Journal of
Health & Medical Informatics

Pooja and Pushpalatha, J Health Med Informat 2019, 10:2

Research Article Open Access

A Comparative Performance Evaluation of Hybrid and Ensemble Machine Learning Models for Prediction of Asthma Morbidity

Pooja MR[1,2*] and Pushpalatha MP[2]

[1]Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India
[2]Department of Computer Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, Karnataka, India

Abstract

One of the chronic respiratory diseases that affect a large proportion of the population is Asthma. Asthma is more prevalent in children of age groups 6-14 years. Early identification of the risk factors is an important intervention towards the management of the disease as the disease is progressive in nature. In our work, we assess the performance of the two machine learning approaches with respect to their accuracy in predicting the outcome of asthma disease after identifying the critical risk factors that help in the prognosis of the disease. We perform an empirical analysis of ensemble and hybrid machine learning models to deduce the best performing approach for the prediction of the outcome of asthma. The Neyveli rural asthma dataset of India, representing cross sectional study data gathered through questionnaires formulated under ISAAC study was used to validate our approach. The outcome is predicted using both, sequential and parallel ensemble learning techniques as well as the hybrid machine learning model and we suggest the best performing ensemble learning technique on the dataset under consideration. The problem of class imbalance is well handled before presenting the data to the model as unbalanced data sets are seen to have a negative impact on classification performance.

Keywords: Ensemble; Hybrid; Cross validation; ISAAC; Feature scoring

Introduction

Asthma is a chronic respiratory disorder of the airways characterized through an obstruction of airflow, which can also be completely or partially reversed with or without specific therapy. The disease is characterized by patterns that are recurrent and the symptoms that characterize it may be time dependent such as occurrence during nights or early mornings, during specific seasons or during or after exercise. Further flare ups in the disease can be observed when exposed to non-allergic triggering factors such as cold/pollutant air and allergic trigger factors such as mites, pets or pollens. It is characterized by symptoms such as shortness of breath, wheezing and chest tightness. Asthma basically causes the airways to go slender which leads to difficulty in breathing, further since in children the size of the airways is comparatively less, they are more susceptible to its severity once affected by it [1].

Reasons that attribute to the asthma care barriers especially in rural children include lack of primary care providers, pulmonary specialists and availability of insurance benefits. Further, lack of personalized asthma interventions also pose an important challenge to handling issues related to asthma in children from rural background. Thus there is a strong need for the effective prediction of asthma outcome in rural children exhibiting a variable set of risk factors. Successful prediction of risk for asthma control deterioration at the individual patient level would enhance self-management and enable early interventions to reduce asthma exacerbations [2,3].

Here, we present our approach to predict the likely outcome of asthma disease in a cohort population comprising of 13-14 years old children residing in the Neyveli, a rural district of Tamilnadu in India. In recent years, ensemble learning methods are preferred over traditional machine learning techniques because they result in machine learning models that yield considerably good and accurate results when compared to the weak learners or standard base classifiers such as support vector machine, K-Nearest Neighbor and logistics regression for the purpose of prediction. Here, we employ the two variants of ensemble techniques, sequential (Stochastic Gradient Descent-Adaboost and Logistic Regression-Adaboost) and parallel (Random forest) ensemble learning techniques to perform this task as they are expected to improve the performance of base learners. Also we infer the choice of optimal parameters for the respective ensemble techniques. A combined feature scoring technique that aggregates the results of different feature ranking methods is deployed to form the reduced feature subset. We use some of the major performance evaluation metrics to assess and conclude the best ensemble approach that can be relied on for the purpose of asthma outcome prediction.

Hybrid machine learning models which use a combination of supervised and unsupervised learning techniques are known to result in good performance when compared to the techniques applied individually. Hence, we have also deployed the hybrid decision model that was developed by us in the previous work for the dataset under consideration and compare the results obtained with that of the ensemble techniques and some of the classic classifiers widely employed for the purpose of predicting the most common diseases [4]. It can be clearly observed from the results that the hybrid model outperforms all the classifiers including ensemble model.

Our work aims at identifying the critical risk factors which are

***Corresponding author:** Pooja MR, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India, E-mail: pooja.mr@vvce.ac.in.

