ON FEATURE SELECTION ALGORITHMS AND FEATURE SELECTION STABILITY MEASURES: A COMPARATIVE ANALYSIS

Mohana Chelvan P[1] and Dr. Perumal K[2]

[1]Dept. of Computer Science, Hindustan College of Arts and Science,Chennai – 603 103, India,
[2]Dept. of Computer Applications, Madurai Kamaraj University, Madurai – 625 021, India,

ABSTRACT

Data mining is indispensable for business organizations for extracting useful information from the huge volume of stored data which can be used in managerial decision making to survive in the competition. Due to the day-to-day advancements in information and communication technology, these data collected from e-commerce and e-governance are mostly high dimensional. Data mining prefers small datasets than high dimensional datasets. Feature selection is an important dimensionality reduction technique. The subsets selected in subsequent iterations by feature selection should be same or similar even in case of small perturbations of the dataset and is called as selection stability. It is recently becomes important topic of research community. The selection stability has been measured by various measures. This paper analyses the selection of the suitable search method and stability measure for the feature selection algorithms and also the influence of the characteristics of the dataset as the choice of the best approach is highly problem dependent.

KEYWORDS

Data Mining, Feature Selection, Feature Selection Algorithms, Selection Stability, Stability Measures.

1. INTRODUCTION

Data mining is essential for getting useful information from huge amount of data stored due to the day-to-day activities of the organizations. These data are mostly high dimensional which makes the data mining task difficult. Feature selection is a scheme which chooses small related feature subsets from the dataset. Feature selection improves accuracy, efficiency and model interpretability of the algorithms. The subsets get by feature selection on the same sample should be similar in subsequent iterations and should be stable even for small perturbations or the addition of new data. Feature selection stability is the robustness of feature selection algorithms for small perturbations in the dataset. Otherwise it will create confusion in researcher's mind about the result and lowers their confidence in their conclusion of research work [1]. Recently selection stability becomes hot topic of research. This paper gives an account of various selection stability measures and their application in various feature selection algorithms.

2. FEATURE SELECTION ALGORITHMS

The feature selection process is mostly based on three approaches viz. filter, wrapper and hybrid [2]. The filter approach of feature selection is by removing features on some measures or criteria and the feature's wellness is examined using intrinsic or statistical features of the dataset. A

DOI:10.5121/ijcsit.2017.93014

feature is referred as a much suited feature based on these properties, and is chosen for machine learning or data mining uses. In the wrapper approach the subset of features is produced and then goodness of the subset is examined with the use of a classifier. The purpose of some classifier here is ranking the features of the dataset depending on which an option is chosen for the desired use. The embedded model combines the advantages of both the above models. The hybrid approach takes benefits of both the approaches by using the various examination criterions of them in various search stages. The review on feature selection algorithms is presented in [3] and the important feature selection algorithms are shown below.

2.1. One-R

One-R algorithm is put-forth by Holte [4] and is simple. The algorithm has each rule for every aspect in the training data and fixes the rule with minimum error. This algorithm considers the mathematically charged features as continuous. This is one of the most primitive techniques. It just separates the series of values to many dis-joint intervals also it is a straightforward method. It treats missing values as a legitimate value called “missing”. Here simple rules are produced depending on one feature only. It can be helpful to determine a standard performance as a target for more learning techniques even though it is a minimal form of classifier.

2.2. Information Gain (IG)

The entropy is a criterion of impurity in a training set *S*. It is defined as a means which provide extra data about *Y* presented by *X* that shows the value by which the entropy of *Y* falls [5]. This scheme is termed as IG which is symmetrical in nature and is given in (1).

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \quad (1)$$

Knowledge obtained regarding *Y* followed by the observation of *X* equals the knowledge obtained regarding *X* followed by observation of *Y*. This is oriented in support of features with more values though it may be less informative and is the weakness of the IG criterion. It computes the worth of an attribute by considering the information gain based on the class as in (2). By considering the variation between the entropy of the feature and the conditional entropy provided the class label, the IG decides the independence within a feature and the class label.

$$IG\ (Class, Attribute) = H\ (Class) - H\ (Class \mid Attribute) \quad (2)$$

2.3. Gain Ratio (GR)

The GR is a non-symmetrical measure. It is coined to make compensation for the IG bias [5]. GR is contributed in (3).

$$GR = \frac{IG}{H(X)} \quad (3)$$

IG is normalized by dividing using the entropy of X, and vice versa when the variable Y has to be determined. The range of GR values lie within the span of [0, 1] because of the normalization. A value of GR = 0 represents the absence of relation among Y and X while GR = 1 shows that the information about X easily predicts Y. The GR favors variables with less values as opposed to IG.

