# Feature Selection Approach Using Symmetrical Uncertainty For Improving Classification Accuracy On Breast Cancer Dataset

G.Vijay Kumar[1], M.Sreedevi[2]
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Green Fields, Vaddeswaram, Guntur(dt), Andhra Pradesh

**Abstract**

Large data is generating with more number of dimensions due to technology advancement in the field of medical databanks. Thousands of attributes are existing for every medical record for better diagnosis of disease. Classification as well as prediction of large dimensional data is a challenging task in front of researchers. Selection of required attribute and reduction of dimensionality is one of the perfect determinations to get better accuracy results of classifying the data as well as prediction. Hence, here we proposed attribute selection/chosen method based on Symmetrical Uncertainty (SU) for breast cancer dataset with K-Clusters and each cluster maintains finite set of attributes without any redundancy. Results are shown with better accuracy than existing methods by selecting various number of features.

**Keywords**:- Classification, Cluster, High-Dimensionality, Feature Selection, Balanced dataset.

## I.INTRODUCTION

Feature assortment is a pre-processing exercise, which is useful for great dimensional datasets. It is used for removing the unwanted features of the dataset. By means of this method, it will reduce the amount of input variables while developing the model[9]. By using this pre-processing technique the curse of multi- dimensionality [4] can be minimised. This technique selects only key fields that are very important to predict the class label. It eliminates the fields, that are unneeded for the prediction of the output. It removes the insignificant features and attempts to find the set of attributes that can best describes the data. [10] Dimensionality curse is the major problem due to which the end result could not be calculated accurately. By using this technique we can overcome this issue. It also eliminates the redundancy problem, increases the accuracy and then reduce the training time by removing the unrelated or unwanted features. Unsupervised feature selection method take the portion of features without taking the target variable into account. Whereas Supervised feature selection methods use the target variable in subset selection.

They are mainly categorised into three groups: Filter, Embedded, Wrapper methods. Filter methods takes quality based on their characteristics rather than using any machine learning algorithm[1][19][20]. Hence it is computationally inexpensive. In turn filter methods are of two types: Univariate and Multivariate. Univariate methods evaluate and rank the features based on some criteria. Whereas, Multivariate methods ranks the entire feature space based on the relations among the attributes. Some basic filter methods include removing constant, quasi-constant and duplicate features. There are also some Correlation Filter methods like Pearson correlation coefficient, Spearman's rank correlation coefficient and Kendall's rank

correlation coefficient. Statistical and Ranking Filter methods include Mutual Information, Chi-squared Score, ANOVA Univariate Test and Univariate ROC-AUC[2]. By the way of alternative Wrapper methods use ML algorithms for estimating a subset of features, later to generate an optimal subset of features. They work like greedy algorithms. These methods first searches for a split of characteristics , then build a learning model on the selected features whose performance is evaluated. This process continues until desired condition is met. Search methods include Forward Feature Selection, Backward Feature Elimination, Exhaustive Feature Selection and Bidirectional search. One of the advantage of Wrapper methods is to identify the communication of connecting variables. Embedded approach results in taking the suitable feature at the time of model training itself rather than first selecting optimal feature subset and then model training it embeds both the steps. It includes the advantages of both the filter and wrapper methods. Regularization and Tree-based methods use embedded feature selection methods.

Selecting appropriate feature selection method based on the data types of attributes is very important. For Numerical Output - Regression Predictive Modeling and for Categorical Output – Classification Predictive Modeling. In case of Numerical Input, Numerical Output: Correlation Feature Selection methods such as Pearson's for linear correlation or rank-based methods for non-linear ones work best. For Numerical Input, Categorical Output: ANOVA correlation coefficient for linear and Kendall's rank coefficient for non-linear correlations. In case of Categorical Input, Numerical Output: methods used of numerical input and categorical output works well. And finally Categorical Input, Categorical Output: Chi-Squared test and Mutual Information works good.

Good number of proved works related to feature selection methods[5] and few on hybrid feature selection methods. They have developed a new combination whichever of wrapper and filter methods or filter and embedded methods (There are three varieties of feature selection approaches Filter, Wrapper and Embedded). Building a proper combination of methods is the main challenge. There are many filter methods like [7] Symmetric Uncertainty (SU), Correlation[6], Statistical and Ranking methods. Filter methods select a subset of features depending on their characteristics rather than using any classification algorithm. But Wrapper methods like Multi-Layer Perceptron (MLP) uses a classification algorithm to evaluate the subsets and finally generates a subset of best features. Main intention is to reduce the quantity of fields, so that improve the efficiency of the data organization model.