Received January 28, 2019; **Accepted** March 20, 2019; **Published** April 05, 2019

Citation: Pooja MR, Pushpalatha MP (2019) A Comparative Performance Evaluation of Hybrid and Ensemble Machine Learning Models for Prediction of Asthma Morbidity. J Health Med Informat 10: 330.

used to predict the likely outcome of the disease. The comparative study performed is used to propose the best possible approach depending on the nature of the data available. If the study scenario involves a balanced class problem containing approximately equal number of healthy and asthmatic subjects, the hybrid decision support system can be used to predict the outcome with a high degree of accuracy. However the approach can best be deployed provided the data has binary attributes unlike the ensemble approach which can handle heterogeneous attributes. Identification of the features *via* feature reduction technique is very crucial as this is an important step that leads to final prediction of the disease outcome irrespective of the approach deployed.

The rest of the paper is organized as follows: In Section 2, we present the related work in the area of machine learning as applied to findings in the context of asthma related issues, followed by a discussion of the various methodologies involved in the ensemble and hybrid approaches in Section 3. Section 4, presents the results obtained and discussions justifying the inferences drawn. The last section provides concluding remarks about the entire work presented in the paper.

Review of Related Work

One of the challenging arenas in computer science is health informatics where a huge number of machine learning approaches and techniques are applied to perform tasks involving improvement of quality of life as well as quality of care [5]. Machine learning approaches are preferred because of their ability to increase prediction accuracy and one such approach involving random forest and boosting ensemble techniques was used by Goto et al. in to efficiently predict clinical outcomes such as critical care and hospitalization with considerably good sensitivity [6]. Ensemble methods perform classification by taking a vote of the predictions obtained by a set of classifiers. They are known to perform better than individual classifiers and more recent algorithms of this variety include boosting and bagging. An ensemble classifier is likely to use a set of individual base learners that are more correctly called as accurate and diverse learners, accurate because the error rate with these classifiers is far less than random guessing and diverse because not all base classifiers converge towards the same classification results [7]. Boosting is likely to enhance the performance of the base learners because they modify the weights given to training data depending on the degree to which the classification/regression is done correctly in the previous stage [8]. The problem of class imbalance can be handled *via* many approaches and is not attributed to the degree of imbalance alone and can be attributed to various other factors like lack of representative data, and contributing factors like overlapping classes and disjoints and ways to handle them can be dealt at both data and algorithmic levels [9]. In ensemble methods involving single decision trees, forest of decision trees and decision tree boosts were used to distinguish between seasonal air qualities and predict indices pertaining to air quality assessment and they were seen to outperform benchmarked machine learning classifier such as SVM [10]. Busatlic et al. employed ensemble techniques based on decision trees that yielded precision accuracy of 73.3% when deployed for the prediction of skin permeability, which was better than the accuracy obtained by other classifiers [11].

Different approaches involving logistic regression, classification/regression trees and spline modeling (basically a non-parametric approach) were deployed to perform asthma phenol typing [12]. In an attempt was made to predict ozone concentrations using artificial intelligence techniques-multiple linear regression, neural networks, support vector machine, random forest, and two ensemble techniques: linear ensemble and greedy ensemble. It was clearly evident that the ensemble methods especially, linear ensemble model outperformed the others [13]. Neural network with fuzzy membership function (NEWFM) method along with Adaboost was applied to four different medical datasets and in all the cases the accuracy of the disease classification and diagnosis was improved as was projected by the results [14]. In a Bayesian approach was proposed by Ananthi et al. [15] for genomic expressions characterizing asthma to find the severity levels of the disease which were classified as low, medium and high and this approach yielded good results in terms of predicting asthma outcome. A hybrid decision support system was used in to predict the outcome of asthma using the ISAAC dataset for the urban cities of India, namely New Delhi and Bombay [4]. The system however used the complete dataset where in there was a clear indication of class imbalance between the number of asthmatic and non-asthmatic subjects.

Methods

In this section, we discuss the dataset deployed and present an overview of the various techniques involved in the ensemble model including the feature scoring methods involved in the preprocessing stage. The approach employed for hybrid model is also discussed here.

