A COMPOSITE MODEL THAT OVERWHELMS OUTLIERS TO ANTICIPATE THE STAGES IN RENAL DISORDER

Raju R[1], Kayalvizhi Devakumaran [2], Aishwarya Balasubramanian[3] and Gayathiri Balan [4]

[1]Assistant professor, B.Tech Information Technology, Sri Manakula Vinayagar Engineering College, Pondicherry, India
rajupdy@gmail.com
[2,3,4] B.Tech Information Technology, Sri Manakula Vinayagar Engineering College,Pondicherry, India
dp.kayalvizhi@gmail.com
mail2aishu0413@gmail.com
gayu.cadbury@gmail.com

ABSTRACT

The emerging concepts of Artificial Neural Network are used to recognize patterns, manage data and learn. ANN has its own significance in the field of medicine. This paper provides support in diagnosing and categorizing the different stages of the renal disorder. We put forth the idea of defining a Hybrid model which is a combination of Gaussian Mixture Model with the Expectation Maximization algorithm and Adaboost technique. The GMM is used for clustering of similar data and the EM algorithm is used for allocation of the weights based on the input parameters. The Boosting technique – Adaboost is being handled in order to boost the classifier into a stronger one. The drawback of handling noisy data i.e. outliers is being resolved by using this hybrid model. This mixture model is well suited for handling the problem of outliers. The system is trained with 75% of the renal samples and the remaining samples are used for testing. As a result, the system provides a probabilistic function which fits the sample in any one of the classified GMM's.

KEYWORDS

ANN – Artifical Neural Network, GMM – Gaussian Mixture Model, EM – Expectation Maximization, Adaboost, Outliers.

1. INTRODUCTION

Artificial neural networks are inspired by the functioning of the brain. They can be used to recognize patterns, manage data and learn [1]. ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the

DOI : 10.5121/cseij.2012.2301

learning phase. The greatest advantage of ANNs is their ability to learn from observed data. However, using them is not so straightforward and a relatively good understanding of the underlying theory is essential. It requires proper choice of model which depends upon the data being used and the application for which it is being used, proper learning algorithm that will be best suited for the application being utilized [1]. The robustness depends on both the appropriate selection of the model and the learning algorithm. In real time ANN has a wide diversity of applications such as gaming and decision making, pattern recognition, medical diagnosis and financial applications. In medical field they find their significance in diagnostic systems, biomedical analysis, image analysis, drug development. In this paper we tend to focus on diagnosis regarding renal disorders.

The primary role of the kidney is to remove metabolic waste and to balance the water and electrolyte levels in the blood [9]. The kidney also plays a major role in regulating levels of various minerals such as calcium, sodium, and potassium in the blood [8]. Renal disorder can be classified into three stages. The stage I is the start of renal disorder i.e. slightly diminished functioning of kidney with abnormalities in blood or urine. The stage II is the chronic renal failure where the functioning deceases gradually over time. The stage III is the end stage renal failure where the patient must undergo dialysis or must plan for a transplant for their survival.

The system involves classification based on the different stages of the renal failure, for which the Gaussian Mixture Model is being utilized. This model is used in clustering similar data into a single entity. Each stage of the renal disorder is classified into 3 stages [9] which are represented as GMM 1, GMM 2 and GMM 3. Even though the system has been categorized into 3 GMM it still remains to be a weak classifier. It can be considered as a weak classifier due to the fact that the system can also be exposed to inaccurate data. This problem can be resolved by boosting the system in order to make it as a strong classifier.

Adaboost is one of the most prominent boosting techniques among the ensemble learning to convert a weak classifier into a strong classifier [2]. The greatest advantage of using adaboost technique is that, it can be combined with any different classifier [6]. Adaboost is termed to be adaptive in the sense that subsequent classifiers built are tweaked in favor of those instances misclassified by previous classifiers. Adaboost overcomes the problem of over fitting [3]. But still it holds back on noisy data i.e. the problem of outliers [4].

2. RELATED WORK

This section briefly describes about the various concepts in the training process such as Gaussian Mixture Model with Expectation Maximization Algorithm and Adaboost technique.

2.1 Gaussian Mixture Model(GMM)

Gaussian Mixture Model is considered as one of the methods for clustering that helps in building soft clustering boundaries [7]. It can also be referred as a probability density function which considers the weights. The parameters for Gaussian Mixture Model are estimated by training the samples using the iterative Expectation-Maximization Algorithm. A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMM are commonly used in biometric system.

2.1.1 GMM-EM Algorithm

The Gaussian Mixture Model with Expectation-Maximization Algorithm (GMM-EM) is an efficient model for solving the parameter estimation problems [7]. The EM algorithm alternates between finding a greatest lower bound to the likelihood function(the “E Step"), and then maximizing this bound (the “M Step"). The EM algorithm is an iterative method for finding the maximum likelihood which involves two steps

- Expectation step
- Maximization step.