Class imbalance is a major problem in pre-processing, So it can be solved by using the available technique referred in the paper[8] , to do avoid the biased towards large class of the databank. In this research paper, filter methods are considered for comparing the feature subsets formed by proposed technique. Table 1 represents various feature selection methods depending on filter based mechanism and its corresponding functional analysis.

| S.No | Name of the Method | Functional analysis |
|---|---|---|
| 1 | Information Gain (IG) | Ranking |
| 2 | ReliefF (Rel) | Ranking |
| 3 | Chi- Square (Chi) | Ranking |
| 4 | Symmetrical Uncertainty (SU) | Ranking |

Table 1 List of Feature Selection methods considered

In paper [17] authors used I-G, CHI-Sq, GR, SU, one-R, REL for generating ranks of every attribute of credit data of Germany and Australia. Each technique created dissimilar ranks to each attribute. Feature Selection method is proposed based on Genetic Algorithm by authors of [18] [16] to progress the classification results of various medical datasets.

The key criteria is form the 'K' clusters of various features is SU,

i.e., SU=2*IG/(H(U)+H(V))
Entropy of U is denoted as H(U)
Entropy of V is denoted as H(V)
SU considers the values in the range (0,1).
Suppose SU value 0, which represents two attributes are uncorrelated.
SU value 1 which represents one attribute can predict totally.

## II.Proposed Algorithm

In the paper [11] authors developed using Symmetric Uncertainty (SU) for DFS, Correlation-based Feature Subset Selection (CFS), Multi-Layer Perceptron(MLP) feature selection methods. Features are disseminated between the a choice of clusters without any repetition using the two filter methods SU and CFS. Out of these only one strong and best feature subset, depending on the outcome of MLP is choosen which has restricted number of features. In the paper [12], a hybrid two-phased feature selection method is proposed. In the first phase, the sample domain analysis is performed and a secondary dataset is drawn from the base dataset. In the first step, the SMOTE technique is applied on the original dataset. In the second phase, the feature space is searched to reach the best subset that results in the best accuracy using Information Gain (a Filter Method) and a Wrapper Method ( using Genetic, Naïve Bayes Algorithms). Their proposed method outperformed other feature selection methods on different datasets with different dimensions. In the paper [13], a method REL and PCA is projected and it is fitting for equally transcript and micro-array datasets. The method is evaluated by using high dimensional medical datasets like Central Nervous System Dataset, ColorTumor Dataset, Leukemia Cancer Dataset etc., This approach reduced almost half of the mislanious and unwanted attributes from the dataset.

In paper [14] a method is proposed based on attribute assortion chose k-best and select percentile whose combination reduced the over fitting, training time and improved the accuracy. This approach develops the classification accuracy for the algorithm like Naïve Bayes, Support Vector Machine, Logistic regression and K- Nearest Neighbour. Comparing to all other algorithms Logistic Regression has achieved best precision with 96.9%. In the paper [15], using principal component analysis, Chi squared testing, ReliefF and symmetrical uncertainty methods distinctive feature sets are created. Then, few classification algorithms are analysed with models to compare the optimal features combinations, to improve the correct prediction of heart conditions. In paper [21] authors used multi-dimensional database to mine popular patterns from vertical fomat.

| S.No | Research Paper | Feature selection Method | Classification Method | Result/Remark |
|---|---|---|---|---|
| 1 | Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance | Symmetric Uncertainty, Correlation-based Feature Subset Selection and Multi Layer Perceptron (MLP) | K-N-N, SVM | 57% success rate against the traditional methods |
| 2 | Exploring feature selection and classification methods for predicting heart disease | Chi squared testing, ReliefF , PCA, symmetrical uncertainty | BayesNet algorithm | 85.0% |
| 3 | Selecting critical features for data classification based on machine learning methods | Boruta, and Recursive Feature Elimination (RFE) | Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (K-N-N) | 93.26% |
| 4 | A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithms | Information Gain, Genetic + Naïve Bayes wrapper methods | Naive Bayes, Logistic, Multilayer Perceptron, Best First Decision Tree | 95% |
| 5 | An Efficient Hybrid Feature Selection model for Dimensionality Reduction | REL and PCA | Evaluated using different thresholds – mean, median &standard deviation | reduces more than 50% of irrelevant and redundant features from the original datasets |
| 6 | A comparative study on the effect of feature selection on classification accuracy | Information Gain , Symmetrical Uncertainty , Relief-F, Chi - square | Naive Bayes, MLP and decision tree | classification accuracy is improved up to 15.55% |

