CLASSIFICATION BASED HYBRID APPROACH FOR DETECTION OF LUNG CANCER

Supreet Kaur[1], Amanjot Kaur Grewal[2]

[1]Research Scholar, Punjab Technical University, Dept. of CSE, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India

[2]Assistant Professor, Punjab Technical University, Dept. of CSE, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India

ABSTRACT: *Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and Clustering. Clustering is most important process, if we have a very large database. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. In proposed work comparison has been done to distinguish the data of lung cancer using algorithms i.e. BFO, NN, SVM and LDA. Result simulations shows that NN has high accuracy having value 93% than other two LDA and SVM algorithms having values nearby 91% and 89% respectively. The whole result comparison has been done in MATLAB environment.*

KEYWORDS: Data Mining, SVM, NN, BFO, LDA, Data set.

1. INTRODUCTION

Data mining name is introduced in very late but the concept and technology is known since long time. Initially this concept was used manually in an organization but now we use these concepts using super computers over massive database tables, stored in flat files, spreadsheets, or some other storage format. Data mining is a mechanism to find appropriate patterns in a database, using distinct approach and algorithms to look into current and past data that can be analyzed to predict future behaviors [1]. Because data mining tools predict future veers and behaviors by reading through databases for hidden patterns, they allow organizations to make positive, knowledge-driven decisions and respond question that were previously too time-consuming to resolve. Mainly the Data Mining conceptions are used for the observation of any collected information, evaluate their behavior, and also find out unknown relationship between critical data in a variety of ways [2]. There are different types of Data Mining technology available and applied over the process for retrieving information from raw data and these technologies are used for different purposes like for security, marketing, and information gathering. This Data mining technology is used for verifying sample data, analyze their paths, to verify and validate their modules for specific general and business utilization. For example an insurance industry uses these techniques for finding out their rate of risks for their customers. In general application of data mining tools are in the area of marketing, fraud fortifications, and observations [3]. Different types of algorithms and tools are used for data evaluation and they are also providing different results depending on specific algorithms .The purpose of data mining is to discover valid novel, possibly helpful and understandable relationships and patterns in the existing data . Data mining is to extract useful information; this is also known with different names like knowledge extraction, information discovery, information harvesting, and data pattern processing. The name "data mining" is mainly used by database researchers, statisticians, and business communities [4, 5, 6].

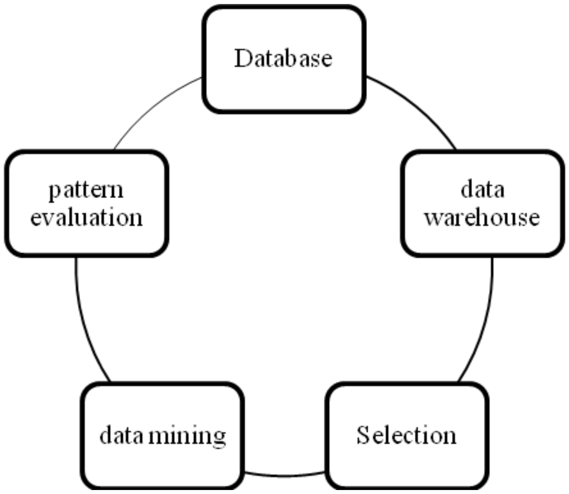


Fig.1 KDD Process in Data Mining

In proposed work three algorithms has been compared for clustering i.e. LDA, SVM and NN in MATLAB 2010a environment [7, 8, 9, 10].

2. SIMULATION WORK

To start with, different algorithms are chosen from data mining and implemented in a programming language. MATLAB will be the implementing language here. Each algorithm will be tested with different types of input data. I am applying different-different clustering algorithms and classification algorithms expect a useful product so as extremely supportive for the creative users and innovative researchers. Technique research effort is as.

Step: 1 Upload data
Step: 2 Reading of file
Step: 3 Training using BFO
Step: 4 Training using NN
Step: 5 Testing using SVM
Step: 6 Testing using LDA.
Step: 7 Find metrics
Step: 8 End

