Machine Learning to Predict the Incidence of Retinopathy of Prematurity

[1]Aniket Ray, [2]Vikas Kumar, [1]Balaraman Ravindran, [3]Dr. Lingam Gopal, [3] Dr. Aditya Verma

[1]Indian Institute of Technology Madras, Department of Computer Science and Engineering, Chennai, India 600036
[2]National Institute of Technology Rourkela, Department of Computer Science and Engineering, Rourkela, India 769008
[3]Medical Research Foundation, Sankara Nethralaya, 18 College Road, Nungambakkam ,Chennai, India, 600006
aniket.ray@gmail.com, ravi@cse.iitm.ac.in

Abstract

Retinopathy of Prematurity (ROP) is a disorder afflicting prematurely born infants. ROP can be positively diagnosed a few weeks after birth. The goal of this study is to build an automatic tool for prediction of the incidence of ROP from standard clinical factors recorded at birth for premature babies. The data presents various challenges including mixing of categorical and numeric attributes and noisy data.

In this article we present an ensemble classifier—hierarchical committee of random trees—that uses risk factors recorded at birth in order to predict the risk of developing ROP. We empirically demonstrate that our classifier outperforms other state of the art classification approaches.

Introduction

Retinopathy of Prematurity (also known as *retrolental fibroplasia*) is a disease of the retina that typically starts developing a few weeks after the premature birth of a child. Its diagnostic test involves dilating the infant's eye using eye drops and then physically checking the condition of the eye. As the child may not have developed ROP by the time of the first test, follow-up tests need to be conducted every two weeks. Our problem involves learning a model so that an accurate prediction can be made as to whether or not the child would contract the disease based on conditions of the child recorded in Neonatal ICU (NICU).

The motivation for this problem stems from the fact that as the disease shows symptoms only after a few weeks, most families would have left the hospitals by then. Also, as a large population in developing countries comes from rural areas they lack the resources to come back to the urban hospitals, even if the child starts showing symptoms. If a system exists that would detect the disease in these infants, the babies could be kept for observation at the hospital itself.

Apart from the large social relevance, this problem also has several challenging issues from a computer science perspective. The problem is a medical data mining task that is characterized by the following challenges:

- Volume and Complexity of the Medical Data
- Physician's interpretation: A lot of natural language problems are involved in medical data mining.
- Sensitivity and specificity: All medical diagnoses are prone to error. Typically, the mining results are proposed as an inexpensive new test, to compete against the original more expensive test (called the hypothesis) which is regarded as definitive. Several evaluation measures have been defined to compare different diagnosis systems.

$$Sensitivity = \frac{\text{True Positive}}{\text{Hypothesis Positive}}$$

$$Specificity = \frac{\text{True Negative}}{\text{Hypothesis Negative}}$$

$$Accuracy = \frac{\text{True Positive} + \text{True Negative}}{\text{Total Data Points}}$$

where True Positive are points that are labeled positive by the system as well as by the original system. Similarly, true negatives are data points that are labeled as negative by both the diagnosis system and the original hypothesis system.

- Non-intuitive mathematical mapping: Most medical concepts can not be intuitively mapped into mathematical models. Terms like inflammation and tiredness lack a formal structure.
- Categorical form: Many medical variables like Blood Group, APGAR values etc are actually categorical in nature.

In spite of these problems, machine learning has been shown to be especially suited for small specialized diagnostic problems, for which independent tests are

extremely costly like oncology, cardiology, neurophysiology, dermatology [9], etc. Traditionally, only rule based learning systems have been used for diagnosis. The reason for the popularity of such systems is that these produce rules which can be easily understood by a physician. Recently, focus has started shifting towards more mathematical approaches like Bayesian Classification and Support Vector Machines.

In the Problem Formulation section, we formulate the task as a machine learning problem. The section on Layered Hierarchical Committee of Trees details a new hierarchical ensemble method for classification. The "Results" section outlines the details of experiments conducted and the results obtained from the learning machine evaluations. The Conclusion section points out the major conclusions which can be arrived at, from the results of the study.

