefF randomly selects an instance Ri , but then searches for k of its nearest neighbors from the same class, called nearest hits Hj, and also k nearest neighbors from each of the different classes, called nearest misses Mj(C) . It updates the quality estimation W[A] for all attributes A depending on their values for Ri , hits Hj and misses Mj(C) . The update formula is similar to that of Relief, except that we average the contribution of all the hits and all the misses. The contribution for each class of the misses is weighted with the prior probability of that class P(C) (estimated from the training set). Since we want the contributions of hits and misses in each step to be in [0,1] and also symmetric (we explain reasons for that below) we have to ensure that misses' probability weights sum to 1. As the class of hits is missing in the sum we have to divide each probability weight with factor 1−P(class(Ri)) (which represents the sum of probabilities for the misses' classes). The process is repeated for m times. Selection of k hits and misses is the basic difference to Relief and ensures greater robustness of the algorithm concerning noise. User-defined parameter k controls the locality of the estimates. For most purposes it can be safely set to 10 (see (Kononenko, 1994) and discussion below). To deal with incomplete data we change the diff function. Missing values of attributes are treated probabilistically. We calculate the probability that two given instances have different values for given attribute conditioned over class value:

− if one instance (e.g., I1) has unknown value:

$$diff(A,I1,I2) = 1-P(value(A,I2)|class(I1)) \quad (1)$$

− if both instances have unknown value:

$$diff(A,I1,I2) = 1- \#values(A) \sum V ¡ P(V|class(I1))\times P(V|class(I2))¢ \quad (2)$$

Conditional probabilities are approximated with relative frequencies from the training set.

COMPUTATIONAL COMPLEXITY

Although ReliefF look more complicated their asymptotical complexity is the same as that of original Relief, i.e., O(m · n · a). The most complex operation within the main for loop is selection of k nearest instances. For it we have to compute distances from

all the instances to R, which can be done in O(n · a) steps for n instances. This is the most complex operation, since O(n) is needed to build a heap, from which k nearest instances are extracted in O(k logn) steps, but this is less than O(n · a).

3) Hierarchical clustering:

Hierarchical clustering algorithm is of two types:
i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and
ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).
Both this algorithm are exactly reverse of each other. So we will be covering Agglomerative Hierarchical clustering algorithm in detail.Agglomerative Hierarchical clustering -This algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed? For this there are many available methods.

Disadvantages:

1) Hierarchical clustering Algorithm can never undo what was done previously.
2) Hierarchical clustering Time complexity of at least O($n^2 \log n$) is required, where *'n'* is the number of data points.
4) No objective function is directly minimized in Hierarchical clustering
5) Sometimes it is difficult to identify the correct number of clusters by the dendogram in Hierarchical clustering.
6)The generality of the selected features is limited
7)The computational complexity is large.
8)Accuracy is not guaranteed.
9)Ineffective at removing redundant features

4. Filter Methods:

These methods select features based on cultivated criteria that are relatively independent of classification.
Several methods use simple correlation coefficients similar to Fisher's discriminant criterion. Others adopt mutual information or statistical tests (t-test, F-test). Earlier filter-based methods evaluated features in isolation
and did not consider correlation between features. Recently, methods have been proposed to select features with
minimum redundancy. The methods proposed use a minimum redundancy-maximum relevance (MRMR) feature selection framework. They supplement the maximum relevance criteria along with minimum redundancy criteria to choose additional features that are maximally dissimilar to already identified ones. By doing this, MRMR expands the representative power of the feature set and improves their generalization properties.

5. **Wrapper Methods:**

Wrapper methods utilize the classifier as a black box to score the subsets of features based on their predictive
power. Wrapper methods based on SVM have been widely studied in machine-learning community. SVM-RFE
(Support Vector Machine Recursive Feature Elimination), a wrapper method applied to cancer research is called, uses a backward feature elimination scheme to recursively remove insignificant features from subsets of
features. In each recursive step, it ranks the features based on the amount of reduction in the objective function.
It then eliminates the bottom ranked feature from the results. A number of variants also use the same backward
feature elimination scheme and linear kernel.

3.PROPOSED SYSTEM

Feature subset selection can be viewed as the process of identifying and removing as many unrelated and unnecessary features as possible. This is because: (i) irrelevant features do not contribute to the predictive accuracy [14], and (ii) 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 [10], [12].