## III. Research Methodology

The main objective of this approach is to reduce the data area. assume the databank contains 'M' number of features. From 'M' number of features, there will be constraint to choose best conventional 'N' number of features exclusive of any replication, in this approach total

number of groups C(M, N) subsets can be created In high dimensional dataset, which one is also tuff task, to analyse those groups. Hence, filter based ranking methods can be utilized to calculate the rank of each attribute and then more popular 'N' attributes can be measured for investigation. Generation of feature selection procedure is given below

Step1: Consider the Balanced Data set
Step2: Calculate the Rank of each feature based on Symmetrical Uncertinity
Step3: Arrange the features According to order of rank
Step4: Select Total number of features based on SU>0
Step5: identify the required subsets and arrange the features from left to right
Step6: Accumulate all the features of subsets in required direction.

**Data set Description:**

To analyse the proposed procedure, medical dataset (Breast Cancer) are collected from UCI repository. This databank has 569-number of records, 30- number of features, and 2 class labels. The Breast Cancer dataset narrative is specified in the Table number 3.

| S.No | Breast Cancer Dataset | Statistics |
|---|---|---|
| 1 | Total No of Records | 569 |
| 2 | Total No of Features | 30 |
| 3 | Total No of Classes | 2 |

Table 3: Breast Cancer Dataset description

The above specified class have 2 diagnosis values (B = benign, M = malignant). Initial class distribution of the breast cancer dataset is specified in Table number 4.

| Class name | No Of samples | Percentage(%) |
|---|---|---|
| B | 357 | 62.75 |
| M | 212 | 32.25 |

Table 4: Class Distribution-Breast Cancer Dataset

It is clear that, class name B has other tuples rather than class name M, So marginal class records need to be bigger to meet the common class instances for the enhanced outcome. To do balance SMOTE runs on the K-N-N algorithm and produce the synthetic records; In this, experiment K-N-N value 5 is considered. Following table 5 gives the balanced class distribution of dataset.

| Class name | No Of samples Increased | No of samples formulated |
|---|---|---|
| B | 0 | 357 |
| M | 60 | 339 |

Table 5: Balanced Dataset-class distribution

## IV. Results Discussion

For investigating the strength of each subset of attributes, an equal number of top attributes formed by I-G, Chi-sq, REL, GR are considered. The same framework is tested with balanced Breast cancer (BBC) databank and as well as Imbalanced Breast cancer (IBBC) databank. SU is applied on BBC and IBBC databanks to discover rank of every attribute.

| Databank Name | Features selection |
|---|---|
| Balanced Breast cancer(BBC) | 28-23-21-8-24-27-7-4-14-3-1-11-13-6-26-17-2-18-22-16-5-25-29-9-30-20-12-19-15-10 |
| Imbalanced Breast cancer(IBBC) | 23- 21- 24- 28- 8- 3- 7- 4- 1- 27- 14- 11- 13- 6- 26- 17- 2- 18- 22- 25- 29- 16- 5- 30- 9- 19- 20- 10- 12- 15 |