The Expectation (E) step computes the current weight estimate for the parameter [5]. The Maximization (M) step computes parameter that maximizes the weights estimated on the E step. The EM algorithm steps are discussed in detail as follows,

(i) *Initialization*: The distribution parameters such as mean and variance are evaluated at each of the GMM levels.

(ii) *Expectation*: The evaluation of weight factors of each sample is done based on the current parameter values.

(iii) *Maximization*: Here the process of re-estimation of the parameters using the weights calculated in the previous step is done.

(iv) *Iterate*: Here the likelihood is re-evaluated and checked for accuracy, if the change is less than the considered threshold then the parameters are set else the process iterates.

2.2 Adaboost

Boosting is one among the ensemble system for decision making. Seeking an additional opinion before making a sound decision is an innate behavior prevailing in the medical domain. In general a classifier or an expert is used to make a decision by picking any one of the option from the previously set of options [1]. In order to provide a sound decision making system a variety of boosting techniques are available. This paper puts forth the concept of using Adaboost technique.

Adaboost is an adaptive boosting technique that improves the classification accuracy [6]. It can be used with different classifiers in order to obtain a strong classifier [2]. The main significance of using this technique is that it overcomes the problem of over fitting [3]. It can use data that is textual, numeric, images, etc. The adaboost algorithm converges to the logarithm of the likelihood ratio. Adaboost is sensitive to outliers.

Outliers can be termed as those points which seem to deviate from other samples from a given data set. Outliers can also be termed as an extreme observation in a given dataset. They may also include sample maximum or the sample minimum. However sometimes they seemed to be invalid because the sample maximum and minimum may always not be far from the other observations. This drawback is being resolved by our proposed system [4].

3. PROPOSED WORK

The proposed system helps the physicians in diagnosing the current stage of renal disorder based on the patient's vital records which serves as input to the system [8]. The sample data required for training the system is obtained from Apollo Specialty Hospital, Chennai, India. The input parameters considered for the system are as follows,

1) Urea
2) Creatinine
3) Potassium
4) Sodium
5) Uric acid
6) Phosphorus
7) Protein Total
8) Albumin.

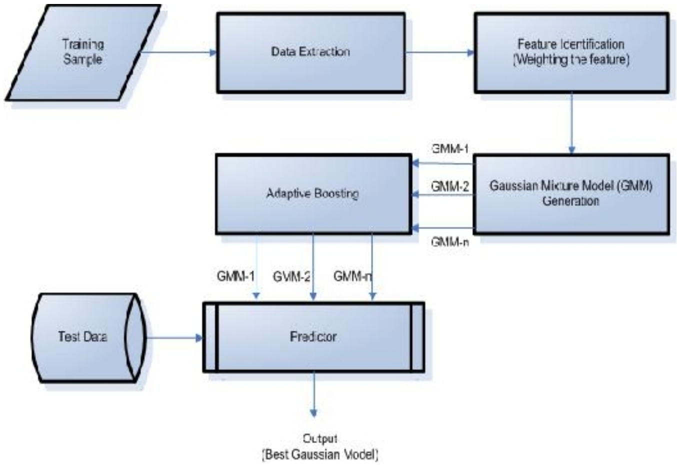

Training Sample
Data Extraction
Feature Identification (Weighting the feature)
GMM-1
GMM-2
GMM-n
Adaptive Boosting
Gaussian Mixture Model (GMM) Generation
GMM-1
GMM-2
GMM-n
Test Data
Predictor
Output (Best Gaussian Model)

Fig 1: Architectural Design

The architectural framework of the proposed system is as shown in the Fig 1.The functioning of the above depicted system is described briefly in two phases namely the Training phase and the Testing phase.

3.1 Training Phase

75 % of the sample data are subjected to training which undergoes the following process. Firstly the process of data extraction from the dataset takes place [9]. Feature identification [5] is done from the extracted data. The identified features are subjected to the GMM generator for the purpose of classification [7].

The mean (M), variance (V) and weight (W) for the model (λ) is computed with the k-means clustering technique. The initial log likelihood is calculated to enhance the computational speed.

