Incapable of identifying suspicious records in CTG data using ANN based machine learning techniques

C Sundar 1*, M Chitradevi 2 and G Geetharamani 3
1Department of CSE, Christian College of Engineering and Technology, Anna University, Chennai, India
2Department of CSE, PRIST University, Tamilnadu, India
3Department of Mathematics, Anna University Chennai, BIT Campus, Tiruchirappalli, India

Received 18 October 2012; revised 17 November 2013; accepted 30 April 2014

Cardiotocography (CTG) is a simultaneous recording of fetal heart rate (FHR) and uterine contractions (UC). It is one of the most common diagnostic techniques to evaluate maternal and fetal well-being during pregnancy and before delivery. By observing the Cardiotocography trace patterns doctors can understand the state of the fetus. We implement a model based CTG data classification system using a supervised artificial neural network (ANN) and support vector machine (SVM) which can classify the CTG data based on its training data. The performance neural network based classification model has been compared with the most commonly used unsupervised clustering methods Fuzzy C-mean and k-mean clustering and supervised clustering method SVM classification. According to the arrived results, the performance of the supervised machine learning based classification (ANN) approach provided significant performance than other compared unsupervised clustering methods and supervised SVM classification method. We used Precision, Recall, F-Measure and Rand Index as the metric to evaluate the performance. Even though the traditional clustering methods can identify the Normal CTG patterns, they were incapable of finding Suspicious and Pathologic patterns and the SVM based classifier provided good performance, it was absolutely incapable of identifying a single suspicious record. It was found that, the ANN based classifier was capable of identifying Normal, Suspicious and Pathologic condition, from the nature of CTG data with very good accuracy. The important finding in this paper is Even though SVM is a well proven technique for classification, it was incapable of identifying Suspicious Records in Cardiotocogram Data - but ANN did considerably good classification of Suspicious Records.

Keywords: Cardiotocography, fetal heart rate, SVM, ANN, Fuzzy c-mean, k-mean clustering.

Introduction

Data Mining (DM) and the technology of Knowledge Discovery from Data (KDD) has brought many new developments, methods, and technologies in the recent decade. Also the improvement of integration of techniques and the application of data mining techniques had contributed in handling of new kinds of data types and applications. However, the field of data mining and its application in medical domain is still young enough so that the possibilities of the application are still limitless 21. One of the major challenges in medical domain is the extraction of comprehensible knowledge from medical diagnosis data such as CTG data. In this information era, the use of machine learning tools in medical diagnosis is increasing gradually. This is mainly because the effectiveness of classification and recognition systems has improved in a great deal to help medical experts in diagnosing diseases 22.

Cardiotocography (CTG)

Cardiotocography (CTG) is a technical means of recording the fetal heart rate (FHR) and the uterine contractions (UC) during pregnancy, typically in the third trimester to evaluate maternal and fetal well-being. FHR patterns are observed manually by obstetricians during the process of CTG analysis. In the recent past fetal heart rate baseline and its frequency analysis has been taken in to research on many aspects 2,6. Fetal heart rate (FHR) monitoring is mainly used to find out the amount of oxygen a fetus is acquiring during the time of labor7. Even then death and long term disablement occurs due to hypoxia during delivery. More than 50% of these deaths were caused by not recognizing the abnormal FHR pattern, even after recognizing not communicating the same without knowing the seriousness and the delay in taking appropriate action7.

Clustering and Classification

Clustering is a machine learning technique used to place data elements into related groups without
advances knowledge of the group definitions. Classification is a technique used to predict group membership for data instances. Data classification may be done in two ways; either supervised or unsupervised.

Supervised classification is done with the help of training data set that are defined by the experts. All the objects available for training are characterized by basic statistical parameters such as mean value vectors, covariance matrix that are collected from the training set. These parameters help us in discriminating. The Bayesian classifier is one of its kinds used for supervised classification. Unsupervised classification is done without the help of experts. This type of classification is mostly used for cluster analysis. A classifier is a device that performs the function of classification. An algorithm that implements classification, with the help of several inputs that are transported with signals carrying information about the objects, is known as a classifier. The output of this system is the information generated about the competence of object into particular class.

The Medical Background of Cardiotocography (CTG)

Cardiotocography is a medical test conducted during pregnancy that records fetal heart rate (FHR) and uterine contractions. The CTG trace generally shows two lines. The upper line is a record of the fetal heart rate in beats per minute. The lower line is a recording of uterine contractions from the TOCO.

