Grenze International Journal of Engineering and Technology, June Issue

A Hybrid Feature Selection Technique to Predict Breast Cancer

Tintu P B[1] and Dr. S Manju Priya[2]
[1-2]Karpagam Academy of Higher Education Coimbatore, Tamil Nadu, India
Email: tintupadikkal@gmail.com, smanjupr@gmail.com

***Abstract* – Medical data mining models are predominant in medical data analysis. In our paper, we have applied the different classifier algorithm; along with a new improved feature selection method is applied, towards an accurate detection of breast cancer, in its early stages. Performance of the classifiers was seen to be noticeably enhanced when irrelevant features were screened out from the modeling process.**

***Keywords*: Classification Algorithms, Feature Selection, Feature Selection, Reliable Features, Information Gain, Relif Method, MIFGRF.**

I. Introduction to Breast Cancer

Now a days breast cancer has been reported as one of the main and leading causes of death in women who are middle aged. Recent reports shows that over and above 7, 00,000 new cases of breast cancer are found all over the world. Throughout their lifetime there is chance that one in twelve women can be affected by this disease. Cancer is curable if diagnosed and treated in its earlier stages. But detection of cancer at early stage if not done can lead to death of large numbers of the female population. This digenesis has changed a lot than previous few years and to this issue so many resolutions are explored, but with a large set of features. The term data mining refers loosely to the process of semi automatically analyzing large databases to find useful patterns [1]. High dimensional datasets usually lead to deteriorate the accuracy and performance of the system by curse of dimensionality. Datasets with high dimensional features have more complexity and spend longer omputational time for classification [2]. Feature selection is a solution to high dimensional data. Feature selection is an important topic in data mining, specifically for high dimensional datasets. Feature selection is a process commonly used in machine learning, wherein subsets of the feature available from the data are selected for application of a learning algorithm. The best subset contains the least number of dimensions that most contribute to accuracy [3]. Apart from the already existing feature selection methods, we have introduced a new hybrid method, which is a combination of feature selection method(s) like univariate selection and recursive elimination, in order to extract the most prominent features. In this data-mining model, the prominent features obtained from hybrid approach can be used to study the diagnostic details of cancer for an effective treatment.

II. Review of Literature

Feature selection aims to improve machine learning performance [4]. Janecek [5] showed the relationship between feature selection and data classification and the impact of applying pca on the classification process. Assareh [6] proposed a hybrid random subspace fusion model that utilizes both the feature space and sample domain to improve

Grenze ID: 01.GIJET.9.2.408_1

the diversity of the classifier ensemble. Hayward [7] showed that data preprocessing and choosing suitable features will develop the performance of classification algorithms. Dhiraj [8] used clustering and kmeans algorithm to show the efficiency of this method on huge amount of data. Xiang [9] proposed a hybrid feature selection algorithm that takes the benefit of symmetrical uncertainty and genetic algorithms. Zhou [10] presented a new approach for classification of multi class data. The algorithm performed well on two kinds of cancer datasets. Fayyad [11] tried to adopt a method to seek effective features of dataset by applying a fitness function to the attributes. Most of the feature selection methods work on feature space and they do not test the effect of filtering and resampling instances on the feature selection process. Furthermore, feature selection methods usually focus on one specific classification algorithm to test their performance. Hence, only one part of the sample space patterns are covered [6].

III. Feature Selection

Feature selection can be defined as a process that chooses a minimum subset of m features from the original set of n features, so that the feature space is optimally reduced according to a certain evaluation criterion [12]. As the dimensionality of a domain expands, the number of feature n increases. Finding the best feature subset is usually intractable [13] and many problems related to feature selection have been international feature selection algorithms may be divided into filters [15], wrappers [13] and embedded approaches [16]. Filters method evaluates quality of selected features, independently from the classification algorithm, while wrapper methods require application of a classifier to evaluate this quality. Embedded methods perform feature selection during learning of optimal parameters.