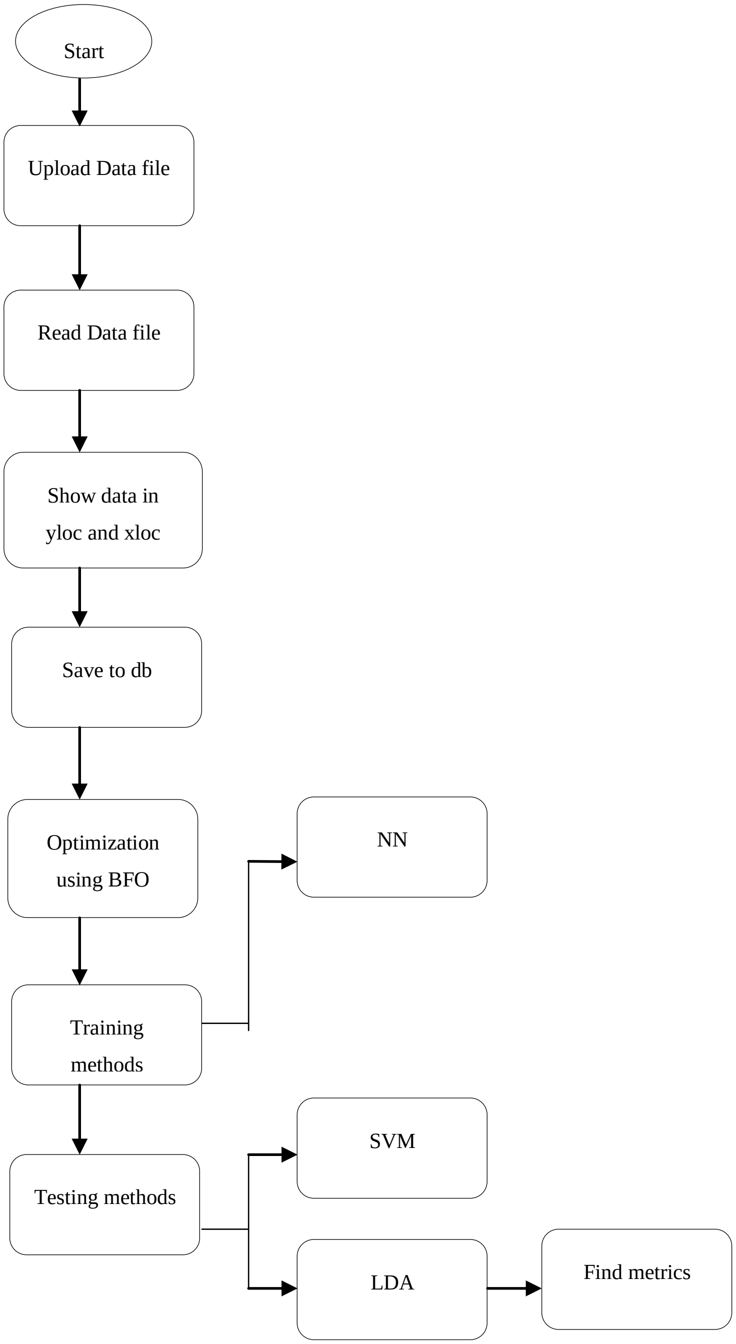


Fig.2 Proposed Flowchart

3. PROPOSED ALGORITHMS

3.1 Data Uploading Algorithm

```
select a file
creating the full path ;
reding the excel file; [data,num,raw]=xlsread(fullpath);%
Show all the data of the excel file
xlabel('Number of samples');
ylabel('Sample Value');
saving the data into a matlab database
Data Upload Successful
```

3.2 BFO Algorithm

```
Load main data
Get size of file
totalbacteria=r;
subbacteria=c;
totalreduced=0;
fori=1:totalbacteria
currentbacteria=data(i,:);
fitness_threshold=mean(data(i,:));
fitness_function=@bacteriafunction;
fit_bacteria=bacteriafunction(currentbacteria);
k=find(fit_bacteria==1);
totalreduced=totalreduced+numel(k);
 (k,currentbacteria(k),'r^');
trainingdata{i}=currentbacteria(k);
training done using BFO
```

3.3 Neural Network Algorithm

```
load all training
count=0;
for k=1:numel(trainingdata)
current=trainingdata{k};
if ~isempty(current)
count=count+1;
Tdata=current;
group=1:numel(Tdata);
net=newff(Tdata,group,20);
net.trainparam.epochs=100;
net=train(net,Tdata,group);
Training using NN done
```

3.4 SVM Algorithm

```
loadalltraining
count=1;
data=[];
for k=1:numel(trainingdata)
current=trainingdata{k};
if ~isempty(current)
data(count) = mean(current);F
count=count+1;
end
end
data=[max(current);min(current)];
```

```
xdata = data';
group = {'Unclassified' 'Classified'}';
svmStruct = svmtrain(xdata,group);
save('SVMTraining','svmStruct');
Training done using SVM
```

3.5 LDA Algorithm

```
loadalltraining
count=1;
data=[];
for k=1:numel(trainingdata)
current=trainingdata{k};
if ~isempty(current)
data(count) = mean(current);
count=count+1;pause(0.15)
end
end
data=[max(current);min(current)];
ldadata = data';
species={'Unclassified' 'Classified'}';
save('LDATraining','ldadata','species');
'LDA training done
```

4. RESULT ANALYSIS

This section contains the brief explanation of obtained results is given.