Problem Formulation

There are 3 different kinds of classes in this problem which are diagnostically significant. The 2 most significant classes are "No ROP" i.e. the class of infants that do not contract the disease and "Progressed ROP" which is the class of infants that contract the disease. The third class is slightly less significant diagnostically; it corresponds to infants who initially show symptoms of ROP but these symptoms later regress.

Class ID	Name Of Class	Description
Class 0	No ROP	Infant shows no symptoms of the disease.
Class 1	Regressed ROP	Infant starts out showing symptoms, but the symptoms eventually wane completely.
Class 2	Progressed ROP	Once the infant starts showing symptoms, the symptoms generally become worse. In some cases the condition improves but the symptoms do not fade completely.

Table 1: Different ROP levels that a premature infant may develop.

This problem is treated as two separate classification problems. In the first problem, using the 2 more significant classes; we attempt to predict whether the child would belong to the No ROP class or Progressed ROP class. Separately, we study it as a 3 class problem where the classes correspond to no ROP, regressed ROP and progressed ROP as shown in Table 1.

It must be noted that the classification problem uses conditions recorded till a few hours after birth, before the child is discharged from NICU and attempts to classify the levels of the infant's ROP that the child may develop at some time in the future.

We used 47 features which are standard measurements and routine tests that are conducted for any premature child after his/her birth. These include nominal features like number of days before the baby is discharged from NICU, gestation period, weight etc. The categorical features include binary valued features like whether blood transfusion was performed, whether the infant was breast fed etc, and multi category categorical features like Blood Group, method of delivery etc. Most of these measurements suffer from experimental error and bias based on the actual test administrator. Some features had too many missing values and had to be ultimately dropped.

In this study, we have tried a variety of classification algorithms for solving this problem. Naive Bayes classifier was used as a base model, against which all other models were tested. Naive Bayes Classifiers are typically advantageous in situations with high dimensionality. Even though the independence assumption is extremely naive, in some medical data mining tasks, Naive Bayes has been known to give results comparable to more complex methods. In particular, each distribution can be independently estimated as a one dimensional distribution which alleviates the need for large data sets. PART and C4.5 [11] decision tree algorithms have been used as they produce classifiers that are easy to understand. Decision Trees assume nothing about the underlying distribution. The other major advantage of decision trees lies in the fact that they can handle both categorical and nominal data and also support missing data. Support Vector Machine (SVM) [2] is a kernel based, maximum margin approach to classification. The training phase involves learning parameters of separating hyperplanes that would maximize the distance of the nearest training points from the separating hyperplane.

Ensemble methods have been used to try and improve the results of the individual classifiers. In ensemble methods of machine learning [7], we use a combination of a set of classifiers to find the final class output. An ensemble method generally improves the accuracy and robustness of the individual classifier. The necessary and sufficient condition for this improvement is that the individual classifiers should be accurate and diverse. A classifier is said to be accurate if it can give better accuracy than random guessing. Two classifiers are diverse if they make different errors on unseen data points.

Bootstrap aggregating or bagging [1] is an ensemble method which is known to reduce the variance in the classification mechanism. Also, it has been seen that over fitting problems can be overcome using bagging. It should be noted that since bagging does an average prediction, as the number of components tends to large values the effect on linear classifiers would become negligible.

Boosting is an iterative process to improve the performance of weak learners in terms of accuracy. It has been proven that as the number of iterations tends to infinity the classifier obtained by boosting is a maximal margin classifier [5].

Random Forest [8] is a classification mechanism that combines the concepts of Bagging with the Random Subspace Method [4]. A data set is created from the original data set **D** using sampling with replacement for each of N different trees. Once the data set is selected, a decision tree is created on the data set. At each node, *m* variables out of the total M decision variables are randomly selected and the best split out of these *m* variables is performed. In this way, each tree is grown to its full capacity without pruning. All trees within a random forest are randomly grown using the same distribution. Once N different trees are grown, simple voting is performed to figure out the final class to be labeled. When the number of decision variables is too large, problems associated with the curse of dimensionality can be avoided with this method.

