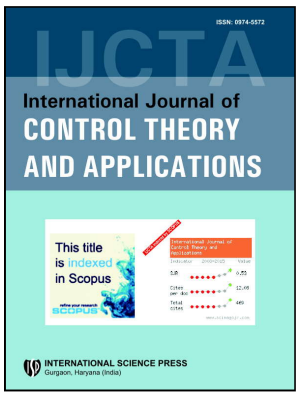

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 18 • 2017

Filter Based Attribute Optimization: A Performance Enhancement Technique for Healthcare Experts

Sushruta Mishra[1], Hrudaya Kumar Tripathy[2] and Brojo Kishore Mishra[3]

[1] *School of Computer Engineering, KIIT University, Bhubaneswar, Odisha, INDIA, Email: mishra.sushruta@gmail.com*
[2] *School of Computing, KIIT University, Bhubaneswar,Odisha, INDIA, Email: hrudayakumar@gmail.com*
[3] *Dept of IT, C.V. Raman college of Engineering, Odisha, INDIA, Email: brojokishoremishra@gmail.com*

Abstract: Now a day's massive amount of data is widely available in information systems. These data are of low quality,unreliable, redundant and are noisy in nature which negatively affects the process of observing knowledge and useful pattern. Machine learning techniques have attracted a big attention to researchers to turn such data into useful knowledge. Further relevant data can be extracted from huge records using filter based feature selection methods. In our study, a comparative analysis is drawn between four different filter based feature selection methods (Information gain method, Consistency based method and Correlation based method) based on Healthcare datasets (i.e., Breast cancer, Diabetes and Hepatitis). Multilayer perceptron were implemented to estimate the performance of the algorithms. The study revealed that filter based feature selection methods enhance the performance of learning algorithms.

Keywords: Multi Layer perceptron, Filter based Feature Selection, F-Measure, RMSE, Correlation based method, InformationGain method, Consistency based method.

1. INTRODUCTION

The health care centers are developed due to health consciousness of people in day today life. But proper disease diagnosis in present life is a very uphill task at manageable cost in such an emergent nation. Due to this, people are facing troubles and at times it causes the death of that person since all doctors may not be able to recognize and identify all diseases in time due to their poor attention and as well as due to the lack of modern instruments. Machine learning methodologies can be of great help in such cases. It forms the basis for knowledge discovery which is depicted in Fig. 1.1. It helps in intelligent analysis and processing of data, thereby minimizing the cost of computational power and thus enables us to use computationally intensive methods for data analysis. Further, with feature ranking using filter based methods researchers can extract relevant and high quality data from huge healthcare records. Feature ranking methods reduce the dimensionality of feature space thereby removing noisy data, enhancing data quality [J. Novakovic, 11].

2. RELATED WORK

In [SarvestanSoltani A, 11] the authors have drawn a comparative analysis of various machine learning methods based on neural networks like Multilayer Perceptron (MLP), Radial Basis Function (RBF) etc and classified WBC and NHBCD data for breast cancer. In [Chang Pin Wei,]Weipin Chang and his colleagues demonstrated genetic technique as the optimizing sarch technique used in breast cancer diagnosis and it produced a high prediction accuracy. K. Rajiv Gandhi and his colleagues published a research paper [Gandhi Rajiv K, 10] in which they used PSO search technique to develop a classification model for breast cancer patients data. In [Jiawei Han, 00] authors proposed a system model that gave an overall accuracy rate of 78.9% on heart Cleveland disease. Authors in [V. A. Sitar-Taut, 09] produced a model which resulted in 83.01% accuracy on the heart Cleveland disease diagnosis. The ANFIS classification [K. Srinivas, 10] with PCA of diabetes disease was classified due to training and test of all the diabetes disease dataset. That produced a classification accuracy of 89.47%. Roslina et al. used SVM to predict hepatitis and used wrapper based feature selection method to identify relevant features before classification. Combining wrapper based methods and Support vector machines produced good classification results [A. H. Roslina, 10]. Sartakhti et al. also presented a novel machine learning method

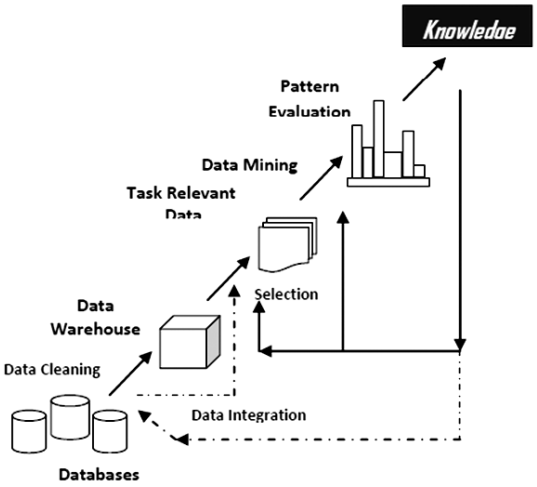

Figure 1: Knowledge Discovery Process

using hybridized SVM and simulated annealing to predict hepatitis and obtained high classification accuracy rates [J. S. Sartakhti, 11] Harb et al. proposed the filter and wrapper methods combined with PSO for medical data. Their proposed model illustrated a very high prediction accuracy among the others [H. Harb, 14].