2.4. Symmetrical Uncertainty (SU)

The Symmetric Uncertainty is given in (4). By dividing IG by the sum of the entropies of X and Y, the Symmetrical Uncertainty criterion compensates for the inherent IG bias [5].

$$SU = 2 \frac{IG}{H(Y) + H(X)} \quad (4)$$

The values taken by SU are normalized to the range [0, 1] due to the correction factor 2. A value of SU = 0 indicates that X and Y are uncorrelated while the value of SU = 1 predicts that information about one feature generally depicts the other. The SU is biased toward features with fewer values as GR.

2.5. Correlation-based Feature Selection (CFS)

The attribute subsets value is evaluated by CFS by considering the degree of redundancy between them together with the distinct predictive capability of each feature. Subsets of features which are greatly correlated with the class with less inter-correlation are given the preference [6]. CFS determines the optimal feature subset and can be brought together with other search mechanisms. Authors have GA as search method with CFS as fitness function. CFS is given in (5).

$$r_{zc} = \frac{k\, r_{zi}}{\sqrt{k + (k-1)\, r_{ii}}} \quad (5)$$

where r_{zc} gives the correlation within the added feature subsets and the class variable, k refers to the number of subset features, r_{zi} is the mean correlations within the subset features and the class variable, and r_{ii} is the mean inter-correlation within subset features [6].

2.6. ReliefF

To evaluate the attributes by assigning a weight to every feature, the feature capacity to differentiate within the classes has been used. Using sampling repetition to sample an occurrence, with the value of particular feature considered for the adjacent occurrence of similar and various classes, the ReliefF attribute contribution [7] examines the value of feature. Relevant features are then selected based on which whose weights cross a user meaned threshold. The weight determination depends on the possibility of two adjacent neighbors belonging to the same class with similar value of the feature and the possibility of the adjacent neighbors from two various classes with various values for a feature. The feature is much important during occurrence of much variation within these probabilities. The measure is described for a two-class problem. However, by parting the challenge to a two class problems series, it could also be expanded to handle various classes. ReliefF trial to increase the margin where it is connected to hypothesis margin maximization. Relief offers a good trade-off between complexity and performance.

2.7. ChiSquare

Like Information Gain, ChiSquare computes the independency among a feature and a class-label using variation within the feature's entropy and the Conditional entropy provide the class-label. ChiSquare analyses if a specific nature does not depend on the class-label [8]. The 2 way Chi-squared test is an analytical technique which identifies closeness of the expected and the actual outcomes. This technique makes an assumption that the parameters are chosen randomly and obtained from a suitable sample of independent values. The resultant Chi-squared values show the deviation of results from the random (expected) outcome. This technique computes the absence of independency within a term and the category. Chi-Squared is the general analytical exam which computes deviation from the distribution anticipated if someone assume the feature happening does not depend on the class value. Being an analytical exam, it shows error values for few small anticipated counts that are normal in classification of test due to presence of rare happening word features and sometimes due to possession of small positive training instances for a theme. Here, the $\chi2$ test is used to examine the independency of both events, where the two events A and B are said to be not dependent only if the condition P(AB) = P(A)P(B) occurs or, its equivalent, P(A|B) = P(A) and P(B|A) = P(B). In feature selection, the two events are term and class occurrence. Feature selection by the $\chi2$ statistic is similar like doing a hypothesis exam on the class distribution because it connects to the feature query values.