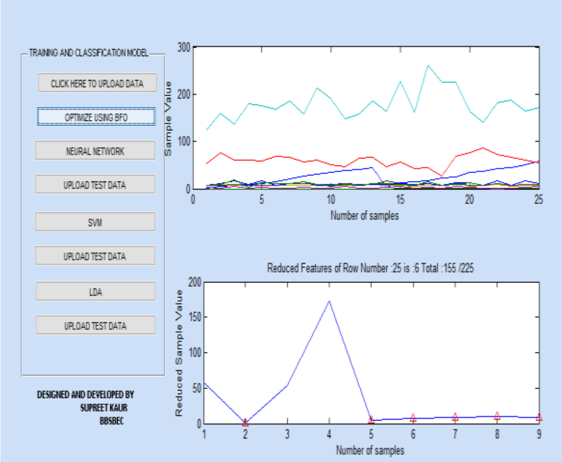
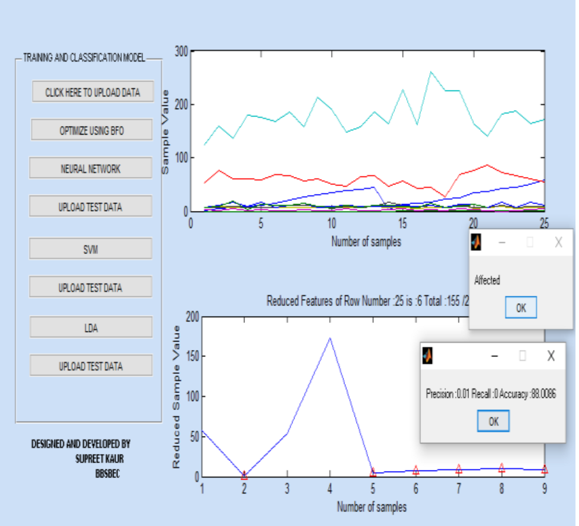


Fig.3 BFO optimization

Above figure 3 shows the BFO optimization using following function.

```
function[ f ]=bacteriafunction(bacteria_value)
threshold=mean(bacteria_value);
z=numel(bacteria_value);
f=zeros(1,z);
fori=1:z
ifbacteria_value(i)<threshold
f(i)=1
end
```

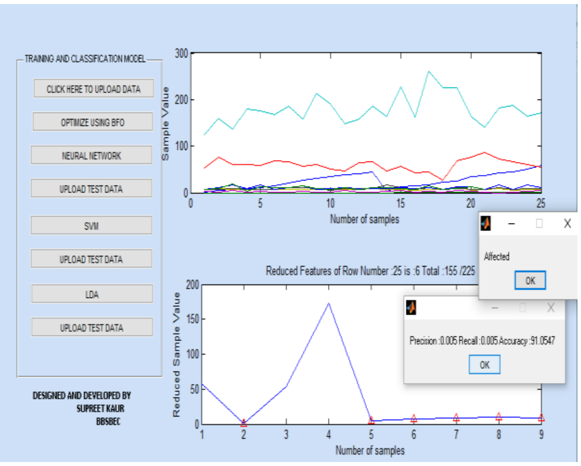


Affected

Precision :0.01 Recall :0 Accuracy :88.0086

Fig.4 Classification using SVM

As above figure 4 shows that for training SVM has been used. And dialog box appeared shows it is saved in the database. The main idea of SVM is that; it finds the optimal separating hyper plane such that error for unseen patterns is minimized and the obtained value using SVM is 89 %

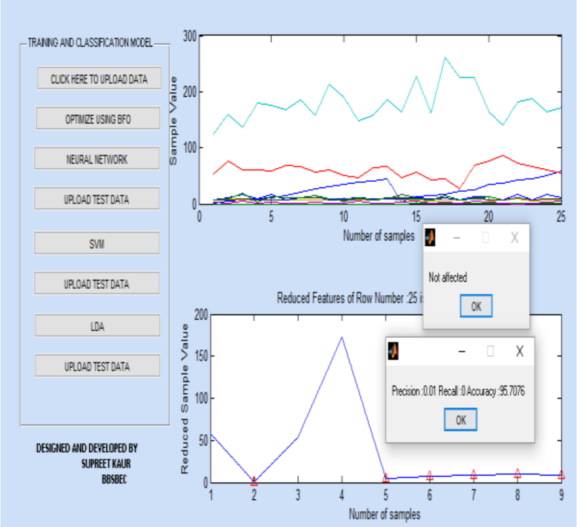


Affected

Precision :0.005 Recall :0.005 Accuracy :91.0547

Fig.5 Classification using LDA

Above figure 5 shows the training as well as testing using LDA method in which main factor that helps to reach that 91 accuracy is training algorithm.