A. Information gain

Information gain is the measure of change in homogeneity or it is a measure of reduction in entropy.information gain of a given attribute x with respect to attribute y is reduction in uncertainty about value y when known the value of x.the information gain for a feature f is calculated as the difference between the entropy in the segment before the split (s 1), and the partitions resulting from the split (s 2). Equation (1) is

$$\text{Infogain}(f) = \text{entropy}(s1) - \text{entropy}(s2). \quad (1)$$

The higher the information gain the better a feature is at creating homogeneous group. Entropy is the measure of impurity in a data set.the information gain of a given attribute x with respect to the class attribute y is the reduction in uncertainty about the value of y when we know the value of x. The uncertainty about the value of y is measured by its entropy, h(y). The uncertainty about the value of y when we know the value of x is given by the conditional entropy of y given x (5) ig is a symmetrical measure [12]. The information gained about y after observing x is equal to the information gained about x after observing y.

B. Relief

Relief is takes a filter-method approach to feature selection that is notably sensitive to feature interactions.relief calculates a feature score for each feature which can then be applied to rank and select top scoring features for feature selection. Relief feature scoring is based on the identification of feature value differences between nearest neighbor instance pairs. If a feature value difference is observed in a neighboring instance pair with the sameClass (a 'hit'), the feature score decreasesalternatively, if a feature value difference is observed in a neighboring instance pair with different class values (a 'miss'), the feature score increases.working of rfe starting with all features in the training data set then remove the irrelevant features one by one in each iteration until reaches to the desired number. Take a data set with n instances of p features, belonging to two known classes.at each iteration, take the feature vector (x) belonging to one random instance, and the feature vectors of the instance closest to x (by euclidean distance) from each class.the closest same-class instance is called 'near-hit', and the closest different-class instance is called 'near-miss'. Update the weight vector such that equation (2) is

$$Wi=wi-(xi-nearhiti)^2+(xi-near). \quad (2)$$

V. Proposed Method

The dataset is passed to mifgrf, which selects the best relevant features using information gain. Feature selected are perimeter worst, area worst, radius worst, concave point worst and mean, perimeter, area, radius and concavity mean and area se, concativity worst, perimeter se, radius se , compactness worst and compactness mean. On applying these features to mifgrf (modified information gain relief) the following features are

selected perimeter se, radius mean, area se, concativity worst, area mean, concave point worst and mean, compactness mean.on selecting these features and comparing with random forest we get the 96.9% accuray.these classifiers were applied on wbdc dataset which consists of 569 instances which was obtained from wisconsin dataset from kaggle repository.

IV. System Design

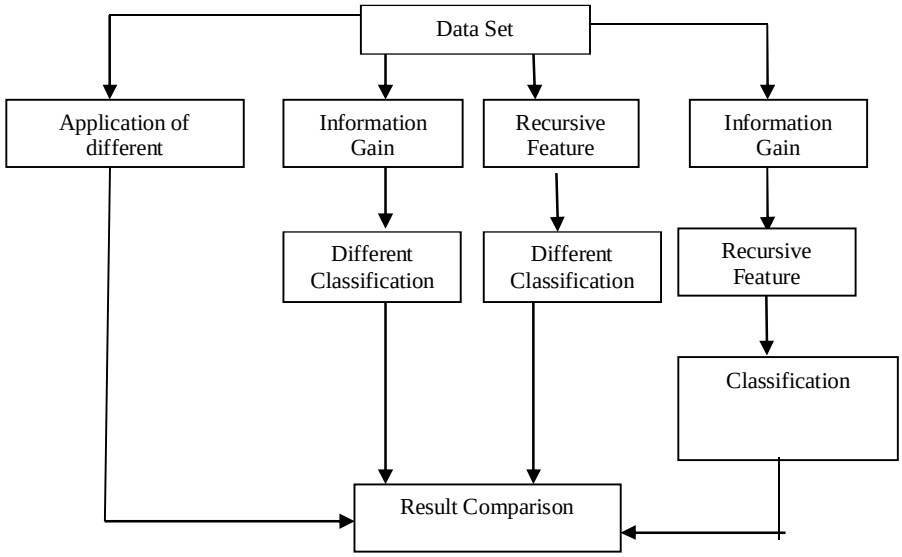


Figure 1. Flowchart

VI. Experimentation Results

Table 1: On applying Information gain and On performing classification we get the following result

Sno	Classifier	Execution time in seconds	Classification accuracy%
1	Random Forest	0.0013s	93.3%

Table II. On applying RELIF algorithm and on Performing classification we get the following result

Sno	Classifier	Execution time in seconds	Classification accuracy%
1	Random Forest	0.0001s	95.0%

Table III. On applying MIFGRF algorithm and on Performing classification we get the following result