Data description

We have used the Neyveli Asthma dataset for the work we have presented, which is publicly made available as part of ISAAC (International Studies of Asthma and Allergies in Childhood) questionnaire for the age group of 13-14 years. The various attributes/features used in the dataset characterize symptoms such as wheeze, shortness of breath, sleep disturbance, frequency of recurrence in symptoms, speech limitation arising from wheeze, nose irritations, presence of hay fever, rashes, breathing problems following exercise and gender.

Feature scoring

We have used the following feature scoring methods for the combined feature scoring technique used in the preprocessing stage.

ANOVA (Analysis of one way variance): the difference between the average values of the feature in different classes, Chi^2: the dependence as measured using chi-square statistics between the class and the feature, Relief: the capacity of an attribute to distinguish between classes with respect to the data instances that are similar and FCBF (Fast Correlation Based Filter): an entropy-based measurement which identifies redundancy that arises because of pairwise correlations between features. Here, we try to assign scores to individual features using each of the feature scoring techniques namely. The average score is obtained for each of the features *via* different techniques by applying weighted averaging and the first 12 features with the highest scores which make up 25% of the original feature space are drawn to constitute the reduced feature set. A weighted average of all the features chosen using the above methods is taken to form the feature subset.

Ensemble models

Parallel ensemble learning-random forest: Random forest is a parallel ensemble technique where, in addition to taking a random subset of the data, a random subset of features is used for the training. However, here we opt to take a feature subset which is obtained by combined features scoring technique rather than a random set of features. In the process many random trees are used and the output of several trees is used to predict the final outcome. We have restricted the number of trees to 10 and the number of attributes at each split to 2. Further, the depth of the individual trees is limited to 3. Random forests can handle the problem of multi dimensionality as well as data with missing values apart

from assuring more accurate results. However, in the case of regression it is not much preferred as the final prediction would be based on the mean predictions.

Sequential ensemble learning-gradient boosting: The Gradient boosting ensemble is a combination of Stochastic Gradient boost and Adaboost. The boosting technique that often uses many weak classifiers and combines them to approximate the Bayes classifier C∗(x) is the Adaboost technique which starts with the unweighted training sample, then builds a classifier, for instance a classification tree, that yields class labels. If a training sample is misclassified, the weight of that sample is boosted. A second classifier is constructed using the new weights, which are no longer the same but may approximate the previous. Again, misclassified training data have their weights boosted and the process is repeated. Typically, one may also construct thousands of classifiers this way. A score is assigned to every classifier, and the final classifier is described as the linear aggregate of the classifiers from every stage. Adaboost results in considerably good results when deployed for a two class, classification problem. However, it is not the case for multi-class problems, although Adaboost used to be additionally proposed to be used in the multi-class case.

Stochastic Gradient Descent uses the concept of batching data over iterations in order to avoid redundancy when handling larger datasets and takes the idea of batching to the extreme level possible where there could be only one sample per iteration. The single sample was chosen and comprising the batch is drawn randomly thus characterizing the stochastic approach. However the samples are to be shuffled after every iteration so as to get a better performance. Also, since the SGD approach tries to optimize a differential function iteratively it can lead to smoother convergences. We have used both the variants of the classification algorithm Stage-wise Additive Modeling using a Multi-class Exponential loss function namely, SAMME and SAMME.R to test the performance and SAMME.R which updates the weights of the base estimators by estimating the probabilities was seen to be more effective when compared to SAMME which updates the weights of the base estimators by just looking into the classification results.

Hybrid model

Here, we use a two-step approach where we initially perform feature clustering to generate the reduced feature set by applying Modified Fuzzy C Means (MFCM) Clustering, that incorporates a correlation based objective function. The cluster containing the asthma attribute is identified as the cluster of interest and the features that coexist with the asthma attribute are extracted to form the reduced feature set. This is followed by subject clustering where the entire set of subjects are clustered into few subject clusters. The number of subject clusters is obtained by applying subtractive clustering and two clusters were formed accordingly. It can be observed that the first cluster contains a higher number of asthmatic subjects while the second contains a higher number of non-asthmatic subjects. Further, Classification and Regression Tree (CART) is used to predict the outcome of asthma morbidity within the individual clusters. The significance of the hybrid model lies in its ability to predict efficiently even when deployed on a smaller dataset that addresses the class imbalance problem by containing approximately equal number of positive and negative classes.