Not affected

Precision :0.01 Recall :0 Accuracy :95.7076

Fig. 6 Classification using NN

Above figure 6 shows the training as well as testing did using neural network. From above results its has been concluded that achieved accuracy is 93.25%.

Table 1: Result Comparison

Technique	Recall rate	Precision rate	Accuracy
NN	0	.01	93.25
SVM	.01	0	89
LDA	.005	.005	91

Table 1 shows the various values of the techniques by considering Recall rate, Precision rate and accuracy. The average of recall rate is 0.005; precision rate has average of 0.005 and the accuracy has the average of 91.083.

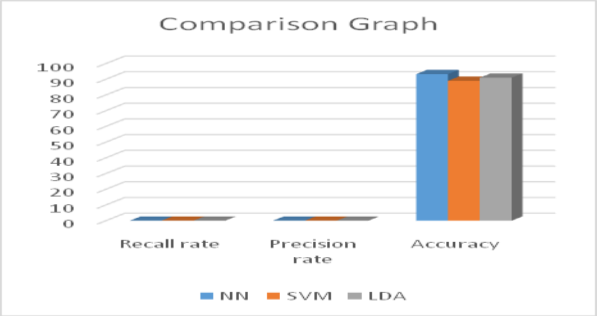


***Figure** Error! No text of specified style in document..8 Comparison Graph*

Above graph compare the recall rate, precision rate and accuracy of NN, SVM and LDA. From results it has been concluded that NN has provided best results having 91.083 accuracy.

5. CONCLUSION

In this proposed work, papers have proposed a system in which we utilized clustering as well classification algorithms to shows there accuracy on lung cancer dataset. In proposed work BFO utilization is done for optimization, NN used for training and SVM, LDA used for classification. From results it has been concluded that NN has provided best results having 93.25 % accuracy.

REFERENCES

[1] V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra. 2013, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," International Journal of Computer Science and Information Technologies, Vol. 4 (1), 2013, 39 – 45.

[2] Ankit Agrawal, SanchitMisra, Ramanathan Narayanan, LalithPolepeddi, AlokChoudhary, "A Lung Cancer Outcome Calculator Using Ensemble Data Mining on SEER Data," BIOKDD 2011, August 2011, San Diego, CA, USA, 2011.

[3] R.Linder, T.Richards and M. Wagner,"Microarray data classified by artificial neural networks," Methods In Molecular Biologyclifton Then TotowA-, 382:345, 2007.

[4] S. Xuejun, Q. Wei, and S. Dansheng, "Three-class classification in computer-aided diagnosis of breast cancer by support vector machine," Proceedings SPIE Med. Imag., vol. 5370, pp. 999–1007, 2004.

[5] Y. Song, W. Cai, S. Eberl, M. J. Fulham, and D. Feng, "Automatic detection of lung tumor and abnormal regional lymph nodes in PET-CT images," J. Nucl. Med., vol. 52, no. Suppl. 1, pp. 211, 2011.

[6] H. Cui, X. Wang, and D. Feng, "Automated localization and segmentation of lung tumor from PET-CT thorax volumes based on image feature analysis," in Proc. IEEE EMBC, 2012, pp. 5384–5387.

[7] Chen Han-ning, Zhu Yun-long, Hu Kun-yuan, Cooperative bacterial foraging algorithm for global optimization,2009 Chinese control and decision conference, Guilin, China, 2009, pp. 3896-3901.

[8] S. B. Patil and Y. S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450-216X, EuroJournals Publishing, Inc., vol. 31, no. 4, (2009), pp. 642-656.

[9] D. Glotsos, J. Tohka, P. Ravazoula, D. Cavouras, and G. Nikiforidis, "Automated diagnosis of brain tumours astrocytomas using probabilistic neural network clustering and support vector machines," Int. J.Neural Syst., vol. 15, pp. 1–11, 2005.

[10] P. Campadelli, E. Casiraghi, and G. Valentini. "Lung nodules detection and classification," In Proceedings. IEEE Int. Conf. Image Processing (ICIP2005), pp. 1117-20, September, 2005.