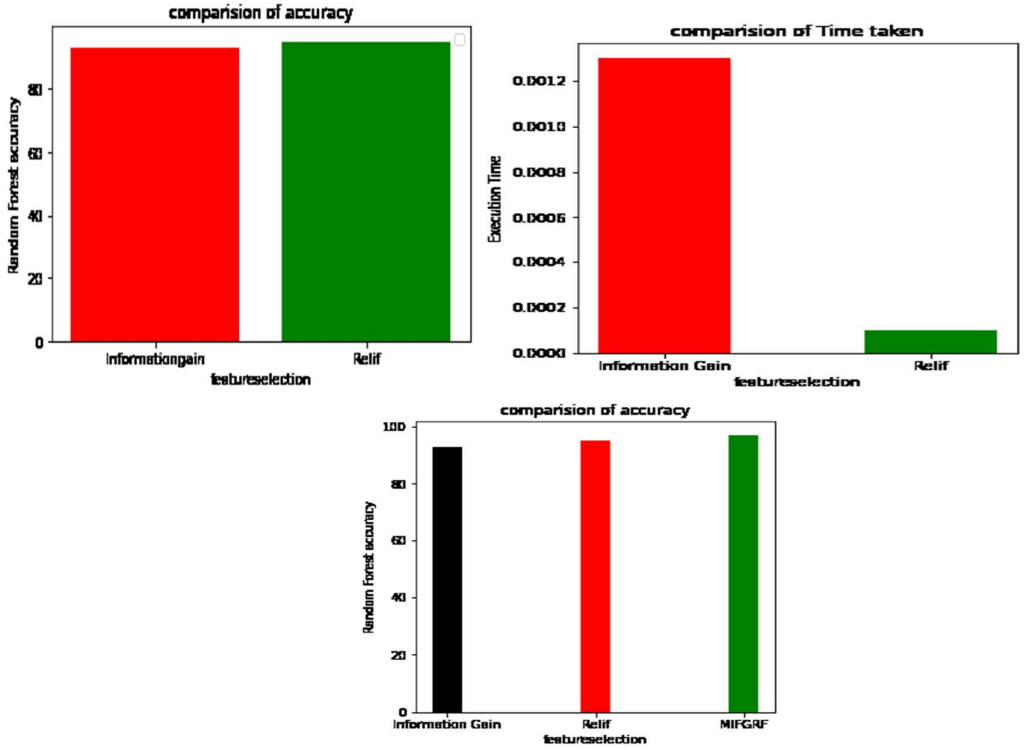
Sno	Classifier	Execution time in seconds	Classification accuracy%
1	Random forest	0.002 s	96.9%

TABLE IV. COMPARISON OF RESULTS

Sno	Feature selection	Execution time in seconds	Classification accuracy%
1	Information gain	0.0013s	93.3%
2	Relif	0.0001s	95%
3	Mifgrf	0.002s	97%

VII. RESULTS & DISCUSSION

In this figure 1 feature selection is done with information gain and relif method, randomforest has the accuracy of 93% and 95%.
In this figure 2, execution time taken by information gain and relif method is represented minimum executiontime is taken by relif.
In this figure 3, feature selection is performed using mifgrf method and accuracy for each classifiers is represented.



VIII. CONCLUSION

In our work we have proposed a new hybrid feature selection method for earlier prediction of breast cancer. Breast cancer is growth of tumors which if detected early can save lives of women and men. By using machine learning algorithms we can classify and predict breast cancer as benign or maligant.from the analyze using wbdc datasets, it is noted that out using hybrid feature selection method can improve the performance of classification. Further works are carried out to provide accurate results in diagnosing breast cancer.

REFERENCES

[1] Silberschatz, a, korth and sudarshan, “database system concepts”, mcgrawhill, 2010.
[2] Bellman,” adaptive control processes: a guided tour”, princeton university press, princeton, 1996.
[3] Ladha, deepa, “feature selection methods and algorithms”, international journal on computer science and engineering, volume 3, 2011.
[4] Liu, and zhao, “manipulating data and dimension reduction methods: feature selection”,journal of computational complexity, page(s) 1790- 1800, 2012.
[5] Janecek, gansterer, demel and ecker,” on the relationship between feature selection and classification accuracy, journal of machine learning and research”. Jmlr: workshop and conference proceedings 4, pages 90–105, 2008.