Baseline Heart Rate

The baseline heart rate helps to evaluate the healthy functioning of the cardiovascular system. The baseline fetal heart rate is determined by approximating the mean FHR rounded to increments of 5 beats per minute (bpm) during a 10-minute window, excluding accelerations and decelerations and periods of marked FHR variability (greater than 25 bpm). Abnormal baseline is termed bradycardia and tachycardia. The fluctuations are visually quantitated as the amplitude of the peak- to-trough in bpm. Using this definition, the baseline FHR variability is categorized by the quantitated amplitude as:

Absant- undetectable
Minimal- greater than undetectable, but less than or equal to 5 bpm
Moderate- 6 bpm - 25 bpm
Marked- greater than 25 bpm

Bradycardia: It is the resting heart rate of under 60 beats per minute, though it is seldom symptomatic until the rate drops below 50 beats/min. It may cause cardiac arrest in some patients

Tachycardia: It typically refers to a heart rate that exceeds the normal range for a resting heart rate (heart rate in an inactive or sleeping individual). It can be dangerous depending on the speed and type of rhythm.

Materials and methods

Fuzzy C-Means Clustering Algorithm

Fuzzy c-means (FCM) is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. This technique was originally introduced by Jim Bezdek in 1981 as an improvement on earlier clustering methods. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters. The Fuzzy c-means algorithm starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely incorrect. Additionally, Fuzzy c-means algorithm assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, Fuzzy c-means algorithm iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point’s membership grade.

The fuzzy c-means (FCM) algorithm was introduced by J. C. Bezdek. The idea of FCM is using the weights that minimize the total weighted mean-square error:

\[ J(w_{kq}, z^{(k)}) = \sum_{(k=1,K)} \sum_{(l=1,K)} (w_{lk})(x^{(ql)} - z^{(k)})^2 \]  

\[ \sum_{(k=1,K)} (w_{lk}) = 1 \text{ for each } q \]  

\[ w_{lk} = \frac{1/(D_{lk})^{(p-1)}}{\sum_{(k=1,K)} (1/(D_{lk})^{(p-1)})}, \text{ } p > 1 \]
K-Mean Clustering Algorithm

One of the most popular heuristics for solving the k-means problem is based on a simple iterative scheme for finding a locally optimal solution. This algorithm is often called the k-means algorithm. There are a number of variants to this algorithm. K-Means algorithm is very popular for data clustering. K-means' goal is to partition data D into K parts, where there is little similarity across groups, but great similarity within a group. More specifically, K-means aims to minimize the mean square error of each point in a cluster, with respect to its cluster centroid.

Formula for Square Error:

\[
\text{Square Error (SE)} = \sum_{i=1}^{k} \sum_{j=1}^{|c_i|} \left( x_j - M_{ci} \right)^2, \quad \ldots \quad (3)
\]

Where, \( k \) is the number of clusters, \( |c_i| \) is the number of elements in cluster \( c_i \), and \( M_{ci} \) is the mean for cluster \( c_i \).

Steps of K-Means Clustering Algorithm

The k Means algorithm is explained in the following steps. The algorithm normally converges in short iterations. But will take considerably long time for a iteration if the number of data points and the dimension of each data is high.

**Step 1:** Choose \( k \) random points as the cluster centroids.

**Step 2:** For every point \( p \) in the data, assign it to the closest centroid. That is compute \( d(p, M_{c*}) \) for all clusters, and assign \( p \) to cluster \( C^* \) where distance

\[
d(P, M_{c*}) <= d(P, M_{c_i}) \quad \ldots \quad (4)
\]

**Step 3:** Recompute the center point of each cluster based on all points assigned to said cluster.

**Step 4:** Repeat steps 2 & 3 until there is convergence.

ANN Based Classification

Here in this classification, we use supervised learning by using a set of training data which is accompanied by class labels. When a new data arrive, then classification of that data will be done based on the training set by generating descriptions of the classes. In addition to training set we also have a test data set that is used to determine the effectiveness of a classification. In general, commonly used and popular neural networks can be trained to recognize the data directly, whereas in simple networks there is a chance of the system being complex and training may be difficult. The time taken and the accuracy of classification depend on the dimension of the input given and also on the dimension in the training data. For input data with high dimension, the process will take a longer time.