3.1.FEATURE SUBSET SELECTION ALGORITHM

3.1.1 Framework and Definatons : Irrelevant features, along with redundant features, severely affect the correctness of the learning machines [12]. Thus, feature subset selection should be able to identify and remove as much of the irrelevant and re- relevant to the target concept; Moreover, "*good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.*"

Keeping these in mind, we develop a novel algorithm which can powerfully and successfully deal with both unrelated and unneeded features, and obtain a good feature subset. We achieve this through a new feature selection framework (shown in Fig.1) which composed of the two connected components of *irrelevant feature removal* and *redundant feature elimination*. The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset.

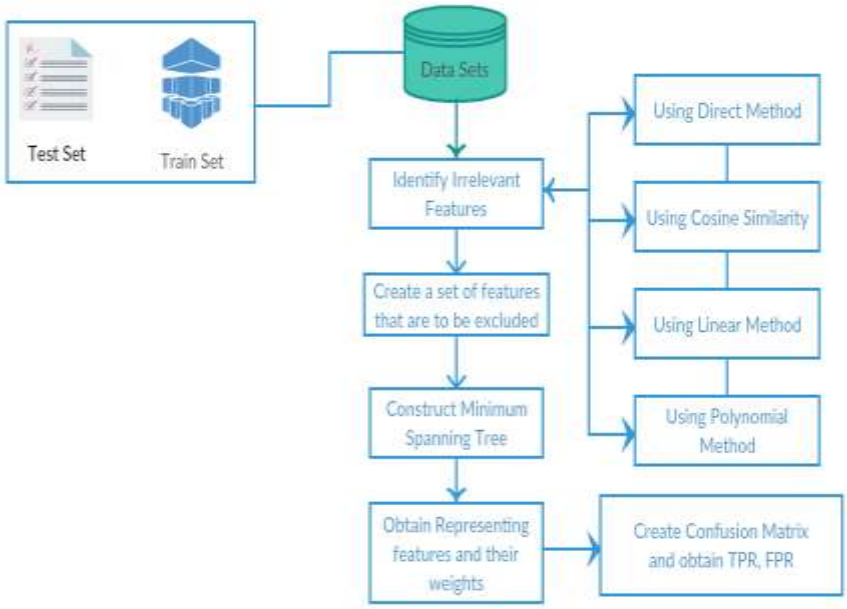

Fig. 1: Framework of the proposed feature subset selection algorithm

The former obtains features relevant to the target concept by eliminating irrelevant ones, and the latter removes redundant features from relevant ones via choosing representatives from different feature clusters, and thus produces the final subset. 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 (I) identify irrelevant features with help of four mehods 1)using Direct method 2)using cosine methods 3)using polynomial method 4)using linear method.(II)create a set of features are to be excluded (III)construct a MST (IV)obtain representative features and their weights (V)create a confusion matrix and obtain TPR and FPR. features In order to more precisely introduce the algorithm, and because our proposed feature subset selection framework involves irrelevant feature removal and redundant feature elimination, we firstly present the traditional definitions of relevant and redundant features, then provide our definitions based on variable correlation as follows.

John et al. [14] presented a definition of relevant features. Suppose F to be the full set of features, $Fi \in$ be a feature, $Si = F - \{Fi\}$ and $S'i \subseteq Si$. Let $s'i$ be a value assignment of all features in S', fi a value-assignment of feature Fi, and c a value-assignment of the target concept C. The definition can be formalized as follows.

The definition can be formalized as follows.

Definition *1: (Relevant feature)* Fi is relevant to the target concept C if and only if there exists some $s'i$, fi and c, such that, for probability $p(S'i = si, Fi = fi) > 0, p(C = c \mid S'i = s'i, Fi = fi) \neq p(C = c \mid S'i = s'i)$.Otherwise, feature Fi is an *irrelevant feature*.Definition 1 indicates that there are two kinds of relevant features due to different $S'i$: (i) when $S'i = Si$,from the definition we can know that Fi is directly relevant to the target concept; (ii) when $S'i \subsetneq Si$, from the definition we may obtain that $p(C \mid Si, Fi) = p(C \mid Si)$.