3. FILTER BASED FEATURE SELECTION

Feature Selection methods are the optimizing agents in a machine learning algorithm. The prime objectives of these methods are to eliminate noisy data from the dataset. Attribute selection methods can be categorized into

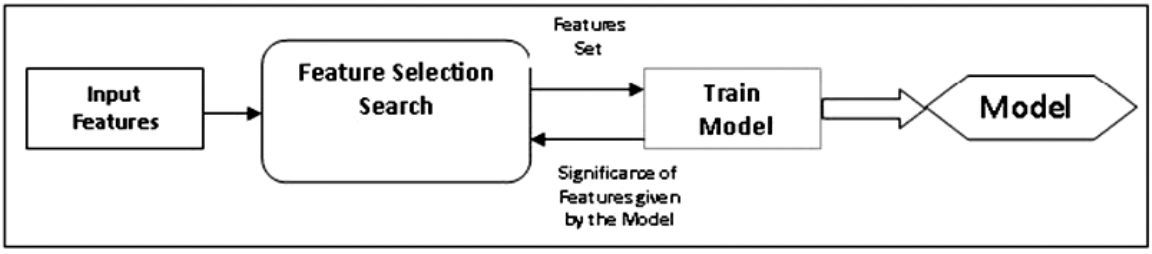

Figure 2: Wrapper based Feature Selection

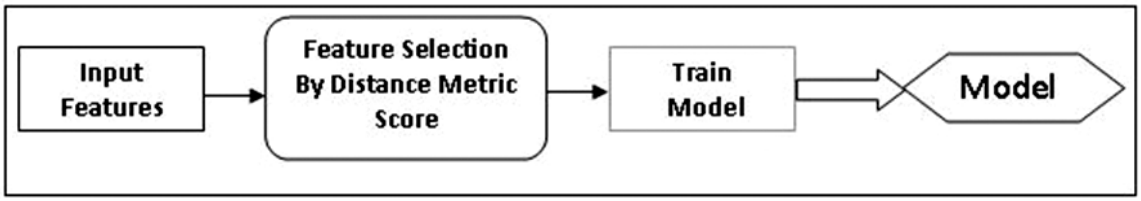

Figure 3: Filter based Feature Selection

two parts: Wrappers and Filters. The Wrapper determines attributes based on accuracy estimates by the target learning algorithm. While a filter method uses the statistical correlation between a set of variable and the target variable. Fig. 1.2 and Fig.1.3 highlights these two methods of Feature selection. These methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. Ranking of features determines the importance of any individual feature, neglecting their possible interactions. The correlation quotient between features and the target attribute computes the significance of target attribute [M. Ashraf, 13], [M. Leach,12]. In our research we have formulated and demonstrated four critical Filter based methods as shown in table 1.1.

4. PROPOSED WORK

4.1. Correlation Based Feature Optimization

It is a heuristics based method to find the goodness of an attribute subset. It correlates various attributes based on the usefulness of the feature set to predict the class label. It assumes that attribute set is considered good if they share a strong correlation with their class but less correlated with each other. Let the relation between every test variable with their extraneous variable is given at prior. Let the correlation among every attribute pair is known. Thus there exist a correlation between the cumulative components and the extraneous variable which may be computed as:

$$r_{zc} = \frac{k\overline{r}_{zi}}{\sqrt{k + k - (k-1)\overline{r}_{ii}}} \tag{1}$$

Where

r_{zc} = Pearson's Correlation coefficient which depicts the relation of the cumulative attributes with the extraneous variable.

k = count of attributes present.

$\overline{r}_{zi}$ = average of correlations between all attributes and the extraneous variable.

$\overline{r}_{ii}$ = mean interrelationship between various attributes.

Three vital observations are inferred from this coefficient which are:

More is the correlations between the components and the extraneous variable, more will be the correlation between the composite and the extraneous variable. The correlation between the composite and the extraneous variable is directly proportional to the number of components in the composite. As the inter-correlation among the components reduces, the correlation between the composite and the extraneous variable is enhanced. The data dimensionality reduction process occurs by using symmetrical uncertainty thereby picking the variable subset that has the maximum coefficient value from equation 1 stated above.

4.2. Information Gain Based Feature Optimization

The basis of information theory domain is Entropy. It is a metric that represents the level of purity of a sample set taken in random. It denotes the unpredictable nature of a system model. For a random variable Y the entropy is represented by:

$$H(Y) = -\sum_{y \in y} p(y) \log_2 \left(p(y) \right) \tag{2}$$

P(y) = Marginal probability density function of the random variable Y.