Results and Discussion

The conventional model evaluation methods do not accurately measure model performance when faced with imbalanced datasets and hence we go with the solution of balancing the classes in the dataset. Random Under-Sampling is the approach we have adopted for our model to handle the problem of imbalance. Random Under sampling aims to balance class distribution by randomly eliminating majority class examples. This is done until the majority and minority class instances are balanced out. Table 1 shows that the hybrid approach is in fact an optimal one to obtain highly accurate results, when balanced class problem is presented to the model. Further, the approach works well in a scenario wherein the attributes are binary in nature.

Initially the dataset containing information of 3281 subjects with respect to 58 features/attributes is checked for the class imbalance by finding the number of samples with asthma outcome as indicating the presence of asthma and those with the outcome as indicating the absence of asthma. However, the samples with the value 9 for the target attribute asthma are ignored as they indicate a lack of clarity with respect to the attribute. The irrelevant and redundant features are filtered resulting in a dataset consisting of 49 features. A total of 78 subjects were identified to have asthma while a total 3203 were identified to not have asthma. This distribution indicates a clear problem of imbalance. Thus, we opted to choose a collection of about 64 samples which constitute nearly 2% of the total samples drawn from the set of subjects with asthma outcome as this leads to a balanced class problem where we finally have a positive class with 78 samples and a negative class with 64 samples. We have used ANOVA, Chi^2, Relief and FCBF feature scoring techniques and finally apply combined feature scoring technique to find the most relevant features that are applicable for the problem of classification. This gives rise to a reduced feature set that characterizes asthma morbidity and is used in the successive stages for the prediction of asthma outcome. About 25% of the features from the original feature space are used in reduced feature set.

Table 2 depicts the reduced feature set for the balanced dataset respectively using the combined feature scoring technique discussed above. The features that constitute the reduced feature set are as explained below:

whezev- Wheeze ever; whez12- Wheeze in the past 12 months; nwhez12- 4 or more attacks of wheeze in the past 12 months; Awake12- Sleep disturbance from wheeze, 1 or more nights a week in the past 12 months; Speech12-Speech limited by wheeze in the past 12 months; Ieyes12-Nose and eye symptoms in the past 12 months; pnosejan, pnosefeb, pnosemar, pnoseapr, pnosejun, pnosenov- Nose symptoms in the respective months.

In Table 3, we present the results for the balanced dataset using 10 fold cross validation. The results for the model are tabulated in terms of classification accuracy, precision, recall and F1measure. The AdaBoost classifier works well when SAMME.R algorithm is used as a classification loss function. However it offers the same performance with all the three regression loss functions namely square, exponential and linear. Further, Ridge regularization is performed with strength of 0.00001. The rate of learning follows an inverse scaling approach with an initial learning rate of 0.0100 and inverse scaling exponent as 0.2500. In all the above cases, ridge regression was used as the regularization function as it was experimentally verified that the Lasso/L1 performs

Cluster 1				Cluster 2			
Predicted		Actual		Predicted		Actual	
		Positive	Negative			Positive	Negative
	Positive	61	0		Positive	17	0
	Negative	0	15		Negative	0	49

Table 1: Confusion matrix obtained using hybrid model.

low when compared to Ridge/L2 regularization.

Table 4 presents the comparative results obtained by applying some of the common classifiers including Naïve Bayes, Support Vector machine, Logistic Regression and KNN along with those obtained by the ensembles deployed in the work.

The results obtained using the Hybrid model are shown in the form of a confusion matrix. Subtractive clustering when applied on the balanced dataset indicates that it is preferable to have two clusters. As such, the subjects were divided into two clusters. The total number of subjects was 76 and 66 in the cluster1 and cluster2 respectively. The results in the table indicate that the model offers a precision and recall of 1.0 as the number of false positives and false negatives is zero in both the clusters.

Figure 1 below shows the ROC (Receiver Operating Characteristic) curves for both, sequential and parallel ensemble models, for the balanced class dataset. The ROC curves in figure represent the merged predictions from folds, where the curves in green and orange represents the ROC's for Gradient boost and Random forest ensemble models respectively.