The Null Hypothesis shows absence of co-relation; every value can have instance in a particular class like other classes. Consider on Null Hypothesis, if p instances has a particular value and q instances occur in a particular group, (p·q)/n instances have a particular value and are in a particular class. This is since p/n instances with the value and q/n instances are in the class, and if probabilities don't depend, then (i.e. the null hypothesis) their total probability is given by their multiple. Provided the Null Hypothesis, the $\chi2$ statistic computes deviation of actual value compared to the expected one.

$$x^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{6}$$

Here in (6), 'r' is the number of various values of the feature in examination, 'c' gives the number of classes in examination (in this work, c = 2), O i,j is the number of instances having value i present in class j, and Ei,j is the anticipated instances number having value I and class j, depending upon (p·q)/n. The greater this Chi-squared statistic, the more un-likely the independency the values and classes distribution; i.e., if they are connected, the feature in examination favours the class.

2.8. Fisher

The Fisher score is sometimes coined as the information score, as this shows the quantity of knowledge which is provided by a variable about a unknown variable on whom it is based [9]. Computation of the score is done by computing the deviation of the knowledge of observed one from the expected one. When variance is reduced, information is increased. As the score expectation is zero, the Fisher information also gives the score variance. Fisher score gives larger scores to the feature which discriminates different samples from different classes easily.

3. Selection Stability Measures

The stability measures have been divided to three main categories depending on the representation of the output of the selection method [10]. The categories are stability by index, stability by rank and stability by weight. Let A and B are features subset, A, B ⊆ X, of the similar dimension i.e., cardinality, k. Let r = |A ∩ B| be the cardinality of the two subset intersections. The desirable properties that each stability measurement should have [11] are as follows:

- **Monotonicity.** For a fixed number of features, n and subset size, k, the higher the intersection within the subsets, the larger the stability index value.
- **Limits.** The index must be destined by constants that don't rely on n or k. When the two subsets are identical, maximum value should be attained, i.e., for r = k.
- **Correction for chance.** The index must be a perpetual value to draw independent feature subsets of similar cardinality, k.

In addition to these requirements, the following important properties should be taken into consideration due to their impact on the selection stability result [1] [12].

- The size of the dataset.
- The amount of chosen features.
- The sample size.
- The data variance.
- The symmetry of the measurement.

4. CATEGORIES OF FEATURE SELECTION STABILITY MEASURES

Stability can be assessed by the pairwise comparison between the resulting subsets obtained by feature selection algorithm on datasets. The stability is higher if the similarity between the resulting subsets is greater. Based on the output of the feature selection technique, the stability measures are of three different representations i.e., indexing, ranking, and weighting [10].

4.1. Stability by Index

In this category of measurements, the selected subset of features is signified as a binary vector with cardinality equivalent to the total features m or as a vector of indices relating to the selected features k. Unlike the other stability measurements i.e., rank or weight based measurements, the index measurements have the possibility for handling subset of feature, i.e., the number of selected features $k \leq m$. The index measurements assess the amount of overlap between the resulting subsets of features for assessing the stability. The examples for stability by index measurement are Dice's Coefficient, Tanimoto Distance, Jaccard Index and Kuncheva Index.

4.2. Stability by Rank

The stability by rank method assesses the stability by evaluating the correlation between the ranking vectors. Unlike the index method, these methods do not deal with partial set of features as they cannot tackle vectors with multiple cardinality i.e., vectors that resemble to various features set. The measurements in this category include Spearman's Rank Correlation Coefficient SRCC.

4.3. Stability by Weight

Similar to the stability by rank, this category of measurement deals with only full subset of features. This method assesses selection stability by evaluating the weight of the full feature set. The stability by weight category of measurement has only one member called the Pearson's Correlation Coefficient PCC. Here the stability is assessed by evaluating the correlation between the two sets of weights w_i and w_j for the whole feature dataset.

5. IMPORTANT FEATURE SELECTION STABILITY MEASURES

5.1 Dice's Coefficient

Dice, Tanimoto and Jaccard are similar index based stability measures. Dice coefficient calculates selection stability by computing the overlap among two subsets of features as in (7) and is used in [5].

$$\text{Dice}(F'_1, F'_2) = \frac{2\left|F'_1 \cap F'_2\right|}{\left|F'_1\right| + \left|F'_2\right|} \tag{7}$$

Dice bounds between the values of 0 and 1, where 0 means the results are unstable i.e., no overlap between the subset of features and 1 means the two subsets are stable or identical.