Definition *2: (Redundant feature)* Let S be a set of features, a feature in S is redundant if and only if it has a minimum cosine simmilarity within S. The *symmetric uncertainty* is defined as follows

$(X, Y) = 2 \times (X \mid)/H(X) + H(Y)$. (1)Where,1) (X) is the entropy of a discrete random variable X. Suppose (x) is the prior probabilities for all values of X, (X) is defined by $(X) = -\Sigma(x) \log 2\ p(x)$. (2)

$x \in X$

2) $Gain(X/Y)$ 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 [55] which is given by

$Gain(X \mid Y) = H(X) - H(X \mid Y)$

$= H(Y) - H(Y \mid X)$.

Where $(X/)$ 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. Suppose (x) is the prior probabilities for all values of X and (x/y) is the posterior probabilities of X given the values of

$(X/)$ is defined by

$(X \mid Y) = -\Sigma(y)\ \Sigma p(x \mid y) \log 2\ p(x \mid y)$.

$y \in Y \quad x \in X$

Definition 3*: (T-Relevance)* The relevance between the feature $Fi \in F$ and the target concept C is referred to as the *T-Relevance* of Fi and C, and denoted by (Fi, C). If (Fi, C) is greater than a predetermined threshold θ, we say that Fi is a strong *T-Relevance* feature.

Definition 4*: (F-Correlation)* The correlation between any pair of features Fi and Fj $(Fi, Fj \in F \wedge i \neq j)$ is called the *F-Correlation* of Fi and Fj , and denoted by $SU(Fi, Fj)$.

Definition 5: *(F-Redundancy)* Let $\mathcal{S} = \{F1, F2, ..., Fi,..., Fk<|F|\}$ be a cluster of features.if $\exists Fj \in \mathcal{S}, SU(Fj, C) \geq SU(Fi, C) \wedge SU(Fi, Fj) > SU(Fi, C)$ is always corrected for each $Fi \in \mathcal{S}$ $(i \neq j)$, then Fi are redundant features with respect to the given Fj (i.e. each Fi is a *F-Redundancy*).

Definition 6*: (R-Feature)* A feature $Fi \in \mathcal{S} = \{F1, F2, ..., Fk\}$ $(k < |F|)$ is a representative feature of the cluster $\mathcal{S}$ (i.e. Fi is a *R-Feature*) if and only if, $Fi = \text{argmax}_{Fj \in \mathcal{S}} SU(Fj, C)$. This means the feature, which has the strongest *TRelevance*, can act as a *R-Feature* for all the features in the cluster.