Let S be the training dataset such that values of Y are partitioned based on the values of another attribute X. Thus the entropy value of Y with respect to X is represented by:

$$H(Y/X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y/x) \log_2 \left(p(y/x) \right) \tag{3}$$

Thus a parameter denoting the relative decrease in entropy of Y can be determined by the extra information that X projects about Y is referred as Information Gain (IG). It is stated as:

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \tag{4}$$

The above equation suggests that the information gained regarding Y after observing X is equal to the information gained regarding X after observing Y. Hence the attribute with highest information gain value is chosen as the basis for classification.

4.3. Consistency Based Feature Optimization

This method determines the worthiness of attribute subsets. It computes a consistency measure to evaluate the best feature subset. Three inferences are used in this method:

Inference 1: A pattern is inconsistent if there is a perfect matching for at least wo instances while their class labelsdiffers. Ex: In the two instances {1, 0, 0} and {1, 0, 1} identical values are noted for the two attributes in bothe instances but their class label is not the same.

Inference 2: Frequency of Inconsistency (FIR) is defined as the difference between the frequency of occurrence of aparticular pattern in data and largest frequency among all class labels. Ex: Let a pattern p occurs in np instances for a attribute subset. Among all np instances class label allotment is done as: c1 $\approx$ label1 : c2 $\approx$ label2 : c3 $\approx$ label3 such that c1 + c2 + c3 = np.

Now let us assume that c2 is the highest among all then: FIR = n – c2.

Inference 3: Rate of Inconsistency (RIs) may be referred as the ratio of cumulative combination of all frequency ofinconsistencies for all patterns in an attribute subset in a dataset to the total number of given instances. It is given as:

$$\text{RIs} = \frac{\text{FIRs of all patterns}}{\text{Total Instances}} \tag{5}$$

Thus to achieve attribute selection process important steps followed are:

Step 1: A candidate feature subset is input.

Step 2: Determine the Rate of Inconsistencies (RIs).

Step 3: The subset S remains consistent only when RIs $\leq \alpha$ (where α is a user defined threshold limit)

In our work we have used three vital clinical datasets which include Breast cancer, diabetes and Hepatitis. Three crucial filter based methods are implemented (CFS, InfoGain and Consistency) while Multilayer Perceptron is used as a classifier in our study.

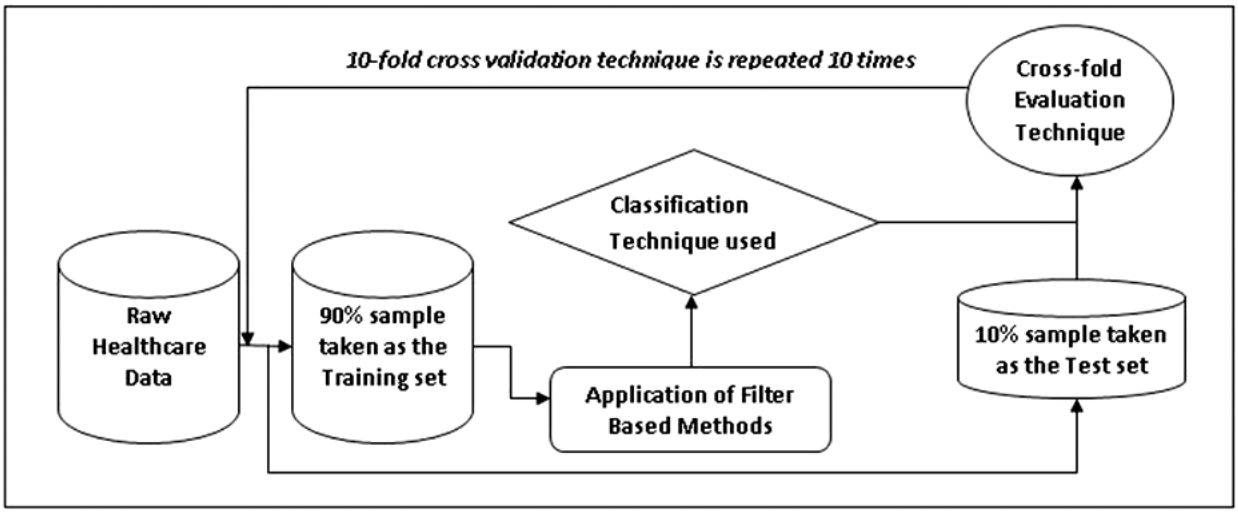

Figure 4: Implementation of Filter based feature selection with Cross validation technique for classification in Healthcare sector

Table 1
Proposed study work summary

Healthcare Datasets used	Breast Cancer, Diabetes and Hepatitis
Filter based Feature Selection methods used	Correlation, Information Gain and Consistency methods
Classification Technique used	Multilayer Perceptron