Feature reduction methods were also evaluated. Best first search method [3] evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class, while having low inter-correlations are preferred. The space of attribute subsets is searched by greedy hill climbing augmented with a backtracking facility. The search procedure starts with the empty set of attributes and searches forward by considering all possible single attribute additions and deletions at a given point. The other method of feature reduction, Principal Components Analysis [14] involves calculation of a linear transformation, so that points in a higher dimensional space can be transformed into those into a space of lesser dimension. Typically, the transformation is selected in such a way that models in the smaller dimensional space can be defined easily.

Layered Hierarchical Committee of Trees

Random Forests give extremely encouraging experimental results, but they are hampered by the fact that they are an ensemble of randomly created trees. Rather than randomly creating the trees and carrying out a single vote, we can create a mechanism where the random trees are divided into separate sets, and the prediction by each set is biased differently. This is done using a new framework of ensemble learning called Hierarchical Ensemble Learning. In this framework, rather than simply taking votes among all the individual components, we actually carry out hierarchical voting. This method may also be viewed as a hierarchical constituency based voting system. The decision at each constituency is taken based on voting by its constituents at the sub-level. Each constituency gets a single vote for affecting the decision at its super-level.

A new machine learning algorithm for modeling data mining tasks based on this framework has been created, which we call the Layered hierarchical Committee of Trees (LCT). We can treat this framework as a tree with the root node as the final decision layer. Every non-leaf node is a random committee comprised of its child nodes as its component classifiers i.e. one set of node and its children form a random committee with each of the children corresponding to the individual random classifiers.

In the LCT model, all the leaf nodes are random forest classifiers themselves which classify each data point as belonging to some class. Each node just above the leaf nodes, would then assign the data point a class based on a majority of votes of its child (leaf) nodes. In this way, class belief is percolated up the tree until we reach the root where the final decision is taken.

The construction of the LCT takes place in 2 phases: the tree building phase (in which data from the data set is used to create a random tree) and a layering phase (in which random trees are combined in a hierarchical fashion to create a layered decision making structure). The detailed construction algorithm has been discussed in Algorithm 1. The Tree Building step would create a random tree using information gain principles finding the best split at each node. The layering step involves the combination of Trees in Forests, then forests into committee of forests, and so on.

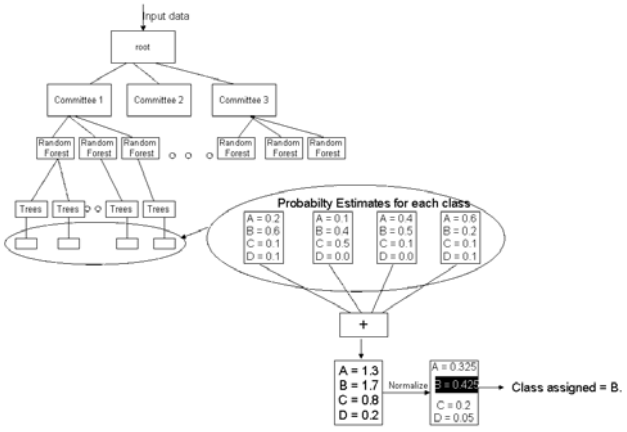

Fig 1: Classification using LCT

Votes of committees at each layer are passed up the tree as a single vote at the higher layers, starting with random

forests at the leaf, thus classifying any data point as shown in Fig 1.

The main advantage of the LCT lies in the fact that the different random forests can use different distributions or can use distributions with different parameters (which we recommend). This leads to lesser bias and better modeling of outliers, leading to a better over all accuracy. The variable interactions are better captured in this method. Thus, problems with large number of features are also reduced.