5.2 Jaccard Index (JI)

The given different results R = {R_1, R_2... R_l} will correspond to l different folds of the sample dataset. By evaluating the overlapping instances between the subsets in R, the stability can be evaluated as in (8). By estimating the number of overlap among the features of the chosen subsets, the JI is to assess the stability for feature subset with selected feature indices [13]. The similarity between finite numbers of subsets is measured by the Jaccard coefficient measures. JI is measured as the intersection dimension of the selected feature subsets divided by the size of their union. JI for two selected subsets is shown by (8) and for a number of subsets in subsequent iterations is shown by (9).

$$S_J(R_i, R_j) = \frac{|R_i \cap R_j|}{|R_i \cup R_j|} \tag{8}$$

$$S_J(R) = \frac{2}{l(l-1)} \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} S_J(R_i, R_j) \tag{9}$$

The Jaccard Index S_J gives a value which bounds within the interval [0, 1] where 0 defines the two subsets R_i and R_j of feature selection consequences are not steady and overlapped and 1 defines the results are very steady and identical.

5.3 Kuncheva Index (KI)

The stability index is an index based measure based on correction for chance and cardinality of the intersection. KI obey all the requirement appeared in [11] and is the only measurement as such.

$$KI(F'_1, F'2) = \frac{|F'_1 \cap F'_2| \,.\, n - k^2}{k\,(n - k)} \tag{10}$$

In (10), k is the total features of the subset and n is the total features of the dataset. KI's results ranges in [–1, 1], where –1 means there is no intersection between the lists and k = n/2. KI achieves 1 when F'_1 and F'_2 are same to define the intersection set cardinality equals k. KI values becomes nearer to zero for drown individually lists. KI becomes desirable measure because of the correction for chance term which was introduced in [11]. Unlike other measurements, larger value of cardinality will not affect the stability in KI. In case of the other stability measures, the larger the cardinality is, the higher the stability will be.

5.4 Spearman's Rank Correlation Coefficient (SRCC)

SRCC is rank based stability measure and is shown in (11) to evaluate the stability of two ranked sets of features' r and r'. It is introduced by A. Kalousis et al. in [10].

$$SRCC(r, r') = 1 - 6 \sum_{t} \frac{(r_t - r'_t)^2}{m\,(m^2 - 1)} \tag{11}$$

The result of Spearman's will be in the range of [–1, 1]. The maximum will be achieved when the two ranks are identical while the minimum is when they are exactly in inverse order and 0 mean no correlation at all between r and r′.

5.5 Pearson's Correlation Coefficient (PCC)

Pearson's is weight based stability measure and is used to measure the correlation between the weights of the features that returned from more than one run and is adapted in [10]. The Pearson's Correlation Coefficient PCC stability will be as in (12).

$$\mathrm{PCC}\,(w, w') = \frac{\sum_i (w_i - \mu_w)\,(w'_I - \mu_{w'})}{\sqrt{\sum_i (w_i - \mu_w)^2 \sum_i (w'_I - \mu_{w'})^2}} \tag{12}$$

Here μ is the mean. PCC takes values between –1 and 1, where 1 means the weight vectors are perfectly correlated, –1 depicts they are anti correlated while 0 means no correlation. The stability will be shown higher when the weight is equal to zero for big number of features. However, this will not be an issue in situations as the algorithm assigns weight between 1 and −1. The PCC is a symmetric measure. It is the only stability measure that handles feature weights.

6. Experimental Results

The two datasets used in the experiments are Census-Income (KDD) dataset and Insurance Company Benchmark (COIL 2000) dataset. The datasets are obtained from the KEEL dataset repository [14]. The dataset characteristics are shown in Table 1.

Table 1. Characteristics of the datasets census and coil 2000.

S. No.	Dataset Characteristics	Dataset	
		Census	Coil 2000
1	Type	Classification	Classification
2	Origin	Real World	Real World
3	Instances	142521	9822
4	Features	41	85
5	Classes	3	2
6	Missing Values	Yes	No
7	Attribute Type	Categorical, Numerical	Numerical

Filter model is mostly stable, very efficient and highly generalizable because it is independent of any classifier [7]. Due to these advantages, most of the analysis in this paper belongs to this model. However it might not be accurate as wrapper model. Well-known Filter algorithms used in the experiment include One-R [9], Information Gain [5], Gain Ratio [5], Symmetric Uncertainty [5], CFS [10], ReliefF [11], ChiSquare [12] and Fisher [13]. All algorithms are filter-based; hence classifier is not included in any selection process. The search methods include BestFirst and Ranker.

