GA BASED FEATURE RANKING MECHANISM TO DETECT NEW BORN INFANTS JAUNDICE WITH AN ENSEMBLE TREE STRATEGY

A. ARULMOZHI, DR.M. EZHILARASI

1Assistant Professor, Department of Printing Technology, Avinashilingam University, Coimbatore, Tamilnadu, India.
2Professor & Head, Department of Electronics and Instrumentation (E & I), Kumaraguru College of Technology (KCT), Saravanampatti (Post), Coimbatore, Tamilnadu, India.
E-mail: aruul2001@gmail.com, ezhilrasi@yahoo.co.in

ABSTRACT

Jaundice is a yellow discoloration of the skin and whites of the eyes that is often seen in newborns. Newborn jaundice can be analyzed by scrutinizing the infant and examining blood levels of bilirubin. Infants with high blood levels of bilirubin called hyperbilirubinaemia, evolve the yellow color when bilirubin acquires in the skin. The major symptom of jaundice is yellow coloring of the skin and conjunctiva of the eyes. A Genetic Algorithm (GA) is used to enhance or optimize the overall behavior by evolving the population. Genetic Algorithms (GAs) are a virtually new criterion for a search, based on the precept on instinctive selection. Ensemble methods are learning algorithms that design a set of classifiers. An ensemble of classifiers is a predetermined classifier whose individual decisions are combined in some way to classify new examples. In this analysis, an ensemble of fitness evaluations would produce an ensemble of fitness values for each individual. The experimental results reveal that the proposed method can act as a supplement to support earlier detection and more effective treatment due to improved jaundice detection.

Keywords: Fitness Evaluation, Jaundice, Hyperbilirubinaemia, Maximal Information Compression Index (MICI), Machine Learning, Kernel Support Vector Machine (SVM), Gray Level Co-occurrence Matrix (GLCM), Genetic Algorithm (GA).

1. INTRODUCTION

Infant jaundice is a yellow discoloration in a newborn baby’s skin and eyes. The color of the skin and sclerae varies depending on levels of substance that causes yellowing of the skin and sclerae. Infant jaundice appears because the baby’s blood contains a surplus of bilirubin, a yellow colored pigment of red blood cells. Infant jaundice is a typical condition, especially in baby’s birth before 38 weeks gestation (preterm babies) and some breast-fed babies. The sign of infant jaundice is the yellowing of the skin and the white of the eyes and it mostly appears between the second and fourth after birth. The yellow coloration of the skin and sclera in newborns with jaundice is the result of accumulation of unconjugated bilirubin. Blood group incompatibilities may increase bilirubin production through increased hemolysis. Initial treatment and persistent monitoring of infants at high risk for jaundice can help to avoid severe hyperbilirubinaemia. Genetic Algorithm (GAs) is randomized search and optimization approaches guided by the principles of evolution and instinctive genetics, having a considerable amount of implicit parallelism. GAs perform search in elaborate, enormous and multimodal landscapes and contributes near optimal solutions for objectives or fitness function of an optimization problem. The process of selection, crossover and mutation continues for a fixed number of generations or till a termination condition is satisfied.

Genetic Algorithms (GAs), a form of inductive learning strategy, are adaptive search techniques which have demonstrated substantial improvement over a variety of random and local search methods. The main issues in applying GAs to any problem are selecting an appropriate representation and an adequate evaluation function. In the feature selection problem the main interest is in representing the space of all possible subsets of the
2. RELATED WORK

Rajesh, et al [1] suggested a genetic algorithm based approach for the selection of a subset from the combination of wavelet packet statistical and wavelet packet co-occurrence textural feature sets. In this paper, user’s accuracy, overall accuracy, producer’s accuracy and kappa co-efficient were used to assess the accuracy of the classified data. 

Hu, et al [2] presented a forward greedy approach for exploring the feature subset, which diminishes the Neighborhood Decision Error Rate (NDER). Feature selection plays an essential role in machine learning and pattern recognition. Feature evaluation and optimal subset search are crucial to the overall process of feature selection. Future evaluation functions were used to measure the quality of the candidate subsets. Rao, et al [3] implied the user oriented mechanism for Content Based Image Retrieval (CBIR) approach based on an Interactive Genetic Algorithm (IGA). The primary disparity between IGA and GA is the construction of the fitness function, i.e. the fitness is resolved by the user’s evaluation and not by the predetermined formula. The IGA desires a population of potential solutions to be computed at the beginning of the GA process.