Algorithm 1. LCT Construction Algorithm

1: Input: Data set *D* with *l* training points in an *M* dimensional input space.
2: Input: Parameter *m*, which is the number of features that would be considered at each node in the tree.
3: Input: Parameter *N*, which is the number of random trees to build.
4: Input: Parameter *R*, which is the number of levels up to which LCT is to be built.

createLCT(*D*, *m*, *N*, *R*):
1: Set *totalComponents* = $N^R - 1$
2: for $i = 0$ to *totalComponents* do
3: Create random forest R_i = createRandomForest(*D*, *m*, *N*)
4: end for
5: \\At this point, we have all the individual components required.
6: Set *currentLevel* = R
7: while *currentLevel* > 1 do
8: Set *componentsAtLevelAbove* = $N^{currentLevel} - 2$
9: Randomly assign *N* different components (random forest or committee) to be part of one component in level above. These *N* sub committee components are combined using voting to form a super committee.
10: *currentLevel* = *currentLevel* – 1
11: end while
12: \\These steps build a tree, where each node in the tree is a random committee formed by combining *N* smaller random committees. This is repeated until the required R (number of levels) is achieved.

RESULTS

In this study, we attempt to model the medical data available to us, corresponding to 358 different prematurely born infants. Out of these, 169 infants did not show any sign of ROP. 77 infants initially showed some signs of retinopathy which later regressed, while for 102 infants the retinopathy fully progressed. As can be seen, compared to the real world this is biased towards the progressed ROP case, as in reality only about 21.3% of premature infants contract the disease [6].

As earlier mentioned, there were 2 different kinds of problems that were attempted. First, we treated it as a 2 class problem differentiating between ROP and No ROP cases. The other as a 3 class classification problem to differentiate between No ROP, Regressed ROP and fully Progressed ROP data. Our study primarily aims to find an ideal learning machine for this classification problem. We have analyzed the performance of different classifiers for this problem. We have also noted the False –ve percentages of the results, which denote the percentage of infants who contract the disease but were predicted to belong to the No ROP class.

Even for the 2 class case, the data was trained as a 3 class problem and then all test data points predicted to be regressed ROP were treated as predictions for No ROP. This approach has been recommended for non-i.i.d. [13] data and application of the method to this problem is a logical extension.

While applying algorithms which use only nominal features, we convert the categorical features into a set of binary features e.g. Blood group can be treated as 4 different features each with 0 or 1 values. All nominal features were normalized between 0 and 1 in an unsupervised fashion for all the classifiers. All results are based on 5 fold cross validation. Naive Bayes (which has shown some good performances in other diagnostic tasks [9, 10]) was used as a base performance classifier. We use the Java implementation (J4.8) of the C4.5 algorithm.

SVMs were used because of their nature to maximize margins. As there is a bias in the number of training points from No ROP class, one experiment was conducted with a counter bias by sampling from the data set accordingly. In each of the 2 biased experiments, points from regressed ROP and progressed ROP classes were respectively doubled.

Classifier	Accuracy(%)	Sensitivity(%)	Specificity(%)	False –ive(%)
No bias	80.42	77.45	82.12	22.54
Bias towards Regressed ROP	80.78	83.33	79.32	16.66
Bias towards Progressed ROP	81.13	85.29	78.77	14.71

Table 2: Performance of Quadratic Support Vector Machines for the 2 class problem with varying bias

Bagging and Boosting with 100 components each were created. In bagging and boosting, each data set was learnt using restricted decision REP trees (expansion of a node is stopped when purity of the node goes above a threshold). Attempts at ADABoost using decision stumps and logistic regression were made but results were not encouraging. Also, some experimental attempts were made by changing the number of iterations but even then, the results for boosting did not improve significantly. Random Forests also gave good results with 100 and 1000 components. We use a 3 Layer hierarchical Committees of Trees (3LCT) in

which each layer comprises 10 subcommittees and the leaves are random forests with 10 trees each. Thus, an ensemble comprising 1000 component classifiers was used for the experiments.