Our proposed work is based on implementation of filter based feature selection in healthcare industry. As seen in the diagram the original medical dataset is the input. It is sub divided into two distinct parts which includes Training set and Test set in the ratio 9:1. The training samples are applied to filter based feature selection methods like Information Gain method or Correlation based method to optimize the raw dataset. The output of implementing filter based techniques is an optimized reduced feature set. This reduced set is applied to a machine learning technique and thus consideration is 10-fold cross validation. We have applied cross validation method as the evaluation technique to categorize the entire medical dataset into training set and test set. Cross-Validation is a statistical process used to evaluate machine learning techniques by partitioning data into two segments: one segment is used to train a model while the other segment is used for model validation. We have used a 10-fold cross validation technique to evaluate healthcare datasets. In such scenario the data is first partitioned into 10 equally sized segments. Eventually 10 iterations of training and testing are performed in such a way that at each iteration it yields a different fold of the data to be held-out for validation while the remaining 9 segments are used for learning.

Table 2
Filter based techniques used in our study

Filter based method	*Evaluation*
Correlation based method	Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.
Information Gain method	Evaluates the worth of an attribute by measuring the information gain with respect to the class.
Consistency based method	Evaluates the worth of a subset of attributes by the level of consistency in theclass values when the training instances are projected onto the subset of attributes.

5. RESULTS AND ANALYSIS

Our entire research is carried out using WEKA 3.12 which is widely popular machine learning software. In the first step the original medical datasets are subjected to classification using Multilayer perceptron classifier. In the second step filter based feature selection techniques are implemented to the datasets before carrying out classification process. An extensive series of results are inferred from the experimental set up. It includes the confusion matrix of every classification process. Various performance parameters like Precision, Recall, RMSE, Latency, F-measure, MCC metric, ROC Area etc are used to evaluate the efficiency of filter based feature selection methods. The details regarding various clinical datasets used in our research are depicted in Table 3 to Table 6.

Table 3
Breast Cancer dataset details

Class:	no-recurrence-events, recurrence-events
age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
menopause	lt40, ge40, premeno
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
node-caps	yes, no
deg-malig	1, 2, 3
breast	left, right
breast-quad	left-up, left-low, right-up, right-low, central
irradiat	yes, no

Table 4
Diabetes dataset details

Class: tested_positive, tested_negative
Number of times pregnant
Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Diastolic blood pressure (mm Hg)
Triceps skin fold thickness (mm)
2-Hour serum insulin (mu U/ml)
Body mass index (weight in kg/(height in m)^2)
Diabetes pedigree function
Age (years)

Table 5
Hepatitis dataset details

Class:	DIE, LIVE
AGE	10, 20, 30, 40, 50, 60, 70, 80
SEX	male, female
STEROID	no, yes
ANTIVIRALS	no, yes
FATIGUE	no, yes
MALAISE	no, yes
ANOREXIA	no, yes
LIVER BIG	no, yes
LIVER FIRM	no, yes

(*contd...Table 5*)

SPLEEN PALPABLE	no, yes
SPIDERS	no, yes
ASCITES	no, yes
VARICES	no, yes
BILIRUBIN	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
ALK PHOSPHATE	33, 80, 120, 160, 200, 250
SGOT	13, 100, 200, 300, 400, 500
ALBUMIN	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
PROTIME	10, 20, 30, 40, 50, 60, 70, 80, 90
HISTOLOGY	no, yes

Table 6
Class distribution of healthcare datasets

Class Value	Number of Instances
Breast Cancer	
no-recurrence-events	201
recurrence-events	85
Diabetes	
tested_negative	500
tested_positive	268
Hepatitis	
DIE	32
LIVE	123

5.1. Classification With Original Healthcare Dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.746	0.588	0.750	0.746	0.748	0.158	0.623	0.790	no-reccurence-events
	0.412	0.254	0.407	0.412	0.409	0.158	0.623	0.410	reccurence-events
Weighted Avg	0.647	0.489	0.648	0.647	0.647	0.158	0.623	0.677	

```
  a   b   <-- classified as
150  51 |  a = no-recurrence-events
 50  35 |  b = recurrence-events
```

Figure 5: Performance evaluation metrics of Breast cancer dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.832	0.392	0.798	0.832	0.815	0.449	0.793	0.850	Tested negative
	0.608	0.168	0.660	0.608	0.633	0.449	0.793	0.667	Tested positive
Weighted Avg	0.754	0.314	0.750	0.754	0.751	0.449	0.793	0.786	

```
  a    b   <-- classified as
416   84 |  a = tested_negative
105  163 |  b = tested_positive
```