Yousef, et al [4] proposed a comparison analysis between one class and two class machine learning for microRNA target detection. Machine learning facilitates one to develop automatic rules based on observation of the applicable examples of the learning machine. Chen, et al [5] presented an improved Self Organization Map (SOM) algorithm and its applications to color feature extraction. In this paper, the quality of features and the data reduced by the conventional SOM algorithm heavily depends on the distribution of the training data. Karakukcu, et al investigated the possible relations between serum bilirubin and CoQ10 concentrations in jaundiced newborns [6].

Jaundice in newborn is caused by neonatal changes in bilirubin metabolism resulting in increased bilirubin production, diminished bilirubin clearance and enhanced enterohepatic circulation. Jaundiced newborns at high and low risk zone were excluded in this paper. According to the nomogram newborns were divided into two groups: infants in the low intermediate risk zone were considered as Group I and in the high intermediate risk zone were considered as Group II. Dani, et al evaluated the various procedures used in the management of hyperbilirubinaemia in preterm infants [7]. There are no uniform guidelines regarding the management of hyperbilirubinaemia in preterm
infants, although these could assist greatly in the management of jaundice in these patients.

Nassef, et al proposed a paper for the interpretation of G6Pd activity and antioxidants status in jaundiced Egyptian neonates [8]. Jaundice with glucose – 6 phosphate dehydrogenase (G6PD) deficiency is one of the most prevalent conditions needing medical attention in newborn babies. Jaundice is one of the most common problems for neonates, including premature infants. Glucose – 6 phosphate dehydrogenase (G6PD) deficiency is an important disorder of the hexose monophosphate shunt in erythrocyte metabolism that catalyses the oxidation of glucose – 6 phosphate to 6 phosphoglucuronate; the rate limiting step of the pathway. Rosario, et al identified how the nurse cares for the newborn infant with jaundice [9].

Neonatal jaundice is characterized by accumulation of bilirubin in the blood above 5mg/dL and it may have a slow or rapid progression. Davanzo, et al [10] suggested a paper for breastfeeding and neonatal weight loss in healthy term infants. The aim of this study was to determine the extent of neonatal weight loss and its combination with elected variables in a population of healthy term infants cared for using a distinct protocol on weight loss. Mansor, et al [11] suggested a Principal Component Analysis (PCA) based feature extraction and k-NN algorithm for jaundice detection. In this paper, a PCA method was employed to study the behavior of the infant. Zappa, et al [12] compared the characteristics of jaundice and hyperbilirubinaemia in the newborn population of both immigrant and Italian mothers.

Song, et al [13] suggested a paper about discriminative data transform for image feature extraction and classification. In this process, while generic feature descriptors were used, data adaptive information was integrated into the feature extraction process based on the optimization objective to enhance the discriminative power of feature descriptors. Chen, et al [14] presented an enhanced Self Organization Map (SOM) algorithm and its application to color feature extraction. Kaynak-Turkmen, et al [15] contemplated a research for transcutaneous measurement of bilirubin in Turkish newborns. Hyperbilirubinaemia is the most common diagnosis that leads to hospital readmission of newborns within the first month of life. Although the most jaundiced infants suffer no lasting ill effects, acute bilirubin encephalopathy may appear at high bilirubin levels.

Onyewuruha, et al [16] determined the occurrence, aetiological and other associated factors of neonatal jaundice. The incidence, aetiological and contributory factors to neonatal jaundice vary according to ethnic and geographical differences. Sakha, et al [17] evaluated the efficacy of clofibrate in reducing total serum bilirubin levels in near term neonates with non-hemolytic jaundice. In the neonates, hyperbilirubinaemia is generally due to a sequence of an enhanced bilirubin load and diminished bilirubin elimination. Aina and Omoigberate [18] determined the risk factors for neonatal jaundice. Bilirubin production in the newborn is two to three times higher per kilogram body weight than in adults. Conjugated bilirubin is not neurotoxic, but it may signify a serious disorder in the newborn. Boskabadi, et al determined the prevalence rate of hemolytic jaundice, predisposing factors and assessment of treatment and complications in hemolytic jaundice [19]. Laddi, et al [20] investigated a non-invasive and instant method of jaundice detection using machine vision technique. In this paper, the proposed technique was developed based upon non-invasive detection of jaundice by using simple image acquisition and processing tools.