The BestFirst method examines the attribute space subsets using greedy hillclimbing augmented with a backtracking capability. The BestFirst method may search backward after beginning with the full set of attributes or search forward after beginning with the empty set of attributes or it may search in both directions by making use of all suitable single attribute additions and deletions at a specific point after starting at any point. The search method BestFrst is well suited for CFS, ChiSquare and Fisher. The Ranker method ranks attributes with particular estimations. It will be used in combination with attribute evaluators One-R, Information Gain, Gain Ratio, Symmetric Uncertainty and ReliefF. The table 2 summarizes the result of the experiments.

In the experiments, dataset census is mostly categorical while the dataset coil 2000 is completely numerical. Some feature selection algorithms are better suited for the dataset census while others are better suited for the dataset coil 2000. The dataset census is consistent for all feature selection algorithms except CFS. But the results with the dataset coil 2000 have some fluctuations. The algorithms ReliefF, Information Gain and Symmetric Uncertainty give good results with coil 2000 dataset while CFS gives the worst result.

S. No.	Feature Selection Algorithm	Search Method	Stability Measure	Bounds	Dataset Census	Dataset Coil 2000
1	One-R	Ranker	Pearson's	[−1,1]	0.966768	0.644042
2	Information Gain	Ranker	Pearson's	[−1,1]	0.998981	0.976296
3	Gain Ratio	Ranker	Spearman's	[−1,1]	0.988328	0.701720
4	Symm. Uncert.	Ranker	Spearman's	[−1,1]	0.986760	0.817628
5	CFS	BestFirst	Kuncheva	[−1,1]	0.275675	0.043290
6	ReliefF	Ranker	Spearman's	[−1,1]	0.996516	0.983154
8	ChiSquare	BestFirst	Dice's	[0,1]	0.666666	0.727272
9	Fisher	BestFirst	Jaccard	[0,1]	0.518272	0.627272

Table 2. Selection stability results for feature selection algorithms for the datasets census and coil 2000.

The index based selection stability measures i.e., Dice's Coefficient, Jaccard Index and Kuncheva Index can work with partial set of features while the rank based and weight based selection stability measures i.e., SRCC and PCC can work with full set of features. The feature selection stability measure value will be improved up to the optimum number of selected features and then decreases and so the number of selected features will be kept at optimum number for index based measures and is not possible for rank based and weight based measures. The Dice's coefficient has slightly higher values of stability measure due to larger value of cardinality of intersection of selected features i.e., for Fisher in comparing with ChiSquare and CFS. However, the minimum value of stability measure for Kuncheva Index i.e., for CFS is due to the correction of chance term. In the case of Kuncheva Index, the larger value of cardinality will not affect selection stability in comparison with other stability measures. Fig. 1 gives the comparison chart of selection stability outcomes for the feature selection algorithms for the experimental datasets.

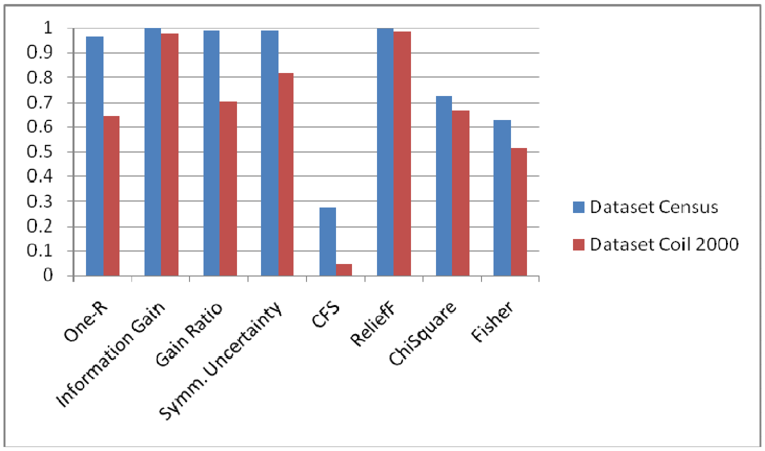

Figure 1. Comparison of selection stability results for the feature selection algorithms for datasets census and coil 2000.