Table 6: Selected Features of balanced and Imbalanced Databanks

| Subset no of features | Subset Id | Balanced Databank (BBC) features | Imbalanced Databank (IBBC) features |
|---|---|---|---|
| **Subsets @5** | BC-5-1 | 28-3-1-16-5 | 23- 27- 14- 25- 29 |
| | BC-5-2 | 23-14-11-22-25 | 21-1- 11- 22- 16 |
| | BC-5-3 | 21-4-13-18-29 | 24- 4- 13- 18- 5 |
| | BC-5-4 | 8- 7- 6- 2- 9 | 28-7- 6- 2- 30 |
| | BC-5-5 | 24- 27- 26- 17-30 | 8- 3- 26- 17- 1 |
| | BC-5-IG | 23- 24- 28- 21-8 | 23- 24- 21- 28- 8 |
| | BC-5CHI | 23- 28- 24- 21- 8 | 23- 28- 24- 21- 8 |
| | BC-5-GR | 28- 23- 21- 8- 27 | 28- 23- 21- 8- 27 |
| | BC-5- REL | 21- 23- 28- 3- 1 | 21- 23- 28- 3- 1 |
| **Subsets @4** | BC-4-1 | 28- 4- 14- 17- 2- 9- 30 | 23- 4- 1- 17- 2- 30 |
| | BC-4-2 | 23- 7- 3- 26- 18- 29- 20 | 21- 7- 27- 24- 18- 5 |
| | BC-4-3 | 21- 27-1- 6- 22- 25- 12 | 24- 3- 14- 6- 22- 16 |
| | BC-4-4 | 8- 24- 11- 13- 16- 5- 19 | 28- 8- 11- 13- 25- 29 |
| | BC-4-IG | 23- 24- 28- 21- 8- 3- 7 | 23- 24- 21- 28- 8- 3 |
| | BC-4-CHI | 23- 28- 24- 21- 8- 7- 3 | 23- 21- 24- 28- 8- 3 |
| | BC-4-GR | 28- 23- 21- 8- 27- 24- 7 | 23- 21- 24- 28- 8- 7 |
| | BC-4- REL | 21- 23- 28- 3- 1- 8- 24 | 21- 28- 23- 22- 1- 3 |
| **Subsets @3** | BC-3-1 | 28- 27-7-11-13-18-22-9-30 | 23- 3- 7- 11- 13- 18- 22- 30- 9 |
| | BC-3-2 | 23-24-4- 1- 6- 2- 16- 29- 20 | 21- 8- 4- 14- 6- 2- 25- 5- 9 |
| | BC-3-3 | 21-8-14-3-26-17-5-25-12 | 24- 28- 1- 27- 26- 17- 29- 16-20 |

| | | | |
|---|---|---|---|
| | BC-3-IG | 23- 24- 28- 21- 8- 3- 7- 1- 4 | 23- 24- 21- 28- 8- 3- 4- 1- 7 |
| | BC-3-CHI | 23- 28- 24- 21- 8- 7- 3- 4- 1 | 23- 21- 24- 28- 8- 3- 4- 1- 7 |
| | BC-3-1GR | 28-23-21—8-27-24-7-4-14 | 23- 21- 24- 28- 8- 7- 27- 3- 4 |
| | BC-3-1REL | 21- 23- 28- 3- 1- 8- 24- 22- 4 | 21- 28- 23- 22- 1- 3- 8- 24- 4 |

Table 7: Features subsets- Where # Subsets are @5, @4 and @3.

The outcome of each classifier (K-N-N- J-Rip- N-B- J-48) against the each subset of features with ranks are specified in this section for discussion. Each subset rank by the selected classifier is denoted by slash.

| **Balanced Databank(BBC) features @5** | | | | | **Imbalanced Databank(IBBC) features @5** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ID | K-N-N | J-Rip | -N-B | J-48 | ID | K-N-N | J-Rip | -N-B | J-48 |
| BBC-5-1 | 93.39/3 | 93.82/5 | 93.96/2 | 92.95/3 | IBBC51 | *93.32/2* | *93.84/2* | *94.02/2* | *94.20/2* |
| BBC-5-2 | **96.69/1** | **96.12/1** | **95.54/1** | 89.22/7 | IBBC52 | 90.33/7 | 92.61/5 | 93.84/3 | 92.44/4 |
| BBC-5-3 | 93.24/5 | 92.81/7 | 93.24/4 | 90.08/6 | IBBC53 | 91.73/5 | 92.97/4 | 91.91/5 | 89.10/7 |
| BBC-5-4 | 91.81/7 | 92.38/8 | 90.22/7 | 90.94/5 | IBBC54 | 92.44/3 | 91.91/6 | 91.21/6 | 89.45/6 |
| BBC-5-5 | 91.54/8 | 93.67/6 | 93.24/4 | 92.24 | IBBC55 | 90.86/6 | 91.03/7 | 91.03/7 | 91.91/5 |
| BBC- IG | 92.67/6 | 95.53/2 | 92.52/6 | 94.25/2 | IBBC- IG | 92.26/4 | 93.67/3 | 92.61/4 | 94.20/3 |
| BBC -CHI | 92.67/6 | 95.53/2 | 92.52/6 | 94.25/2 | IBBC-CHI | 92.26/4 | 93.67/3 | 92.61/4 | 94.20/3 |
| BBC -GR | 93.53/4 | 95.11/3 | 93.67/3 | **94.97/1** | IBBC- GR | 92.26/4 | 93.67/3 | 92.61/4 | 94.20/3 |
| BBC – REL | 94.39/2 | 94.25/4 | 93.10/5 | 92.24/4 | IBBC - REL | **95.95/1** | **94.55/1** | **94.20/1** | **95.07/1** |

Table 8: Performance of Balanced and Imbalanced databanks @ features 5.