Structuring the Network

The number of layers and the number of processing elements per layer are important decisions. These parameters to a feed forward, back-propagation topology are also the most ethereal - they are the "art" of the network designer. There is no quantifiable, best answer to the layout of the network for any particular application. There are only general rules picked up over time and followed by most researchers and engineers applying this architecture to their problems.

**Rule One:** As the complexity in the relationship between the input data and the desired output increases, the number of the processing elements in the hidden layer should also increase.

**Rule Two:** If the process being modeled is separable into multiple stages, then additional hidden layer(s) may be required. If the process is not separable into stages, then additional layers may simply enable memorization of the training set, and not a true general solution effective with other data.

**Rule Three:** The amount of training data available sets an upper bound for the number of processing elements in the hidden layer(s). To calculate this upper bound, use the number of cases in the training data set and divide that number by the sum of the number of nodes in the input and output layers in the network. Then divide that result again by a scaling factor between five and ten. Larger scaling factors are used for relatively less noisy data. If you use too many artificial neurons the training set will be memorized. If that happens, generalization of the data will not occur; making the network useless on new data sets. A single-layer network of \( S \) logsig neurons having \( R \) inputs is shown below in full detail on the left and with a layer diagram on the right\(^{17}\). Feed forward networks (Fig. 1) often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons\(^{11-12}\). Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors. The linear output layer lets the network produce values outside the range \(-1\) to \(+1\). On the other hand, if you want to constrain the outputs of a network (such as between \( 0 \)
and 1), then the output layer should use a sigmoid transfer function.

**SVM Classifier**

Classification can be viewed as a supervised learning scenario. Here a training data set of records is accompanied by class labels. New data can be classified based on the training set by generating descriptions of the classes. In addition to the training set, there is also a test data set which is used to determine the effectiveness of a classification. In principle, the support vector machines can be trained to recognize the data directly. The classification stage follows the feature extraction stage. The main objective of this stage is to classify the FHR signals as normal or at risk represented as '+1' and '-1' respectively. The entire process involves a training set and a testing set of data instances. A training set consists of features extracted from the decomposed FHR signal, also called the attributes, and the labels '+1' (normal) or '-1' (at risk) which are the target outputs.

SVM should produce a model to classify the data instances in the testing set which consists of only features. The classification depends on the inner-product kernel used that produces different learning machines and hence different decision boundaries. In this work, based on the statistical features extracted from FHR signals the radial basis function (RBF) kernel was used equation (5) to obtain reliable results. Parameters C and $\gamma$ are specified by the user.

$$k(x_i,x_j) = \gamma \exp \left( \frac{1}{2} \right)$$  \hspace{1cm} ... (5)

**Steps of SVM Classification Algorithm**

*Step 1:* Scaling the training and testing data sets.

*Step 2:* General grid search method is considered an intractable problem and estimating accuracy for all possible combinations of C and $\gamma$ is a time consuming process. Therefore, exponentially increasing values were considered initially to find a better possible set of values for C and $\gamma$ that yielded better accuracy. Finally, C and $\gamma$ values thus obtained are varied slightly to gain the best possible accuracy. The estimated set of values for C and $\gamma$ are 4 and 2, respectively.

*Step 3:* Training SVM using the chosen C and $\gamma$ values to achieve the best cross-validation accuracy (CSV) possible.

*Step 4:* Predicting the output of the testing set.

*Step 5:* Estimating the accuracy of classification.

Since unbalanced data are used in this work the ratio of C+ / C- are set to the inverse of the corresponding cardinalities of the classes.

**The Metrics Used for the Evaluation**

Precision, recall and F-Measure are computed for every (class, cluster) pair. But Rand index is a metric which will consider all the classes and the clusters as the whole.

**Rand Index**

The Rand index or Rand measure is a commonly used technique for measure of such similarity between two data clusters.

**Recall**

Recall roughly answers the question: "Did all of the documents that belong in this cluster make it in?" In other words, recall is the fraction of actual objects that were identified.

The recall is calculated as:

$$R(L_r, S_i) = n_r / n_i$$

**F-Measure**

F-Measure is the harmonic mean of Precision and Recall and will tries to give a good combination of the two. It is calculated with the equation:

$$F(L_r, S_i) = \frac{2 \times R(L_r, S_i) \times P(L_r, S_i)}{R(L_r, S_i) + P(L_r, S_i)}$$

**Validating the Performance of the Classification**

Classifier performance depends on the characteristics of the data to be classified.
Performance of the selected algorithms is measured for Rand Index, Precision, Recall and F-Measure. Various empirical tests can be performed to compare the classifier like holdout, random sub-sampling, k-fold cross validation and bootstrap method. Here we did Holdout Cross validation for ANN and SVM based classifiers.