Figure 6: Performance evaluation metrics of Diabetes dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.563	0.138	0.514	0.563	0.537	0.411	0.823	0.531	DIE
	0.862	0.438	0.883	0.862	0.872	0.411	0.823	0.930	LIVE
Weighted Avg	**0.800**	0.376	0.807	0.800	0.803	0.411	0.823	0.848	

```
  a    b   <-- classified as
 18   14 |  a = DIE
 17  106 |  b = LIVE
```

Figure 7: Performance evaluation metrics of Hepatitis dataset

Table 7
Actual Breast cancer dataset results details

Total number of Instances	286
Correctly Classified Instances	185
Incorrectly Classified Instances	101
Classification Accuracy	64.68%
Root Mean Square Error	0.5423
Model Build up time	4.59 sec

Table 8
Actual Diabetes dataset results details

Total number of Instances	768
Correctly Classified Instances	579
Incorrectly Classified Instances	189
Classification Accuracy	75.39%
Root Mean Square Error	0.4215
Model Build up time	0.97 sec

Table 9
Actual Hepatitis dataset results details

Total number of Instances	155
Correctly Classified Instances	124
Incorrectly Classified Instances	31
Classification Accuracy	80%
Root Mean Square Error	0.4154
Model Build up time	0.54 sec

5.2. Classification With Correlation Based Method On Healthcare Dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.846	0.588	0.773	0.846	0.808	0.279	0.619	0.749	no-reccurence-events
	0.412	0.154	0.530	0.412	0.464	0.279	0.619	0.457	reccurence-events
Weighted Avg	0.717	0.459	0.701	0.717	0.705	0.279	0.619	0.662	

```
  a   b   <-- classified as
170  31 |  a = no-recurrence-events
 50  35 |  b = recurrence-events
```

Figure 8: Performance evaluation metrics of Breast Cancer dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.832	0.388	0.800	0.832	0.816	0.453	0.809	0.869	Tested negative
	0.612	0.168	0.661	0.612	0.636	0.453	0.809	0.676	Tested positive
Weighted Avg	0.755	0.311	0.752	0.755	0.753	0.453	0.809	0.802	

```
  a   b   <-- classified as
416  84 |  a = tested_negative
104 164 |  b = tested_positive
```

Figure 9: Performance evaluation metrics of Diabetes dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.594	0.089	0.633	0.594	0.613	0.517	0.863	0.632	DIE
	0.911	0.406	0.896	0.911	0.903	0.517	0.863	0.956	LIVE
Weighted Avg	0.845	0.341	0.842	0.845	0.843	0.517	0.863	0.956	

```
  a   b   <-- classified as
 19  13 |  a = DIE
 11 112 |  b = LIVE
```

Figure 10: Performance evaluation metrics of Hepatitis dataset

Table 10
Breast Cancer dataset results with Correlation method

Total number of Instances	286
Correctly Classified Instances	205
Incorrectly Classified Instances	81
Classification Accuracy	71.67%
Root Mean Square Error	0.4863
Model Build up time	1.96 sec

Table 11
Diabetes dataset results with Correlation method

Total number of Instances	768
Correctly Classified Instances	580
Incorrectly Classified Instances	188
Classification Accuracy	75.52%
Root Mean Square Error	0.4075
Model Build up time	0.62 sec

Table 12
Hepatitis dataset results with Correlation method

Total number of Instances	155
Correctly Classified Instances	131
Incorrectly Classified Instances	24
Classification Accuracy	84.51%
Root Mean Square Error	0.369
Model Build up time	0.32 sec

5.3. Classification with Information Gain based method on Healthcare Dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.866	0.671	0.753	0.866	0.806	0.226	0.646	0.797	no-reccurence-events
	0.329	0.134	0.509	0.329	0.400	0.226	0.646	0.505	reccurence-events
Weighted Avg	0.706	0.511	0.681	0.706	0.685	0.226	0.646	0.710	

```
a    b    <-- classified as
174  27  |  a = no-recurrence-events
57   28  |  b = recurrence-events
```

Figure 11: Performance evaluation metrics of Breast Cancer dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.862	0.410	0.797	0.862	0.828	0.472	0.806	0.868	Tested negative
	0.590	0.138	0.696	0.590	0.638	0.472	0.806	0.692	Tested positive
Weighted Avg	0.787	0.315	0.762	0.767	0.762	0.472	0.806	0.806	

```
  a   b   <-- classified as
431  69 |  a = tested_negative
110 158 |  b = tested_positive
```

Figure 12: Performance evaluation metrics of Diabetes dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.500	0.114	0.533	0.500	0.516	0.396	0.842	0.587	DIE
	0.886	0.500	0.872	0.886	0.879	0.396	0.842	0.945	LIVE
Weighted Avg	0.806	0.420	0.802	0.806	0.804	0.396	0.842	0.945	

```
  a   b   <-- classified as
 16  16 |  a = DIE
 14 109 |  b = LIVE
```