3.1.3.RESULT: we test three dataset for FAST Algorithm

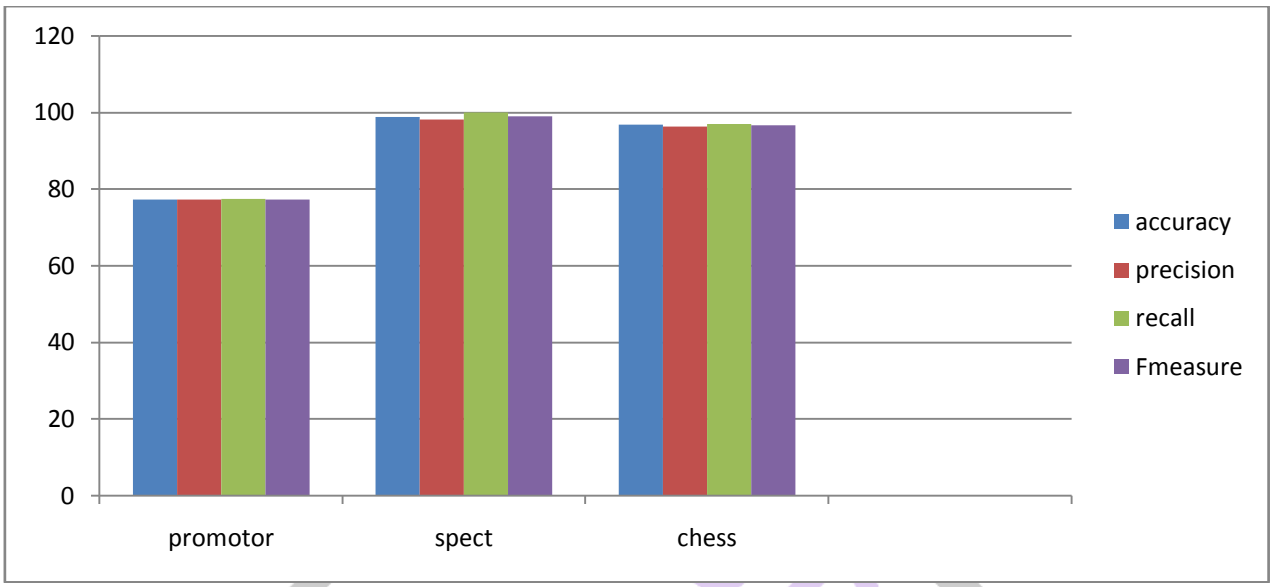

Fig.2.result of dtaaset

Algorithm	*Advantage*	*Disadvantage*
Wrapper Approach	High Accuracy	Large computational complexity
Filter Approach	Suitable for very large features	Accuracy is not guaranteed
Relief Algorithm	Improve efficiency, Reduces cost	Powerless to detect redundant features
FAST Algorithm	Efficient, Effective	Takes more time

Table 1: Comparison of different feature section methods

4. SIGNIFICAN FEATURES:

- It reduces the dimensionality of the feature space, to limit storage requirements and increase algorithm speed;
- It removes the redundant, irrelevant or noisy data.
- The immediate effects for data analysis tasks are speeding up the running time of the learning algorithms.
- Improving the data quality.
- Increasing the accuracy of the resulting model.
- Feature set reduction, to save resources in the next round of data collection or during utilization;
- Performance improvement, to gain in predictive accuracy;
- Data understanding, to gain knowledge about the process that generated the data or simply visualize the data

4. CONCLUSION

Obtaining a suitable rank as to where the FAST algorithm exactly stands amongst few other existents. In this, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data.In the proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. The FAST algorithm works in two steps. In the first step, it involves (I) identify irrelevant features with help of four mehods 1)using Direct method 2)using cosine methos 3)using polynomial method 4)using linear method.(II)create a set of features are to be excluded (III)construct a MST (IV)obtain representative features and their weights (V)create a confusion matrix and obtain TPR and FPR..For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space. In feature we are going to classify the high dimensional data. In this paper, we have presented a novel clustering-based. Each cluster is treated

as a single feature and thus dimensionality is drastically reduced. Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime. For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space.

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] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279- 305, 1994.
[3] 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.
[4] 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.
[5] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.
[6] 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.
[7] Das S., Filters, wrappers and a boosting-based hybrid for feature Selection,In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.
.[8] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res., 3, pp 1265-1287, 2003.
[9]Dougherty, E. R., Small sample issues for microarray-based classification. Comparative and Functional Genomics, 2(1), pp 28-34, 2001.
[10]Forman G., An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 3, pp 1289-1305,2003.
[11] Guyon I. and Elisseeff A., An introduction to variable and feature selection, Journal of Machine Learning Research, 3, pp 1157-1182, 2003. Learning, Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.Learning Research, 3, pp 1157-1182, 2003. [29] Hall M.A., Correlation-Based Feature Subset Selection for Machine Learning, Ph.D. dissertation Waikato, New Zealand: Univ. Waikato, 1999.
[12]Hall M.A., Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, In Proceedings of 17th International Conference on Machine Learning, pp 359-366, 2000.
[13] Jaromczyk J.W. and Toussaint G.T., Relative Neighborhood Graphs and Their Relatives, In Proceedings of the IEEE, 80, pp 1502-1517, 1992.
[14]John G.H., Kohavi R. and Pfleger K., Irrelevant Features and the Subset Selection Problem, In the Proceedings of the Eleventh International Conference on Machine Learning, pp 121-129, 1994.
[15] Kohavi R. and John G.H., Wrappers for feature subset selection, Artif. Intell., 97(1-2), pp 273-324, 997.

Varsha S. Sonwane, I have completed BE in Computer Science and Engineering in 2011 from MSS Engineering College, JALNA.Three years I worked as a Assistant Professor in Jawaharlal P.E.S. Engineering College, Aurangabad (2012-2015) I am doing ME in Computer Science and Engineering from Shreeyash College of Engineering and Technology, Aurangabad.I also worked as a lecturer in S.B.N.M.Polytechnic,Aurangabad.