A Probabilistic Neural Network Approach for Classification of Datasets Collected From North Coastal Districts of AP, India Using MatLab

T. Panduranga Vital\textsuperscript{a,}\textsuperscript{*}, G.S.V Prasada Raju\textsuperscript{b}, K.Sreeramamurthy\textsuperscript{c}, V.P.Venkata Charan\textsuperscript{a}

\textsuperscript{1}Department of Computer Science and Engineering &I.T., Raghu Engineering College, Visakhapatnam, India.
\textsuperscript{2}Department of Computer Science, Andhra University, Visakhapatnam, India.
\textsuperscript{3}Department of IT, Sreenidhi Institute of Science and Technology, Yamanampet, Ghatakeswar, Hyderabad.

Abstract
Data mining is an important tool to analyze the data for diseased datasets. The data has been collected from North coastal districts of AP, India during 2011 to 2014 with 504 instances and 56 attributes. The methods like Confusion matrix, ROC, Best validation performance, R value, SOM Topology, Hits, SOM Neighbor Connections, Neighbor weight distances and SOM weight positions were analysed using MatLab version 7.6.0 (R2008a) from the collected dataset from north coastal districts of AP, India in the present study.

Keywords: Data mining; Cancer; PNN; North coastal districts of AP, India

1. Introduction:
Data mining is the process of extracting the patterns from data and converting to useful information [1, 2]. Data mining is becoming an increasingly important tool in the present decades to transform the data into information [3]. Analysis of cancer datasets is one of the important research in data mining techniques [4, 5]. Data mining softwares like MatLab has a number of analytical tools used for analyzing information from different outlooks like machine learning and database systems, summarizing the data into useful information [6,7]. The analysis and summarization can be used to increase accuracy of the data. Researchers in many fields like computer sciences,
communication networks, business management and biology have shown great interest in data mining.

Data mining in Cancer research is one of the important research topics in biomedical science. Bioinformatics was provided logic due to availability of bigdata for developing novel data mining methods [8]. Biomedical research applies a wide range of designs from various questionnaires from patients to solve problems in laboratory, clinical, and population settings.

2.Methodology:

The main aim of processing the data in the present experimentation is to discriminate healthy people from those with cancer with a two-decision classification problem. Matlab version 7.6.0 (R2008a) was applied in the present experimentation for analysis of network approach. The methods like Confusion matrix, ROC, Best validation performance, R value, SOM Topology, Hits, SOM Neighbor Connections, Neighbor weight distances and SOM weight positions were analysed from the collected dataset from north coastal districts of AP, India in the present study.

The data has been collected from North coastal district of AP, India during 2011 to 2014 with 504 instances and 56 attributes (Place, Age, Cancer Type, Family History, Drinking, Smoking, Tea, Coffee, Job, Perfumes, Soap, Tel: +91 9491936489 E-mail: vital2927@gmail.com

Shampoo, Morning Eat, Lunch Eat, Dinner Eat, Travel, Living, Fruits, Vegetables, Flowers, Sleep, Wake Up, Tensions, Cooking, Cool Drinks, Icecream, Study, Height, Weight, Bp, Pains, Hair Loss, Gutkha, Marital, Milk, Bloodgroup, Treatment Type, Bathing, Oils, Fastfood, Otherdisease, Morningwalk, Usemobile, Born, Drug, Treatment_Mode, Diagnosis, Meditation, Games, Mosquitorepellents, Injuries, Speaklevel, Seetv, Think And Phone).

Source Code:
clc; close all; clear all; clc;
q = csvread ('C:\Users\User\Documents\MATLAB\cancerAndHealth.csv');
q = q';
for i = 1:57
    c1(i)= 1;
end
for i=1:951
    c2(i)= 2;
end
tc = [c1 c2]
target = ind2vec(tc)
net=newpnn(q,target)
y=sim(net,q)
yc=vec2ind(y);
plotroc(target,y);
plotconfusion(target,y);
3. Results:

To analyze the probability neural network response, confusion matrix is computed by considering the outputs of the trained network and comparing with the expected results (targets), shown in Fig. 1.

![Confusion Matrix](image1.png)

Fig. 1. Confusion matrix

The diagonal cells show the number of true sets that were correctly classified for each class of patients. The off-diagonal cells show the number of residue positions that were misclassified. The following results presents the accuracy obtained by training the probabilistic neural network using dataset and got 100% of data for training as positives (correctly classified) using Matlab.

The Receiver Operating Characteristic (ROC) curve, a plot of the true positive rate (sensitivity) versus the false positive rate (1 - specificity) is also drawn and shown in Fig. 2.

![ROC](image2.png)

Fig. 2. ROC

Fig. 3 was shown best validation performance as 0.036.
Fig. 4 was shown the R value as 0.692.

The training of Neural Network (nntrain tool) at 200 iterations taken 45 seconds and the SOM topology, Hits, SOM Neighbor Connections, SOM Neighbor weight distances and SOM weight positions was presented in Fig. 5 to 9 respectively.
Fig. 5. SOM Topology

Fig. 6. Number of Hits

Fig. 7. SOM Neighbor Connections
Classification, clustering and association are very important among the techniques of data mining [9, 10]. Accuracy is the main objective to estimate the performance of the algorithms over cancer datasets [11, 12].

4. CONCLUSION:

The PNN method using Matlab was provided data validation as 100% correct classification. The probability approach was provided good results for cancer datasets and further analysis can provide understanding for diagnosis and cure of cancer.

Acknowledgment
The authors would like to thank Andhra University, GITAM University and Raghu Engineering College for providing computational facility and access to e-journals to carry out this research.

References