3. PROPOSED METHOD

The aim of this study is to improve the classification rate and the accuracy level by using the machine learning algorithm and the ensemble tree for fitness value evaluation. The development of a jaundice detection system involves the following tasks:

- Skin Detection
- Preprocessing
- Color and Texture Feature Extraction
- Feature Selection
- SVM Classification

3.1 Skin Detection

Newborn jaundice is an apparent yellowing of the sclera or yellowish skin in newborn infants. In this paper, a combination of preprocessing and the skin detection method to detect jaundice infants is proposed. Skin detection can be defined as the process of selecting which pixels of a given image correspond to human skin. Skin detection is useful in face detection and face tracking. YCbCr has been widely used for the skin pixels form a compact cluster in the Cb-Cr plane. Conversion to normalized color space \((r, g)\) chromaticity diminishes brightness dependence. Where,

\[
\begin{align*}
    r &= R / (R + G + B) \\
    g &= G / (R + G + B)
\end{align*}
\]
Actual measurements have shown that dark, yellowish and pale skin has almost the same chromaticity.

3.2 Preprocessing
In this analysis, a novel filter structure is used, namely, Hybrid Median Filter, which constitutes a natural extension of the nonlinear rational type hybrid filters. A technique for processing an image, comprises a foreground and a background to yield a greatly compressed and specific portrayal of the image. In hybrid median filtering, the neighboring pixels are ranked according to brightness (intensity) and the median value becomes the new value for the central pixels. In the median filtering operation, the pixel values in the neighborhood window are ranked according to the intensity and the middle value (median) becomes the output value for the pixel under evaluation. The hybrid median filter is more expensive to compute than a smoothing filter.

Given a set of random variables $M = (M_1, M_2, \ldots M_N)$, the order statistics $M_{(1)} \leq M_{(2)} \leq \ldots \leq M_{(N)}$ are random variables, defined by sorting the values of $M_i$ in an increasing order. The median value is then given as,

$$\text{Median} (M) = \begin{cases} M_{(\frac{N}{2})} & \text{for } X = 2A + 1 \\ \frac{1}{2} (M_{(A)} + M_{(A+1)}) & \text{for } X = 2A \end{cases} \quad (3)$$
Where, \( f = 2A + 1 \) is the median rank. The median is considered to be a robust estimator of the location parameter of a distribution and has found numerous applications in smoothing. For a grayscale input image with intensity values \( x_{ij} \), the hybrid median filter is defined as,

\[
Y_{ij} = \text{median}_{(a,b)\in G}(x_{i+r, j+s})
\]

Where, \( G \) is a window over which the filter is applied. The symmetric square window of size \( A \times A \) with \( A = 2C + 1 \), i.e., the median rank \( f \) equals \( f = (A^2 + 1)/2 \). This is probably also the most widely used form of this filter.

### 3.3 Color and Texture Feature Extraction

The extraction of image features is the fundamental step for image classification. There are various types of features for image classification’s aim as follows: color, statistical features, shape features and transform coefficient features. Feature extraction is a method of capturing the visual content of images for indexing and retrieval. Texture is one of the most imperative characteristics used in identifying objects or regions of interest in an image. Selection of a feature extraction approach is commonly the single most crucial factor in accomplishing high recognition performance in character recognition system.

A common technique in texture analysis involves the computation of GLCM as a second order texture measure. In statistical texture analysis, texture features are computed from the statistical distribution of observed combinations of intensities at specified positions relative to each other in the image. According to the number of intensity points (pixels) in each combination, statistics are classified into first order, second order and higher order statistics. The GLCM method is a way of extracting second order statistical texture features. The features of GLCM are,

1. Contrast
2. Correlation
3. Cluster prominence
4. Cluster shade
5. Dissimilarity
6. Energy
7. Entropy
8. Homogeneity
9. Local homogeneity
10. Maximum Probability
11. Sum of Squares
12. Auto correlation

For extraction of color, texture features, RGB and HSV color spaces are used. HSV (Hue, Saturation and Value) color space is a nonlinear transformation of the RGB color space. Feature vector is computed for every Hue (H), Saturation (S) and Value (V) channel:

\[
FV = [\text{mean (H)}, \text{mean (S)}, \text{mean (V)}, \text{variance (H)}, \text{variance (S)}, \text{variance (V)}, \text{skewness (H)}, \text{skewness (S)}, \text{skewness (V)}]
\]

Where, \( FV \) – Feature Vector. GLCM describes the frequency of one gray tone appearing in a specified spatial linear relationship with another gray tone, within the area under investigation.