7. CONCLUSION

Feature selection can play a major role in data mining and it has been identified as a challenging research problem for the academicians, industrialists and researchers. Researchers of feature selection must understand the underlying sampling techniques because the dataset distribution is generally un-known. There is really no "best" method for feature selection because different data sets have different measures of correlation. Even when the fold is sampled from the similar dataset, with no overlap with other folds, may tend with different related features. The behaviors of the feature selection algorithm will vary on different datasets with different characteristics. In addition, selection of suitable stability measure for the feature selection algorithm is also an interesting research problem.

REFERENCES

[1] Salem Alelyani, Huan Liu, “The Effect of the Characteristics of the Dataset on the Selection Stability”, 1082-3409/11, IEEE DOI 10.1109/ International Conference on Tools with Artificial Intelligence.2011.167, 2011.

[2] K. Mani, P. Kalpana, “A review on filter based feature selection”, International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE) ISSN: 2320-9801, Vol. 4, Issue 5, May 2016

[3] K. Sudha, J. JebamalarTamilselvi, "A Review of Feature Selection Algorithms for Data Mining Techniques", International Journal on Computer Science and Engineering (IJCSE) ISSN: 0975-3397, Vol. 7, No.6, pp. 63-67, June 2015.

[4] Holte, R.C., “Very simple classification rules perform well on most commonly used datasets”, Machine Learning, 11: 63-91, 1993.

[5] Hall, M.A., and Smith, L. A., “Practical feature subset selection for machine learning”, Proceedings of the 21st Australian Computer Science Conference, 181–191, 1998.

[6] Mark A. Hall, ”Correlation-based Feature Selection for Machine Learning”, Dept. of Computer science, University of Waikato. http://www.cs.waikato.ac.nz/ mhall / thesis.pdf, 1998.

[7] Marko, R.S., and Igor, K., “Theoretical and empirical analysis of relief and reliefF”, Machine Learning Journal, doi: 10.1023/A: 1025667309714, 53: 23–69, 2003.

[8] I. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann Pub, 2005.

[9] R. Duda, P. Hart, and D. Stork. Pattern Classification. John Wiley & Sons, New York, 2 edition, 2001.

[10] A. Kalousis, J. Prados, and M. Hilario, “Stability of feature selection algorithms: a study on high-dimensional spaces”, Knowledge and Information Systems, 12(1):95 – 116, May 2007.

[11] L. I. Kuncheva, "A stability index for feature selection", In Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi Conference: artificial intelligence and applications, Anaheim, CA, USA,. ACTA Press, 390 – 395, 2007.

[12] Salem Alelyani, Zheng Zhao, Huan Liu, "A Dilemma in Assessing Stability of Feature Selection Algorithms", 978-0-7695-4538-7/11, IEEE DOI 10.1109/ International Conference on High Performance Computing and Communications. 2011.99, 2011.

[13] Y. Saeys, T. Abeel, and Y. Van de Peer, Robust feature selection using ensemble feature selection techniques, 2008.

[14] J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García,L.Sanchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework", J. Multiple-Valued Logic Soft Comput., 17(2): 255–287, 2010.

AUTHORS

P. Mohana Chelvan working as an Assistant Professor in Department of Computer Science at Hindustan College of Arts and Science, Chennai, India since 2015. His educational qualifications are MCA, NIELIT C Level (IT), MPhil. (CS) and UGC NET. He is currently Ph.D. rese arch scholar in computer science from Madurai Kamaraj University in the area of Privacy preserving data mining.

Dr. K. Perumal working as an Associate Professor in Department of Computer Applications at Madurai Kamaraj University, Madurai, India since 1990. He awarded his Ph.D. degree in computer science from Madurai Kamaraj University in the area of Digital image processing. He has contri buted more than 40 papers in the International Journals and Conferences and also editor of proceedings for National Conference on Contemporary Developments in Information and communication technologies. He has guiding 9 scholars. His research interest includes Data Mining, Big Data and image processing especially in medical image processing.