The LCT uses 3 levels with each layer having 10 subcommittees and leaves consisting of Random Forest with 10 trees each thus the model uses a total of 1000 iterations to predict the accuracy up to 71.7877% (three classes).

Classifier	Accuracy(%)	Sensitivity(%)	Specificity(%)	False – ive(%)
Naïve-Bayes	63.69	64.70	77.09	35.29
PART	65.36	70.59	78.77	37.25
C4.5	60.06	61.76	78.77	38.24
Random Forest (100 trees)	70.39	71.57	89.39	28.43
Random Forest(1000 trees)	71.23	70.59	90.5	29.41
SVM:RBF kernel	67.04	68.62	84.92	31.37
SVM:Polynomial kernel	66.76	65.69	83.24	34.31
3 LCT	71.79	72.55	91.62	16.67

Table 3: 3-class accuracy of each classifier

Classifier	Accuracy(%)	Sensitivity(%)	Specificity(%)	False – ive(%)
Naïve-Bayes	77.09	64.7	82.03	35.39
PART	82.68	70.59	86.32	37.25
C4.5	75.42	61.76	80.86	38.24
Random Forest (100 trees)	82.96	71.57	87.50	28.43
Random Forest(1000 trees)	83.24	70.59	90.23	29.41
SVM:RBF kernel	84.08	68.62	90.23	31.37
SVM:Polynomial kernel	81.84	65.69	88.28	34.31
3 LCT	84.36	72.55	91.62	27.54

Table 4: 2-class accuracy (Regressed ROP taken as No ROP)

A reduced subset of features was calculated using hill climbing. These were { Gestation Period, Age, Birth Weight, Sex, Whether Serum Bilirubin is elevated (yes/no) , Number of days O_2 was given, Whether ultrasound of the brain was conducted (yes/no), Whether apneic episodes happened (yes/no), whether the baby was kept in incubator (yes/no), Whether the baby was breast fed (yes/no), Whether the baby suffered from Hypoglycemia (yes/no), Whether the baby developed Septicemia (yes/no) }. The SVM trained on these features seemed to bias the classification towards the no ROP case, as can be seen in the high specificity and markedly low sensitivity. Principal Components Analysis (PCA) was done and the 12 most prominent eigenvectors were chosen (the number 12 was chosen to compare with the hill climbing results). The results are shown in Table 5. The high specificities in the reduced features were the reason that reduced features were not used for the final classification.

Classifier	Accuracy(%)	Sensitivity(%)	Specificity(%)	False – ive(%)
All features	80.42	77.45	82.12	22.54
Hill Climbing	81.18	76.47	84.91	23.52
PCA	81.13	66.66	89.38	33.33

Table 5: Performance of Feature Selection methods for the 2 class problem over SVMs Both the methods reduced space to 12 dimensions

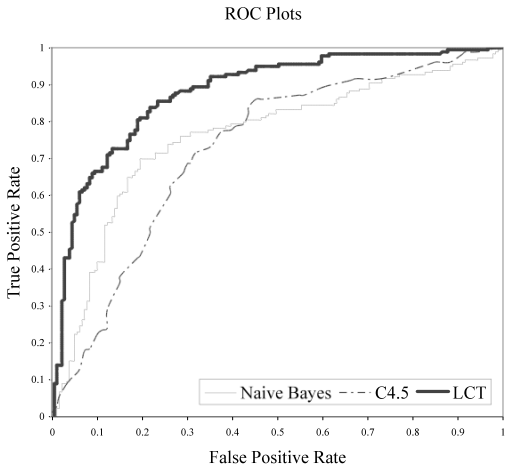

Fig 2: ROC plots for Naïve Bayes, C4.5 and 3 LCT.