### 3.4 Feature Selection

Feature selection is one of the key topics in machine learning, it can remove the irrelevant, even noisy features and hence improve the quality of the data set and the performance of learning systems. The wrapper methods wrap the feature selection, using cross validation to predict the benefits of adding or removing a feature from the feature subset used. In feature selection, a wrapper estimates the accuracy of adding each unselected feature to the feature subset and chooses the best feature to add according to this criterion. Feature selection is a process of finding a subset of features, from the original set of features forming in a given data set, optimal according to the given goal of processing and criterion. Since each feature used as part of a classification procedure can increase the cost and the running time of a system, there is a strong motivation within the image processing community to design and implement systems with small feature sets. Feature ranking is useful to gain knowledge of data and identify relevant features. One of straightforward feature selection process is based on an evaluation of predictive power of individual features, then ranking such evaluated features and eventually choosing the first best features. The ranking of individual features will be done, in order to reduce the GA optimization complexity. After ranking of features, the top most features will be inserted into GA.

Applying Genetic Algorithms (GAs) to the feature selection problem is straightforward: the chromosomes of the individuals contain one bit for each feature and the value of the bit determines whether the feature will be used in the classification. Using the wrapper approach, the individuals are evaluated by training the classifiers. Ensemble methods are learning algorithms that
construct a set of classifiers and then classify new data points by taking a vote of their predictions.

The main discovery is that ensembles are often much more accurate than the individual classifiers that make them up. In this paper, a recent and generic machine learning algorithm based on decision tree ensembles that has been shown to perform remarkably well on a variety of tasks. Feature selection gives the chance to attain the same or better performance by using fewer features. So, two terms are adequate in fitness function, accuracy and the number of selected features. With these parameters the fitness function should be maximized during optimization operation.

$$\text{Fitness function} = \text{Error} + \mu \text{Ones}$$

(6)

Where, Error corresponds to the classification error for a specific subset of features and Ones corresponds to the number of features selected. $\mu$ is a coefficient that expresses the weight of second parameter of fitness function (Ones) versus the first one (Error). A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate.

Feature selection for ensembles has shown to be an adequate approach to ensemble creation due to its ability of generating good subsets of features, which make the classifiers of the ensemble disagree on crucial cases. Initially, it performs feature selection in order to develop a set of classifiers and then it elects the best team of classifiers. An ensemble of classifiers has been extensively used to diminish model uncertainty and enhance generalization performance.

3.5 Kernel SVM Prediction

The Support Vector Machine is a theoretically superior machine learning methodology with great results in pattern recognition. The function can be a classification function or the function can be a general regression function. The kernel based method is based on mapping data from the initial input feature space to a kernel feature space of higher dimensionality and then determining a non-linear problem in that space. These techniques allow us to interpret learning algorithms geometrically in the kernel space, thus incorporating statistics and geometry in an efficient way.

4. PERFORMANCE ANALYSIS

The experimental results are used to detect the jaundice in newborn babies. Jaundice is due to an escalation in the blood of bilirubin, a yellow pigment which comes from the breakdown of old red blood cells. It is normal for red blood cells to break-down, but the bilirubin organized does not ordinarily cause jaundice because the liver metabolizes it and gets rid of it into the gut. Newborn babies are often affected by jaundice, which makes their skin and eyes have a yellowish tinge. Other symptoms may incorporate drowsiness, feeding troubles and dark urine. Bilirubin is a waste product of the body’s breakdown of old and damaged red blood cells.

4.1 Preprocessing

Fig 2. (A) depicts the input image. Baby image 16 is given as the input image. The newborn danger signs are defined as the signs which occur in the newborn within 30 days of life, such as pathological jaundice, poor feeding and excessive weight loss. Fig 2. (B) shows the skin detection. The preeminent demand in skin detection is to constitute the recognition robust to the large variations in appearance of skin that may occur, like shape, intensity, color, etc. Skin color and texture are important cues that people use consciously or unconsciously to infer a variety of culture-related aspects about each other. Skin detection is the process of finding skin colored pixels and regions in an input image.

