International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 6 (2013), pp. 479-484
© International Research Publications House
http://www. irphouse.com /ijict.htm

Evaluating Bioinformatics Dataset Accuracy using Pattern Classification

Manu Banga

Department of Computer Science & Engineering,
Raffles University, Neemrana, Rajasthan.

Abstract

Pattern Classification is the science of making inferences from perceptual data, using tools from statistics, probability, computational geometry, machine learning, signal processing, and algorithm design. It is a supervised technique in which the patterns are organized into groups of pattern sharing the same set of properties and to solve the classification problem at the attribute level and return to an output space of two or more than two classes. Probabilistic Neural Networks (PNN) and K-Nearest Neighbors are the effectively used. They uses training and testing data samples to build a model but with PNN it is very difficult to handle huge data so K-Nearest Neighbors (KNN) algorithms are used to improve the performance accuracy and the convergence ratebut the computational cost becomes expensive so genetic algorithms are used to design a classifier in which samples are divided into different class boundaries and for each generation accuracy of algorithm continued till we get our desired accuracy or desired number of generation. In this paper, comparative study of Probabilistic Neural Network, K-Nearest Neighbors and Genetic Algorithm as a classifier is done. We have tested these different algorithms using instances from lung cancer dataset and Iris dataset taken from UCI repository and efficiency of three techniques are compared on the basis of the performance, convergence time and on the implementation complexity.

1. Introduction

Classification is the process in which we use some of the data to categorize the present dataset into different classes for future use. Classification is done in such a way that the properties of each data samples in a particular class are similar to each other. All approaches known use the knowledge of the already present datasets. Usually a subset of the present dataset is used as the training sample for the classification technique. Different data mining approaches can be used forclassification.Probabilistic Neural Network is a type of neural network which can classify any number of input patterns to any number of classifications. A PNN learns more quickly than any other Neural Network. PNN is as a supervised neural network that can be used in system classification and pattern recognition. K Nearest Neighbor can be considered as statistical learning algorithms and it is extremely simple to implement and leaves itself open to a wide variety of variations. Here the training sample stores the data points which are given to it and then when the sample dataset is given to it, it finds the closest training data sample according to some distance formula(sometimes it is Euclidean distance) and gives the category of the closest data sample as the class for the sample dataset. Genetic Algorithm is a model inspired by evolution. Here an initial population is selected which can be considered as a solution to a problem and then different process of crossover and mutation is applied which keeps most of the next population having characteristics of their previous generation and improvised. Genetic algorithm as a classifier is used to obtain the class boundaries and once the training samples are divided in different class boundaries then the sample dataset can be tested and depending upon the region it belong its class can be found.

2. Approach Used:

2.1 Probabilistic Neural Network

It is a supervised classifier that can map any number of input to any number of output. We give a subset of our data set to it as an input and the data is then processed through various layers. It was derived from the Bayesian network and a statistical algorithm Kernel Fisher discriminant analysis. Similar to other neural networks it also has different layers. Probabilistic neural network (PNN) is closely related to Parzen [4] window probability distribution function estimator. A PNN consists of several smaller sub-networks; all of them are a Parzen window probability distribution function estimator for each of the classes.

The different layers of Probabilistic Neural Network are:

- **Input Layer:** It consists of the input to the neural network the training sample whose class has to be found out.
- **Pattern Layer:** It consists of the Gaussian function with the training samples are the center to it. We then find the Gaussian distance for all of them using the formula.

- **Output Layer:** Here all the outputs from the summation layers are taken and a selection of the largest value is done and that particular class of the selection determines the class of the sample.
- **Summation Layer:** Here all the output of the pattern layer is received and then they are added and send to the output layer for class determination. It is also the hidden layer in the Neural Network. A general structure of a probabilistic neural network is given below:

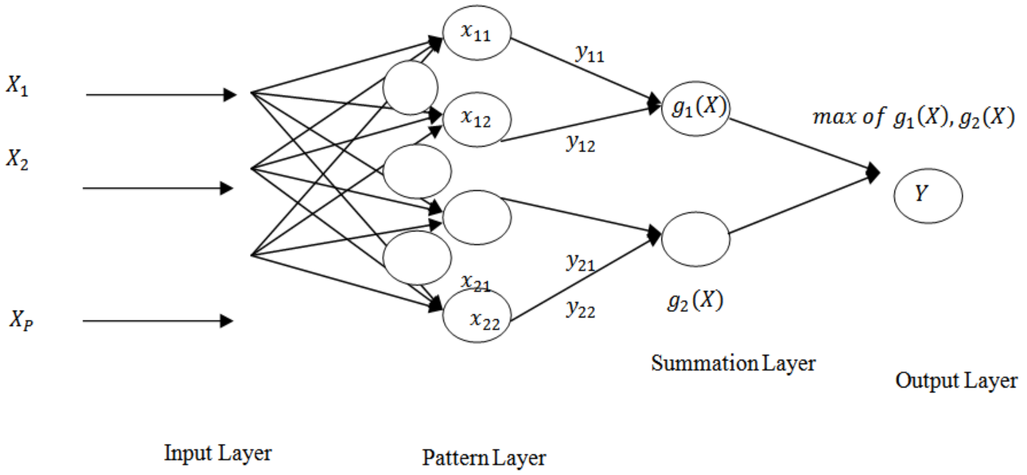

Figure 1: Probabilistic Neural Network.

2.2 K-Nearest Neighbour

It is one of the most trivial classification algorithms which use the neighbor characteristics to determine the class of the sample data. It is also a supervised classification technique where we need to give a particular subset of the dataset to examine the rest of the data set.KNN is a non-parametric lazy learning algorithm. By non-parametric, we mean that it does notmake any assumptions on the given data distribution. This characteristic of KNN is very usefulin real world because in real world most of the data doesn’t obey any theoretical assumptionsmade (Gaussian mixtures, linearly separable, etc.).It works on the principal where we find the Euclidean distance from all the neighboring elements and the Kthnearest neighbor is taken as the class for the given sample data. All the sample data represent to points in an n-Dimensional space. Nearest neighbor for each dataset is found using the Euclidean distance. The target function can be either real valued or discrete in nature. For discrete-valued, the k-NN returns the most common value among the k training examples nearest to the given sample. Vonoroi diagram: the decision surface induced by 1-NN for a typical set of training examples.

2.3. Genetic Algorithm:

Genetic algorithm can be used because they are more robust then the general AI techniques. They can take a large amount of input data as their input and can handle a large amount of data. Unlike other techniques Genetic Algorithm has a tendency to even adjust with a data which is noisy and have inconsistency. Genetic Algorithm although random uses the information from previous generations to improve the search. It is based on four concept of selection, crossover, mutation and acceptance.

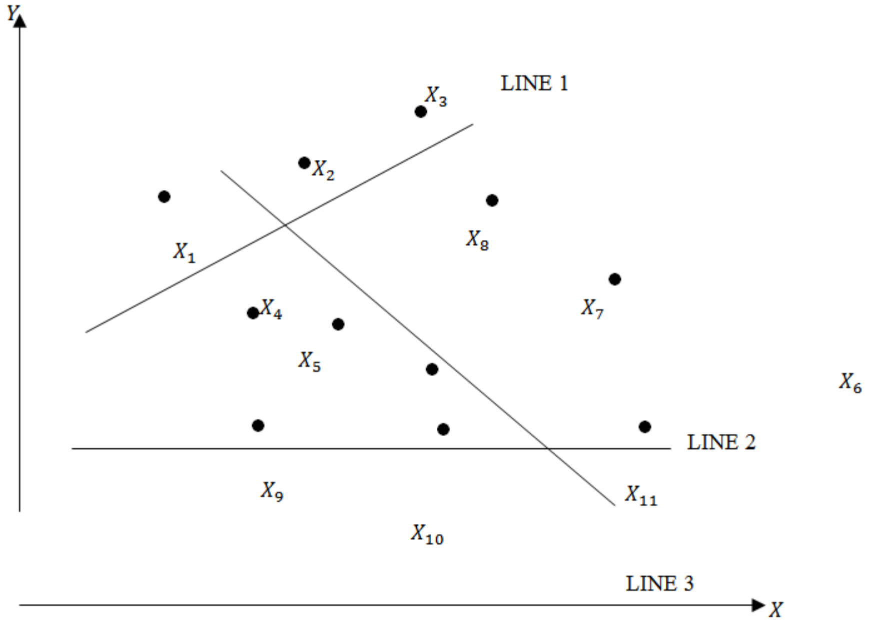

Figure 2: Dividing the 2-D space.