The formula for Log Likelihood (L) is given in (1),

```
L=L+W(i)*(-0.5*n*log(det(2*pi*V(i))-0.5*(n-1)*(trace(iV*S)+(U-M(i))'*iV*(U-M(i))));
```

Where,

L - Log Likelihood
X - Dataset
W-Weight
M - Mean
V-Variance
U-Mean(X)
S-Covariance of X (1)

The formula for Mean (M(i)) is given in (2),

```
M(i) = E(j,i)*X(j);
```

Where,

X-Dataset
E- Expectation (2)

The formula for Expectation (E) is given in (3),

```
E = ( exp(-0.5*dXM'*iV(j)*dXM)/(a*S(j))) * W
```

Where,

W -Weight
dXM - Difference between mean (initial) and weight

S(j) = sqrt(det(V(j))); (3)

The formula for Variance (V(i)) is given in (4),

```
V(i) = E(j,i)*dXM*dXM' / W;
```

Where,

W - Weight
dXM - Difference between mean (initial) and weight

E-Expectation (4)

Expectation-maximization (EM) algorithm is a method that iteratively estimates the likelihood [10]. It is similar to the k-means clustering algorithm for Gaussian mixtures since they both look for the center of clusters and refinement is done iteratively.

For a given data set of X, the best fitting model (λ) is the one that maximizes. The algorithm of Expectation and Maximization is as follows

1. *Given an initial model (λ) with mean (M), variance (V), weight (W) and Log Likelihood (L).*
2. *Find the expectation, E for the model (λ) with M, W, and V*
3. *Compute the new model (λ') using the expectation, E.*
4. *Calculate the likelihood (L') for the new model (λ').*
5. *Repeat Step2, if abs (L'-L)/L' > 0.1 & iteration count 1000.*
6. *The best fitting Gaussian Mixture Model (λ) is arrived for the data set*

Here, the samples are classified into 3 different models based on the functioning of the kidney.

GMM 1 - The initial stage is the slightly diminished functioning of kidney with abnormalities in blood or urine tests. This stage is differentiated as GMM 1 of the probability density function.

GMM 2 - The next stage is the chronic renal failure where the functioning of the kidney deceases gradually over time maybe from months or years. This stage is represented as the GMM 2 of the probability density function.

GMM 3 - The final stage is the end stage renal failure where the survivability of the patient is possible by treating them with dialysis or kidney transplantation. This stage is noted as GMM 3 of the probability density function.

The mean and variance of each dimension (i.e. in each stage) has been calculated for each GMM [7]. Even though the system has been classified, the system still remains to be a weak classifier because of the fact that the system can also be subjected to inaccurate data. So in order to overcome this we need to boost the system to make it a strong classifier. For this purpose the Adaboost technique has being used which combines all the weak classifiers into a strong classifier [2]. The problem of outliers cannot be handled by the Adaboost technique, so in order to overcome this drawback we propose a hybrid model which is the combination of Gaussian with the Adaboost which can easily resolve the outliers.This training method makes the system to learn rather than to memorize [3]. If the sample data contains any outbound data or noisy data [4] they are resolved by this system. As a result the system is trained to handle renal samples.

3.2 Testing Phase

The remaining sample data are used for the testing process. The test data is fed to the predictor along with the three GMMs. Predictor computes the best probabilistic fit among the three GMMs. Thus a system is produced which is less prone to noise,thereby helping in predicting the nature of the disease accurately.

4. Experimental Results

75% of the sample data has been used to generate the Gaussian Mixture Models (GMM). Matlab R2011a is being used for this project. Table I shows the sample input data that are used for the generation of Gaussian Mixture Models (GMM).

TABLE I. SAMPLE DATA

Urea	**Creatinine**	**Uric acid**	**Calcium**	**Phosphorus**	**Potassium**	**Protein**	**Albumin**
37	3.2	6.6	8.5	3.8	3.4	6	3.5
56	6	6.5	9.2	4.5	4.2	5.9	3.6
75	7.8	7.3	8	2.9	4	6.3	3.4
122	9.2	7.7	8.6	4	4.5	7	3.9
109	9.7	7.9	9	5.7	4.3	7.2	4
90	2.7	4.9	7.1	2.9	3.5	6	3.7
. .	. .	. .	. .	. .	. .	. .	. .

Table II shows the mean, variance and weight of each GMM. The mean, variance and the weight are calculated from the above mentioned formulae. The mean, variance of each GMM are used to plot a 2D graph, (see Fig 2).

TABLE II. GMM PARAMETERS

GMM	**Mean**	**Variance**	**Weight**
Gmm 1	1.0249	-4.238	0.74574
Gmm 2	1.034	-4.715	0.2333
Gmm 3	1.023	-4.2373	0.2543

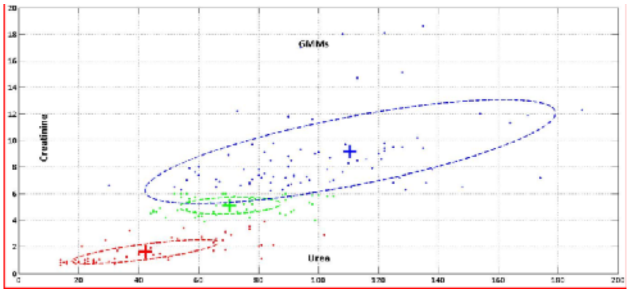


Figure 2: Plot of 3 GMMs

Remaining 25% of the data is used for testing and they are predicted accurately by the predictor. 5 out of the 100 test-dataset were plotted in-correctly by predictor. Overall, the efficiency of the EM with adaboost training is 95%. The same renal dataset was trained using Support Vector Machine learning technique. They were classified into three different renal failure stages. Below are the plots for each stage.