Figure 13: Performance evaluation metrics of Hepatitis dataset

Table 13
Breast cancer dataset results with Information Gain method

Total number of Instances	286
Correctly Classified Instances	202
Incorrectly Classified Instances	84
Classification Accuracy	70.62%
Root Mean Square Error	0.4677
Model Build up time	1.97 sec

Table 14
Diabetes dataset results with Information Gain method

Total number of Instances	768
Correctly Classified Instances	589
Incorrectly Classified Instances	179
Classification Accuracy	76.69%
Root Mean Square Error	0.4081
Model Build up time	0.61 sec

Table 15
Hepatitis dataset results with Information Gain method

Total number of Instances	155
Correctly Classified Instances	125
Incorrectly Classified Instances	30
Classification Accuracy	80.64%
Root Mean Square Error	0.3913
Model Build up time	0.29 sec

5.4. Classification with Consistency based method on Healthcare Dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.816	0.553	0.777	0.816	0.796	0.273	0.646	0.770	no-reccurence-events
	0.447	0.184	0.507	0.447	0.475	0.273	0.646	0.676	reccurence-events
Weighted Avg	0.706	0.443	0.697	0.706	0.701	0.273	0.646	0.676	

```
  a   b   <-- classified as
164  37 |  a = no-recurrence-events
 47  38 |  b = recurrence-events
```

Figure 14: Performance evaluation metrics of Breast Cancer dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.866	0.455	0.780	0.866	0.821	0.437	0.801	0.871	Tested negative
	0.545	0.134	0.685	0.545	0.607	0.437	0.801	0.671	Tested positive
Weighted Avg	0.754	0.343	0.747	0.754	0.746	0.437	0.801	0.801	

```
  a   b   <-- classified as
433  67 |  a = tested_negative
122 146 |  b = tested_positive
```

Figure 15: Performance evaluation metrics of Diabetes dataset

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.375	0.089	0.522	0.375	0.436	0.325	0.714	0.407	DIE
	0.911	0.625	0.848	0.911	0.878	0.325	0.714	0.902	LIVE
Weighted Avg	0.800	0.514	0.781	0.800	0.787	0.325	0.714	0.800	

```
  a   b   <-- classified as
 12  20 |  a = DIE
 11 112 |  b = LIVE
```

Figure 16: Performance evaluation metrics of Hepatitis dataset

Table 16
Breast cancer dataset results with Consistency method

Total number of Instances	286
Correctly Classified Instances	202
Incorrectly Classified Instances	84
Classification Accuracy	70.62%
Root Mean Square Error	0.5047
Model Build up time	3.28 sec

Table 17
Diabetes dataset results with Consistency method

Total number of Instances	768
Correctly Classified Instances	579
Incorrectly Classified Instances	189
Classification Accuracy	75.39%
Root Mean Square Error	0.4146
Model Build up time	0.6 sec

Table 18
Hepatitis dataset results with Consistency method

Total number of Instances	155
Correctly Classified Instances	124
Incorrectly Classified Instances	31
Classification Accuracy	80%
Root Mean Square Error	0.4264
Model Build up time	0.35 sec

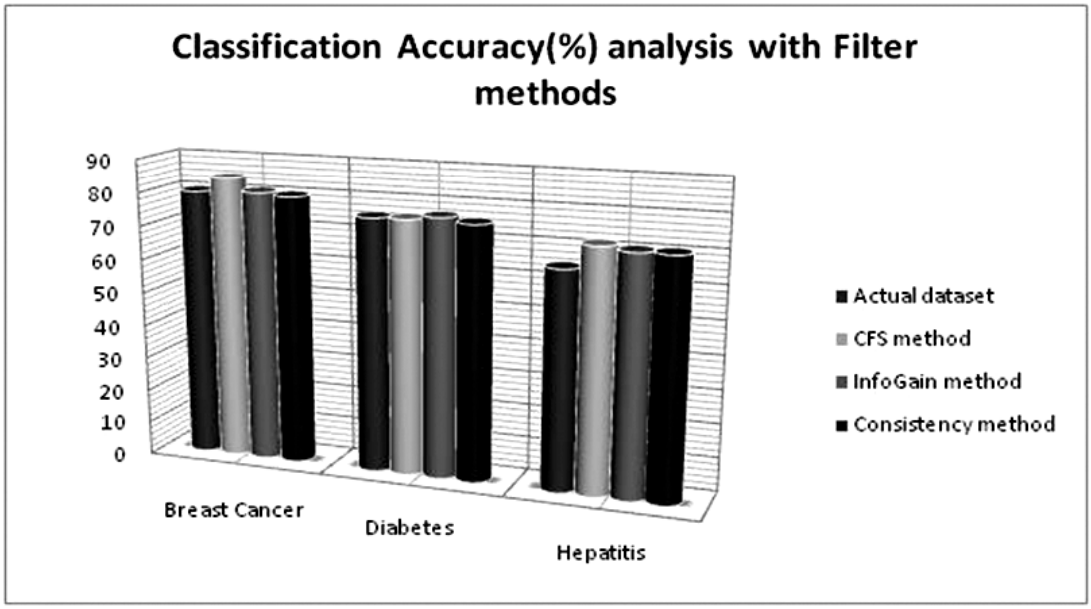

Figure 17: Classification Accuracy analysis with Filter based methods on Healthcare datasets