From the above Table, it specifies that, BBC-5-2 made the maximum performance than all existing methods with K-N-N, J-Rip, N-B over the Balanced dataset. Similarly BBC-5-1 feature subset made better performance than existing BBC-IG, BBC-GR, BBC-CHI with all Classifiers in imbalanced dataset.

| **Balanced Databank(BBC) features @4** | | | | | **Imbalanced Databank(IBBC) features @4** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ID | K-N-N | J-Rip | -N-B | J-48 | ID | K-N-N | J-Rip | -N-B | J-48 |
| BBC-4-1 | 95.25/2 | **95.54/1** | 94.97/2 | 93.96/3 | IBBC-4-1 | 93.49/4 | 94.20/4 | 92.61/5 | 93.32/5 |
| BBC-4-2 | 92.24/5 | 92.67/5 | 93.96/3 | 92.95/5 | IBBC-4-2 | *94.55/2* | *94.72/3* | 92.97/4 | *94.37/3* |
| BBC-4-3 | **96.26/1** | 94.82/2 | **95.11/1** | **95.11/1** | IBBC-4-3 | 92.79/5 | 92.61/6 | *93.67/2* | 92.61/6 |
| BBC-4-4 | 92.09/6 | 92.09/7 | 92.09/7 | 90.22/6 | IBBC-4-4 | 90.15/6 | 92.79/5 | 91.21/6 | 92.44/7 |
| BBC- I-G | 94.68/3 | 94.39/3 | 93.1/5 | 93.96/3 | IBBC IG | 94.20/3 | 95.25/2 | 93.14/3 | 93.67/4 |
| BBC -CHI | 94.68/3 | 94.39/3 | 93.1/5 | 93.96/3 | IBBC CHI | 94.20/3 | 95.25/2 | 93.14/3 | 93.67/4 |
| BBC – GR | 94.54/4 | 93.24/4 | 93.39/4 | 94.39/2 | IBBC GR | 93.49/4 | 94.20/4 | 92.61/5 | 94.55/2 |
| BBC -REL | 94.54/4 | 92.95/6 | 92.81/6 | 93.39/4 | IBBC REL | **95.43/1** | **95.43/1** | **94.20/1** | **94.72/1** |

Table 9: Performance of Balanced and Imbalanced databanks @ features 4.

From the above Table, it has been observed that Rel method performed better than all methods. IBBC-4-3 features subset made improved performance than IG, GR, and CHI with K-N-N over the imbalanced dataset. J-Rip performed well with BBC-4-1 subset of features. K-N-N, J-48, N-B recorded.

| Balanced Databank(BBC) features @3 | | | | | Imbalanced Databank(IBBC) features @3 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ID | K-N-N | J-Rip | -N-B | J-48 | ID | K-N-N | J-Rip | -N-B | J-48 |
| BBC-3-1 | *95.68/2* | 92.38/6 | 91.81/6 | 91.23/4 | IBBC-3-1 | 93.14/6 | 93.32/4 | 93.67/3 | 91.91/4 |
| BBC-3-2 | 93.67/5 | 93.1/5 | *93.24/3* | *93.24/3* | IBBC-3-2 | *95.25/2* | **95.43/1** | **96.3/1** | 92.97/3 |
| BBC-3-3 | *95.68/2* | *93.67/3* | *93.39/2* | *93.24/3* | IBBC-3-3 | 94.37/3 | 93.49/3 | 93.49/4 | 93.67/2 |
| BBC- IG | 95.25/3 | 93.53/4 | 93.10/4 | 93.53/2 | IBBC IG | 94.20/4 | 93.49/3 | 92.26/6 | **94.02/1** |
| BBC - CHI | 95.25/3 | 93.53/4 | 93.10/4 | 93.53/2 | IBBC CHI | 94.20/4 | 93.49/3 | 92.26/6 | **94.02/1** |
| BBC -GR | 94.54/4 | 93.95/2 | 93.1/5 | **93.82/1** | IBBC GR | 93.32/5 | 92.09/5 | 92.79/5 | **94.02/1** |
| BBC – REL | **97.27/1** | 95.11/1 | **94.97/1** | *93.24/3* | IBBC REL | **95.78/1** | 94.55/2 | 94.90/2 | **94.02/1** |

Table 10: Performance of Balanced and Imbalanced databanks @ features 3.