Holdout Cross validation

We used Holdout Cross validation because, the dataset contains sufficient amount of samples which can be separated and used for training and testing (50%, 50%). In this work, instead of doing holdout cross validation for a single time, the data set is randomly permuted and the training and testing records were randomly taken for 10 times and the average result of 10 such holdout cross validations only considered.

Results and Discussion

Data Set Information

For evaluating the algorithms under consideration, we used cardiotocograms data from UCI Machine Learning Repository. This data set contains 2126 fetal cardiotocograms belonging to different classes. The data contains 21 attributes and two class labels. The CTGs were classified by three expert obstetricians and a consensus classification label assigned to each of them. Classification was both with respect to a morphologic pattern (A, B, C, ...) and to a fetal state (N, S, P). Therefore the dataset can be used either for 10-class or 3-class experiments. Here we use this data set for these evaluations.

Class Information

We used the data for a three class classification problem. The descriptions for the three classes are

Normal
A CTG where all four features fall into the reassuring category.

Suspicious
A CTG whose features fall into one of the non-reassuring categories and the reassuring category and the remainder of features are reassuring.

Pathological
A CTG whose features fall into two or more of the Non-reassuring the reassuring category or two or more abnormal categories.

The Visualization of Data Space

The image (Fig. 2) shows the projection of this 21 attribute (dimension) data in to a virtual three dimensional data space. We used three principal components of the data for this projection. In this plot, the normal CTG data points are shown in black dots, the suspicious data points are shown as blue dots, and the Pathologic data points are shown as red ‘x’ mark. This figure roughly shows the distribution of the data in the virtual space.

The Analysis of Results

The arrived results (Table. 1) shows, in terms of Rand Index, the performance of ANN based classifier is better than the other three. In terms of time, the performance of ANN was poor than the other two. But 2.5 seconds is not a big figure to consider and will not be an obstacle in practical use of the method in real world application. The Comparison of Precision under four different methods, the ANN based classifier provided good Precision in all the cases (Normal, Suspicious and pathological). Even though the performance of SVM in terms of Precision is good while classifying the Normal and Pathologic records, it was not good in identifying the suspicious cases. The Comparison of Recall under four different methods, the ANN based classifier provided good Recall in all the cases. In terms of recall, SVM was
not good in identifying the suspicious cases. The Comparison of F-Measure under four different methods, the ANN based classifier provided good F-Measure in all the cases. In terms of F-Measure, SVM was not good in identifying the suspicious cases. In terms of classification performance, Fuzzy C-Mean algorithm gives good precision for normal records and poor performance in all other cases, K-Mean algorithm gives good precision for normal records and poor performance in all other cases, ANN based classifier gives good precision, recall and F-Measure for normal as well as pathological records but giving poor performance in the case of suspicious records and SVM was not good in identifying the suspicious cases from the dataset. The arrived results obviously show that supervised machine learning based methods can be used for the classification of CTG data. We realize that there are some training glitches in the case of suspicious records which caused some unexpected poor results while classifying the CTG data class "suspicious". Even, the Fuzzy C-Mean algorithm provided little bit of better result in the case of ‘suspicious’ category of CTG data. This should be noted while designing a improved algorithm for the classification of CTG data. From the results obtained, we find that the rand index value is slightly less than 1 which confirms the maximum formation of similar clusters. By having the recall value and F-Measure value it is clear that the relevance for the query submitted and the test accuracy is high respectively.

### Conclusion

We have evaluated the performance of the three methods with respect to four different metrics. The performance neural network based classification model has been compared with the clustering methods Fuzzy C-mean and k-mean. According to the arrived results, the performance of the supervised machine learning based classification approach provided significant performance than other compared unsupervised clustering methods. Even though the traditional clustering methods can identify the Normal CTG patterns, they were incapable of Suspicious and Pathologic patterns, so that, the traditional unsupervised methods provided very poor accuracy in predicting the different classes. It was found that, the ANN based classifier was capable of identifying Normal, Suspicious and Pathologic condition, from the nature of CTG data with very good accuracy. If we see the performance of ANN with respect to Rand Index, then we can say that it almost provided double the performance of the other two compared methods.