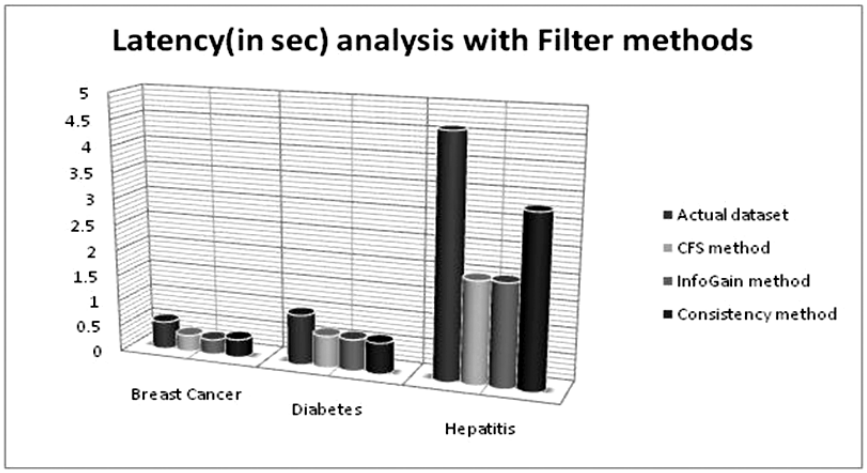

Figure 18: Latency Accuracy analysis with Filter based methods on Healthcare datasets

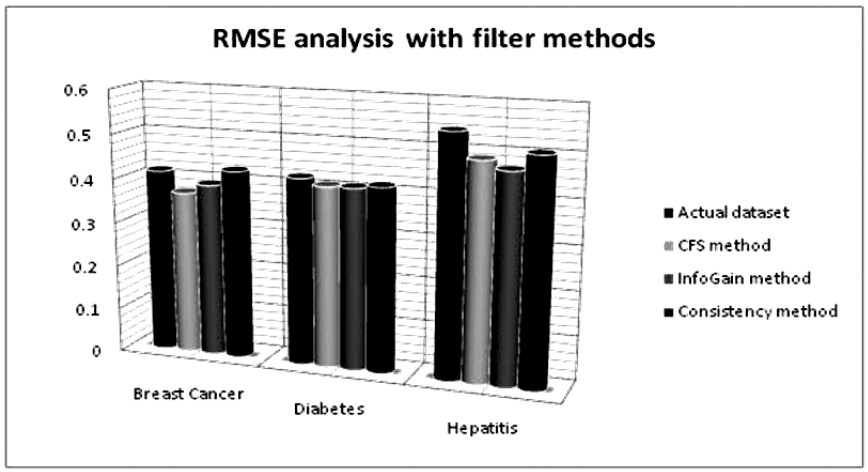

Figure 19: RMSE metric analysis with Filter based methods on Healthcare datasets

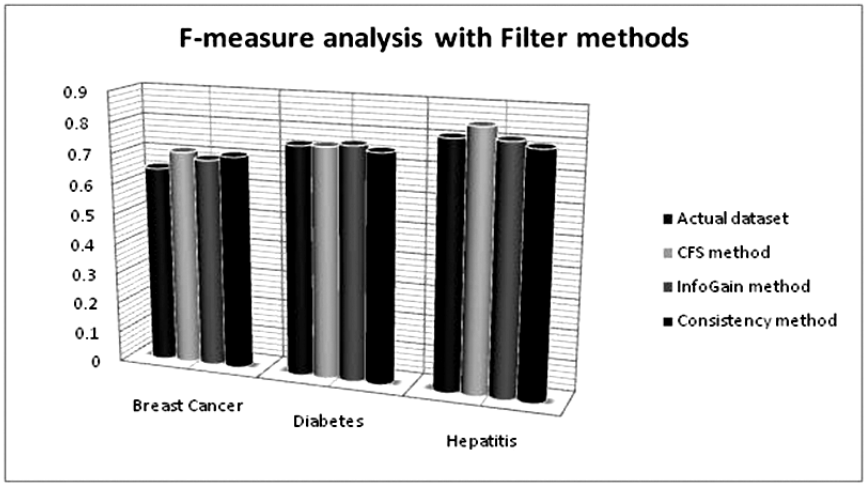

Figure 20: F-measure metric analysis with Filter based methods on Healthcare datasets

As per the observation from the graphs it is clearly visible that the overall effectiveness of disease risk prediction gets highly optimized and precise when classification is done with filter based feature selection methods. Rather classification with Correlation based feature selection yields an optimal performance in classification process in terms of classification accuracy, Latency, Root mean square error and F-measure metrics. The overall accuracy is optimum while the error rate is the least with Correlation based filter technique. The delay in disease risk prediction is very low in Correlation based filter technique thereby facilitating for real time applications. The F-measure value is also maximized if classification is undertaken with Correlation based filter technique.