From the above table it is specified that, all the existing methods performed well over the imbalanced dataset. REL performed better than all methods with the K-N-N and J-48, but BBC-3-1 and BBC-3-3 recorded excess accuracy other than REL over the balanced dataset. IBCREL subset of features got positive performance with the K-N-N. IBC-3-2 performing better than other existing methods except REL. IBC-3-2 recorded improved performance with J-Rip and J-48.

## IV. CONCLUTION

The proposed method is analyzed using real time balanced Breast Cancer dataset collected from UCI databank. In this paper- ranking framework with K-cluster of dimensions reduction is applied. This methodology can be applied for any sort of datasets. The result of this method is specified with good accuracy prediction.

## References

1.Jason brownlee " How to choose feature selection method for machine learning"-On Machine learning Masterly-2019.
2.Kargupta-et.al "Data Mining: Next Generation Challenges and Future Directions"-PHI Learning- 2009.
3.SP Potharaju- M Sreedevi" Ensembled rule based classification algorithms for predicting imbalanced kidney disease data" Journal of engineering science and technology review 9 (5)-201-207
4.Vijay Kumar- G.- Krishna Chaitanya- T.- Pratap- M." Mining popular patterns from multidimensional database" Indian Journal of Science and Technology- 2016- 9(17)- 93106
5.J Novaković - Yugoslav "Toward optimal feature selection using ranking methods and classification algorithms"Journal of Operations Research- 2016 - yujor.fon.bg.ac.rs
6.P Malji- S Sakhare " Significance of entropy correlation coefficient over symmetric uncertainty on FAST clustering feature selection algorithm" on Intelligent Systems and Control (ISCO)- 2017 - ieeexplore.ieee.org
7.B Singh- N Kushwaha- OP Vyas – "A feature subset selection technique for high dimensional data using symmetric uncertainty" Journal of Data Analysis and …- 2014.
8.Chawla- N.V.- Bowyer- K.W.- Hall- L.O. and Kegelmeyer- W.P.- 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research-16- pp.321-357.

9.Vijay Kumar- G.- Bharadwaja- A.- Nikhil Sai- N. "Temperature and heart beat monitoring system using IOT" Proceedings - International Conference on Trends in Electronics and Informatics- ICEI 2017- 2018- 2018-January- pp. 692-695
10.SP Potharaju- M Sreedevi "A Novel M-Cluster of Feature Selection Approach Based on Symmetrical Uncertainty for Increasing Classification Accuracy of Medical Datasets." Journal of Engineering Science & Technology Review 10 (6)
11.Sai Prasad Pothuraju- M.Sreedevi "Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance- Clinical Epidemiology and Global Health-Volume 7- Issue 2- June 2019- Pages 171-176
12.Mehdi Naseriparsa, Amir-Masoud Bidgoli, Touraj Varaee "A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithms "- International Journal of Computer Applications-Vol 69-No 17-pp 28-35-2013.
13.D Jain- V Singh- "An Efficient Hybrid Feature Selection model for Dimensionality Reduction"- Procedia Computer Science- 2018 – Elsevier.
14.R Nair- A Bhagat-" Feature Selection Method To Improve The Accuracy of Classification Algorithm"- International Journal of Innovative Technologyand Explore. Eng- 2019
15.Robinson Spencer-et.al- "Exploring feature selection and classification methods for predicting heart disease" Digital health- volume 6- Jan-Dec 2020.
16.Sravya- G.- Sreedevi- M." Genetic optimization in hybrid level sentiment analysis for opinion classification" International Journal of Advanced Trends in Computer Science and Engineering- 2020- 9(2)- pp. 1440-1445- 81
17.Novaković- J.- 2016. Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research-21(1).
18.Singh- D.A.A.G.- Leavline- E.J.- Priyanka- R. and Priya- P.P.- 2016. Dimensionality reduction using genetic algorithm for improving accuracy in medical diagnosis. International Journal of Intelligent Systems and Applications- 8(1)- p.67.
19. Vijay Kumar- G. Sreedevi-M., Bhargav- K and Mohan Krishna- P.-2018 Incremental mining of popular patterns from transactional databases. International journal of Engineering and Technology-7 –pp 636-641.
20. Vijay Kumar- G. Sreedevi-M. Vamsi krishna- G and Sai Ram-N. 2018 Regular frequent crime pattern mining on crime datasets. International journal of Engineering and Technology-7 –pp 972-975.
21. G. Vijay Kumar, T. Krishna Chaitanya, M. Pratap, "Mining Popular Patterns from Multidimensional Database", Indian Journal of Science and Technology, Vol 9(17), DOI: 10.17485/ijst/2016/v9i17/93106, May 2016.