ANN based classifier provided excellent performance in terms of Rand Index, Precision, Recall and F-Measure. It was capable of identifying Normal and Pathologic condition with almost equal accuracy. But if we carefully see the comparative chart of ANN (the last figure), we can tell that, it’s performance to identify the Suspicious CTG pattern is little bit poor than the other two classes. So future works may address the way to improve the system to recognize the Suspicious CTG patterns with the same accuracy. Even though the SVM based classifier provided good performance, it was not good in identifying suspicious record. It means, even though we train the system with all the classes of samples, the trained system was incapable of identifying suspicious record. It is a major weakness we faced during using SVM classifier. And even some trials with ANN, the same problem was identified and that should be overcome in future design. The important finding in this paper is Even though SVM is a well proven technique for classification, it was in capable of identifying Suspicious Records in Cardiotocogram Data - but ANN did considerably good classification of Suspicious Records. One may address the way to improve the system for getting proper training with different classes of CTG patterns. Future works may

### Table 1—Comparison of Results with average of 10 Trials

<table>
<thead>
<tr>
<th>Method</th>
<th>RI</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCM</td>
<td>0.4675</td>
<td>0.7965</td>
<td>0.1739</td>
<td>0.1410</td>
<td>0.3423</td>
<td>0.4251</td>
<td>0.4841</td>
<td>0.4739</td>
<td>0.2432</td>
<td>0.2154</td>
<td>0.3234</td>
</tr>
<tr>
<td>k-Mean</td>
<td>0.4793</td>
<td>0.8056</td>
<td>0.1238</td>
<td>0.0946</td>
<td>0.3257</td>
<td>0.3559</td>
<td>0.3102</td>
<td>0.4575</td>
<td>0.1776</td>
<td>0.1403</td>
<td>0.0375</td>
</tr>
<tr>
<td>ANN</td>
<td>0.9328</td>
<td>0.9663</td>
<td>0.5897</td>
<td>0.9706</td>
<td>0.9910</td>
<td>0.3688</td>
<td>0.9745</td>
<td>0.9784</td>
<td>0.4514</td>
<td>0.9724</td>
<td>2.5891</td>
</tr>
<tr>
<td>SVM</td>
<td>0.8757</td>
<td>0.9266</td>
<td>0.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.0000</td>
<td>1.0000</td>
<td>0.7340</td>
<td>0.9619</td>
<td>0.0000</td>
<td>0.8466</td>
</tr>
</tbody>
</table>

In the following table, RI is the Rand Index. P1 is the precision for normal record, P2 is the precision for suspicious record, P3 is the precision for pathological records. R1 is the recall for normal record, R2 is the recall for suspicious record, R3 is the recall for pathological records. F1 is the F-Measure for normal record, F2 is the F-Measure for suspicious records, F3 is the F-Measure for pathological records. Time is the CPU time taken for the algorithm.
address hybrid models using statistical and machine learning techniques for improved classification accuracy.

Acknowledgements
The authors would like to express their thanks and gratitude to Dr.B.Ramados, who has valuable suggestions and evaluation of the training data. The authors would like to thank Dr.S.Sujath for her valuable suggestion and subjective evaluation of the methods. The authors also thank the anonymous referees for their valuable suggestions.

Appendix - I
Attribute Information
1. LB - FHR baseline (beats per minute)
2. AC - # of accelerations per second
3. FM - # of fetal movements per second
4. UC - # of uterine contractions per second
5. DL - # of light decelerations per second
6. DS - # of severe decelerations per second
7. DP - # of prolonged decelerations per second
8. ASTV - percentage of time with abnormal short term variability
9. MSTV - mean value of short term variability
10. ALTV - percentage of time with abnormal long term variability
11. MLTV - mean value of long term variability
12. Width - width of FHR histogram
13. Min - minimum of FHR histogram
14. Max - Maximum of FHR histogram
15. Nmax - # of histogram peaks
16. Nzeros - # of histogram zeros
17. Mode - histogram mode
18. Mean - histogram mean
19. Median - histogram median
20. Variance - histogram variance
21. Tendency - histogram tendency
22. CLASS - FHR pattern class code (1 to 10)
23. NSP - fetal state class code (Normal=1; Suspect=2; Pathologic=3)

References