Receiver Operating Characteristic (ROC) curves plot points that correspond to the number of True Positives and False Positives that result from setting various thresholds on the probability of the positive class. The Receiver Operating Characteristic Plots of Naive Bayes, C4.5 and LCT have been shown in Fig 2. The ROC curve for Naïve Bayes was plotted by changing the decision threshold for each point. The plot for C4.5 was created using an equivalent Probability Estimation Tree and varying the probability thresholds [12]. On the other hand, the ROC curve in LCT model is calculated by varying the weight of the votes that each random forest can provide.

These show the advantage of using LCT over traditional machine learning methods used for the ROP diagnosis task. The ROC curve for LCT is closer to the ideal ROC characteristics; hence it would exhibit greater sensitivity at higher levels of specificity, than the other methods.

Conclusions

The problem of predicting incidence of ROP on premature babies has been carried on most of the efficient classifiers and results observed concludes that highest accuracy (84.36%) achieved, is by 3 layered LCT model, which also outperforms other classifiers in terms of accuracy in both the problems i.e. two class problem and a three class problem. Each of the classifiers with their best accuracy configuration has been bundled in the diagnosis application "ROP Classification Machine" which can be used directly as a tool to preprocess and classify the ROP data.

The lower false-negative rate of LCT in three class problem supports its use as an efficient diagnosis system, as the progressed ROP will have a much smaller probability of misclassification.

As the medical diagnoses system needs higher confidence of sensitivity instead of accuracy, sensitivity being low in our observation shows that the prediction models still needs to be worked on and more concentration should be on how to handle the biased data. More work is required with missing feature values as well.

Ensemble classifiers are giving more accurate results thus further work can be carried on implementing a more efficient classifier like LCT to increase the accuracy. Still then, the increase in accuracy to 84% from around 60% reported in earlier studies gives enough support to ophthalmologists to use LCT as the diagnoses system in the "ROP Classification Machine" tool to predict ROP.

References

[1] L. Breiman. 1996. *Bagging predictors*. Machine Learning, 24(2):123–140.

[2] C.J.C. Burges. 1998. *A tutorial on support vector machines for pattern recognition*. Data Mining and Knowledge Discovery, 2(2):121–167.

[3] M.A. Hall. 1998. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato.

[4] T.K. Ho. 1998. *The random subspace method for constructing decision forests*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(8):832–844.

[5] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. 1998. *Boosting the margin: A new explanation for the effectiveness of voting methods*. The Annals of Statistics, 26(5):1651–1686.

[6] N. Hussain, J. Clive, and V. Bhandari. 1999. *Current incidence of retinopathy of prematurity, 1989-1997*. Pediatrics 104(3):e26-33.

[7] T.G. Dietterich. 2000. *Ensemble methods in machine learning*. Lecture Notes in Computer Science. 1-15: Springer-Verlag.

[8] L. Breiman. 2001. *Random forests*. Machine Learning, 45(1):5–32.

[9] I. Kononenko. 2001. *Machine learning for medical diagnosis: history, state of the art and perspective*. Artificial Intelligence in Medicine, 23(1):89–109.

[10] I. Rish. 2001. *An empirical study of the naive bayes classifier*. In Proceedings of IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence.

[11] S. Ruggieri. 2002. *Efficient C4.5*. IEEE transactions on knowledge and data engineering, 14(2):438–444.

[12] C. Fe Ferri, P. Flach, P. and J. Hernandez. 2003. *Improving the AUC of Probabilistic Estimation Trees*. Proceedings of the 14th European Conference on Machine Learning, 121-132.

[13] M. Dundar, B. Krishnapuram, J. Bi, and B. Rao. 2006. *Learning classifiers when the training data is not iid*. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'06), 756-761.

[14] R.O. Duda, P.E. Hart, and D.G. Stork. eds 2000. *Pattern Classification*: Wiley Interscience.

Acknowledgements

The authors would like to acknowledge the support they received from the Sankar Nethralaya Eye Hospital. They would like to thank Dr. R. R. Sudheer and Dr. Krishnendu Nandi for their crucial help in validation of the ROP data. The authors would also like to thank Abhishek Ghose for his contributions to the final outcome of this paper.