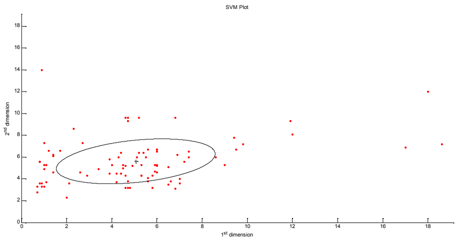
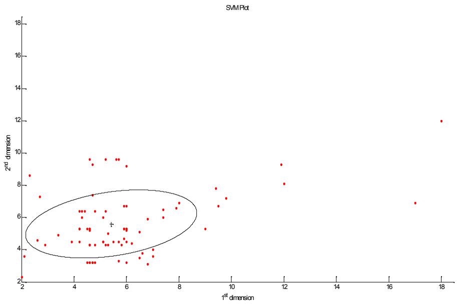


Figure 3: Stage 1 SVM plot



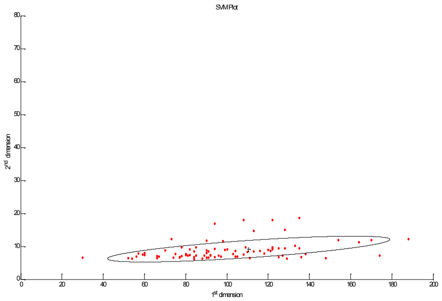


Figure 4: Stage 2 of SVM plotFigure 5: Stage 3 of SVM plot

The efficiency of SVM turned out to be poor with 60% because of its inability to handle the outliers. Thus our proposed system seems to be more efficient than the SVM thereby helping in better predictive rate.

5. CONCLUSION

In this paper, we have employed a Hybrid model of Adaboost technique and Gaussian Mixture model along with Expectation Maximization algorithm to overcome the problem of noisy data set, i.e. which resolves the problem of outliers and thereby lending a hand in predicting the stages in the renal disorder, thus making a best fit for the data. This supports the physicians in diagnosing and making a sound decision.

REFERENCES

[1] Amandeep Kaur, J K Sharma and Sunil Agrawal, “Optimization of Artificial Neural Networks for Cancer Detection”, IJCSNS International Journal of Computer 112 Science and Network Security, VOL.11 No.5, May 2011.

[2] Shigang Chen, Xiaohu Ma, Shukui Zhang, “AdaBoost Face Detection Based on Haar-like Intensity Features and Multithreshold Features”, 2011 International Conference on Multimedia and Signal Processing.

[3] Yufeng Li, Haijuan Zhang, Yanan Zhang, “Research on Face Detection Algorithm in Instant Message Robot”, 2011 International Conference on Uncertainty Reasoning and Knowledge Engineering.

[4] Amit P Ganatra, Yogesh P Kosta, “Comprehensive Evolution and Evaluation of Boosting”, International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010.

[5] Chensheng Sun, Jiwei Hu, Kin-Man Lam, “Feature Subset Selection For Efficient Adaboost Training” in IEEE 2011.

[6] JareeThongkam, GuandongXu and Yanchun Zhang, “AdaBoost Algorithm with Random Forests for Predicting Breast Cancer Survivability”,2008 International Joint Conference on Neural Networks (IJCNN 2008).

[7] KaushikNagarajan, KulandaiAnandBatlagunduRajagopalan, “Emotion Recognition using Glottal Waveform”.

[8] Adenike O. Osofisan, Omowumi O. Adeyemo, Babatunde A. Sawyerr, OluwafemiEweje, “Prediction of Kidney Failure Using Artificial Neural Networks”, European Journal of Scientific Research, ISSN 1450-216X Vol.61 No.4 (2011).

[9] M S SSai, P.Thrimurthy, Dr.S.Purushothaman ,“Implementation of Back-Propagation Algorithm For Renal Datamining”, International Journal of Computer Science and Security, Volume 2, Issue 2.

[10] Wikipedia. (2011, June) Expectation-maximization algorithm.[Online].
http://en.wikipedia.org/wiki/Expectation_maximization_algorithm

[11] A.P. Dempster, N.M. Laird, and D.B. Rubin.Maximimum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 39(1):1{38, 1977}.