Fig 2. (C) represents the region of interest. A region of interest is a portion of an image that performs the filter operation on the image. Fig 2. (D) illustrates the preprocessed image. In this work, the hybrid median filtering technique is used for preprocessing. The aim of preprocessing is an advancement of the image data that conceals unwanted distortions or enlarges some image features essential for further processing.
4.2 Genetic Algorithm

Figure 3 shows the genetic algorithm with fitness evaluation. In this graph, the x-axis represents the generation and the y-axis represents the fitness value. The average fitness of the whole population is the fitness of each genotype multiplied by its frequency, this is known as mean fitness.

4.3 Confusion Matrix

Fig 4. Shows the confusion matrix. The confusion matrix function allows comparison of a classified image. A confusion matrix contains information about actual and predicted classifications done by a classification system. The performance of such systems is generally classified using the data in the matrix.

4.4 Analysis

The comparison between the normal babies and jaundice affected babies are shown in Fig 5. In this analysis, the classification rate and the level of accuracy are increased by using the ensemble tree. Kernel SVM classifier is used to classify the normal babies and the jaundice detected babies.
4.5 Performance Measure for PSO with MICI

The performance measure for PSO with MICI is shown in Table 1. In this analysis, the accuracy level was increased by 93.75%.

Table 1. Performance Measure for PSO with MICI

<table>
<thead>
<tr>
<th>Performance</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Rate</td>
<td>0.9375</td>
</tr>
<tr>
<td>Error Rate</td>
<td>0.0625</td>
</tr>
<tr>
<td>Last Correct Rate</td>
<td>0.9375</td>
</tr>
<tr>
<td>Last Error Rate</td>
<td>0.0625</td>
</tr>
<tr>
<td>Inconclusive Rate</td>
<td>0</td>
</tr>
<tr>
<td>Classified Rate</td>
<td>1</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>1</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.7500</td>
</tr>
<tr>
<td>Positive Predictive Value</td>
<td>0.9231</td>
</tr>
<tr>
<td>Negative Predictive Value</td>
<td>1</td>
</tr>
<tr>
<td>Positive Likelihood</td>
<td>4</td>
</tr>
<tr>
<td>Negative Likelihood</td>
<td>0</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.7500</td>
</tr>
</tbody>
</table>

4.6 Performance Measure for Tree Bagging

The performance measure for PSO with MICI is shown in Table 2. In this analysis, the accuracy level was increased by 100%. In this paper, the level of accuracy and classification rate are increased compared than the previous approach.

Table 2. Performance Measure for PSO with MICI

<table>
<thead>
<tr>
<th>Performance</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Rate</td>
<td>1</td>
</tr>
<tr>
<td>Error Rate</td>
<td>0</td>
</tr>
<tr>
<td>Last Correct Rate</td>
<td>1</td>
</tr>
<tr>
<td>Last Error Rate</td>
<td>0</td>
</tr>
<tr>
<td>Inconclusive Rate</td>
<td>0</td>
</tr>
<tr>
<td>Classified Rate</td>
<td>1</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>1</td>
</tr>
<tr>
<td>Specificity</td>
<td>1</td>
</tr>
<tr>
<td>Positive Predictive Value</td>
<td>1</td>
</tr>
<tr>
<td>Negative Predictive Value</td>
<td>1</td>
</tr>
<tr>
<td>Positive Likelihood</td>
<td>NaN</td>
</tr>
<tr>
<td>Negative Likelihood</td>
<td>0</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.7500</td>
</tr>
</tbody>
</table>

5. CONCLUSION

In this paper, an improved machine learning algorithm is proposed for jaundice detection. A common condition in newborns, jaundice refers to the yellow color of the skin and whites of the eyes caused by excess bilirubin in the blood. In this research, preprocessing is done by using the hybrid median filtering technique and GLCM is used for texture feature extraction and color feature extraction. In this analysis, an ensemble of fitness evaluations would produce an ensemble of fitness values for each individual. Feature ranking is useful to gain knowledge of data and identify relevant features. The ranking of individual features will be done, in order to reduce the GA optimization complexity. After ranking of features, the top most features will be inserted into GA. Finally, the Kernel Support Vector Machine (SVM) is used to classify the normal babies and jaundice detected babies.

REFERENCES