3. Result and Conclusions:

All the datasets were taken from the UCI repository. As they were raw data, first cleaned and then normalized the data to be used in our algorithm. Different datasets used are: Lung cancer data setIt has 57 Attributes and 32 data samples which are divided into three class. Class I: 9 samples, Class II: 10 samples, Class III: 13 Samples and on Iris Data set which has 5 attributes and 150 data samplesthe experiment was performed on matlab on which we get 81.5% accuracy using genetic algorithms and reason for lost in accuracy due to mutation process

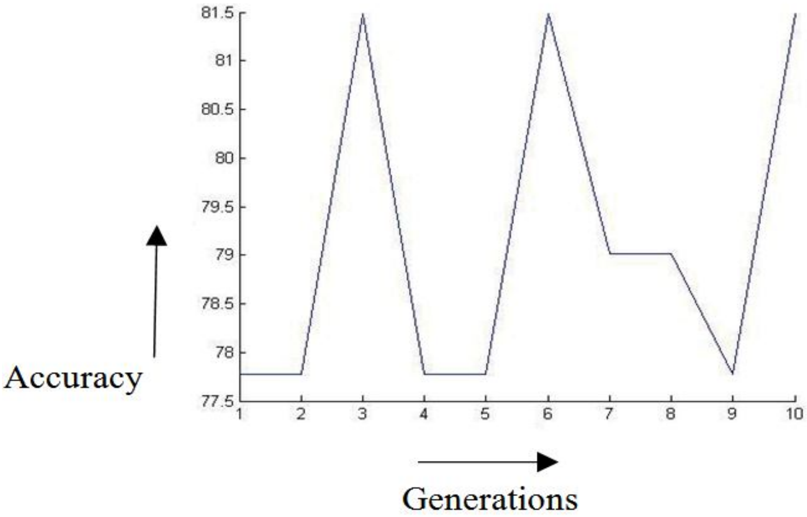

Figure 3: Showing Accuracy v/s Generation.

4. Conclusion

For a smaller data sample PNN works efficiently as compared to the other two techniques. But as the size of the training dataset increases the network complexity of PNN grows as a result computation takes too much time. The same problem also arises when KNN as a classifier is used for large training and test dataset. But from the above implementations Genetic Algorithm is efficient when used as a classifier as it divides the n-dimensional plane into various parts thereby classifying the dataset into respective classes. Overall GA classification technique is effective in classification of data samples. It is observed that for a large number of H (lines) the accuracy of the GA classifier further increases

5. Acknowledgement

I would like to thank Dr. S.C Aggarwal, Principal School of Engineering, for his valuable supportin carrying this work.

References

[1] Ibrahiem M.M.E.I Emary and S. Ramakrishnan, “On the application of various probabilistic neural networks in solving different pattern classification problems", World Applied Sciences Journal, 4(6):772 780, 2012.

[2] Leila FallahAraghi, Hamid Khaloozade , Mohammad Reza Arvan , “Ship Identification using probabilistic neural networks”, Proceedings of the International Multi Conference of Engineers and Computer Scientists, 2009 Vol II , IMECS 2009, March 18 - 20, 2009, Hong Kong.

[3] Khalafkhatatneh, Ibrahiem M.M El Emary and Basem Al- Rifai, “Probabilistic Artificial Neural Network for Recognizing the Arabic Hand Written Characters", Journal of Computer Science, 2 (12): 879-884, 2006 ISSN 1549-3636.

[4] R.O.Duda, P.E.Hart, D.G.Stork, “Pattern Classification”, 2nd Edition, Wiley-Interciencepublication,John Wiley & Sons, Inc, 2006.