6. CONCLUSION

Healthcare information systems comprise heaps of unstructured data records. Machine learning algorithms embedded with filter based attribute optimization methods helps to analyze and process such massive and noisy data efficiently. In our research, a comparative detailed analysis was carried out on the basis of three vital filter based feature selection algorithms to predict the risks of various diseases while their performance was computed by using Multilayer Perceptron classifier. The results were evaluated based on different performance measures. It was observed that using filter based techniques enhance the overall accuracy of classification in healthcare sector. Among the feature ranking methods Correlation based feature selection method outperforms other techniques in accurately predicting a disease risks when evaluated with various performance metrics. Thus our study asserted that filter based attribute optimization methods improve the performance of learning algorithms and more importantly Correlation based method can successfully act as a guide to healthcare experts in identifying disease risks. The results of this study can be successfully employed in the prediction and diagnosis of disease risks in medical research. As a future work, a study will be planned to investigate the impact of multi dimensional attributes of medical datasets in the performance of feature selection methods and classification accuracy. Besides a hybrid filter based meta-variable selection model can be developed in future.

REFERENCES

[1] Machine learning algorithms enable discovery of important regularities in large datasets November 1999/Vol. 42, No. 11 COMMUNICATIONS OF THE ACM

[2] [J. Novakovic, 11] J. Novakovic, P. Strbac, and D. Bulatovic, "Toward optimal feature selection using ranking methods and classification algorithms," Yugoslav Journal of Operations Research, vol. 21, no. 1, pp. 119-135, 2011.

[3] [SarvestanSoltani A, 11] SarvestanSoltani A., Safavi A. A., Parandeh M. N. and Salehi M., "Predicting Breast Cancer Survivability using data mining techniques", Software Technology and Engineering (ICSTE), 2nd International Conference, pp. 227-231, Vol.2, 2010.

[4] [Chang Pin Wei,] Chang Pin Wei and Liou Ming Der, "Comparison of three Data Mining techniques with Genetic Algorithm in analysis of Breast Cancer data", Available: http://www.ym.edu.tw/~dmliou/Paper/com par_threedata.pdf.

[5] [Gandhi Rajiv K, 10] Gandhi Rajiv K., Karnan Marcus and Kannan S., "Classification Rule Construction Using Particle Swarm Optimization Algorithm for Breast Cancer Datasets", Signal Acquisition and Processing, ICSAP, International Conference, pp. 233 – 237, 2010.

[6] [Jiawei Han, 00] Jiawei Han, Jian Pei, Yiwen Yin, " Mining frequent patterns without candidate generation", Proceedings of the 2000 ACM SIGMOD international conference on Management of data, p.1-12, May 15-18, 2000, Dallas, Texas, United States

[7] [V. A. Sitar-Taut, 09] V. A. Sitar-Taut.,"Using machine learning algorithms in cardiovascular disease risk evaluation", Journal of Applied Computer Science and Mathematics, 2009.

[8] [K. Srinivas, 10] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering (IJCSE), vol. 2, no. 2, pp. 250-255, 2010.

[9] [A. H. Roslina, 10] A. H. Roslina and A. Noraziah, "Prediction of hepatitis prognosis using support vector machine and wrapper method," in Proc. the Fuzzy Systems and Knowledge Discovery, Yantai Shandong, pp. 2209-2211, 2010.

[10] [J. S. Sartakhti, 11] J. S. Sartakhti, "Hepatitis disease diagnosis using a novel hybrid method," Computer Methods and Programs in Biomedicine, vol. 108, issue 2, pp. 570-579, 2011.

[11] [H. Harb, 14] H. Harb and A. S. Desuky, "Feature selection on classification of medical datasets based on particle swarm optimization," International Journal of Computer Applications, vol. 104, no. 5, pp. 14-17, 2014.

[12] [M. Ashraf, 13] M. Ashraf, G. Chetty, and D. Tran, "Feature selection techniques on thyroid,hepatitis, and breast cancer datasets," International Journal on Data Mining and Intelligent Information Technology Applications(IJMIA),vol. 3, no. 1, pp. 1-8, 2013.

[13] [M. Leach,12] M. Leach, "Parallelising feature selection algorithms," University of Manchester, Manchester, 2012.

[14] [I. T. Jolliffe,02] I. T. Jolliffe. (2002). Principal Component Analysis.[Online]. Available: http://books.google.com.au