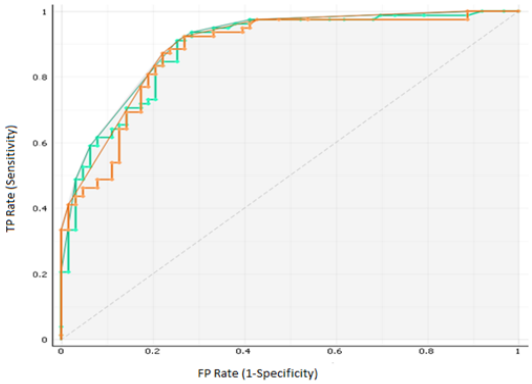

Figure 1: ROC obtained by merged predictions from folds.

Features	ANOVA	Chi²	ReliefF	FCBF
whez12	68.46506	32.9873	0.34975	2.44E-05
speech12	46.0072	15.62986	0.166	1.90E-05
awake12	45.95865	16.20579	0.16875	1.95E-05
nwhez12	45.2193	15.86914	0.17675	2.28E-05
ieyes12	39.45029	19.85013	0.05275	0.18139
whezev	36.38954	21.7419	0.05325	0.3369
pnoseapr	35.25883	15.92126	0.015	1.42E-05
pnosejun	35.25883	15.92126	0.01925	1.42E-05
pnosenov	34.94908	8.769292	0.00775	1.38E-05
pnosemar	34.92739	16.59564	0.01075	1.43E-05
pnosefeb	34.79292	15.9749	0.00725	1.42E-05
pnosejan	34.7286	13.6062	0.00975	1.39E-05

Table 2: Feature subset with respective scores using different feature ranking techniques.

Method	F1	Precision	Recall
SAMME			
Gradient boost	0.744	0.768	0.768
Random Forest Learner	0.774	0.803	0.803
SAMME.R			
Gradient boost	0.787	0.818	0.817
Random Forest Learner	0.774	0.803	0.803

Table 3: Classification results for the ensemble models using stratified 10-fold cross validation.

Method	F1	Precision	Recall
Gradient boost	0.787	0.818	0.817
Random Forest Learner	0.774	0.803	0.803
Naive Bayes	0.783	0.797	0.789
SVM Learner	0.773	0.813	0.810
Logistic Regression	0.708	0.733	0.732
KNN	0.717	0.761	0.761

Table 4: Comparative results for various classifiers *via* stratified 10-fold cross validation.

Conclusion

Since, Asthma is a chronic respiratory disease which is irreversible in nature, it is at most important to identify morbid features characterizing the disease at an earlier stage. In our work, we suggest a hybrid machine learning model that can be used as a reliable model for the effective prediction of asthma morbidity by performing a comparative analysis of ensemble and hybrid models. The performance of the models was tested using 10 fold cross validation and the important performance evaluation metrics characterizing classification of asthma morbidity were evaluated. The Gradient boost ensemble model offers a precision and recall of 82% while the hybrid model offers a precision and recall of 100% when deployed on the balanced dataset. Overall, the hybrid approach deployed in our work can be used as an important and efficient tool to infer the likely asthma outcome by identifying the features characterizing asthma morbidity.

References

1. Do Q, Son TC, Chaudri J (2017) Classification of asthma severity and medication using tensorflow and multilevel databases. Procedia Comput Sci 113: 344-351.
2. Luo G, Stone BL, Fassl B, Maloney CG, Gesteland PH, et al. (2015) Predicting asthma control deterioration in children. BMC Med Inform Decis Mak 15: 84.
3. Amaral JL, Lopes AJ, Veiga J, Faria AC, Melo PL (2017) High-accuracy detection of airway obstruction in asthma using machine learning algorithms and forced oscillation measurements. Comput Methods Programs Biomed 144: 113-125.
4. Pooja MR, Pushpalatha MP (2015) A hybrid decision support system for the identification of asthmatic subjects in a cross-sectional study. Emerging Research in Electronics, Computer Science and Technology (ICERECT).
5. Nithya B, Ilango V (2017) Predictive analytics in health care using machine learning tools and techniques. International Conference on Intelligent Computing and Control Systems (ICICCS).
6. Goto T, Camargo CA Jr, Faridi MK, Yun BJ, Hasegawa K, et al. (2018) Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. Am J Emerg Med 36: 1650-1654.
7. Dietterich TG (2000) Ensemble methods in machine learning. International workshop on multiple classifier systems. Springer, Berlin, Heidelberg.
8. Cao DS, Xu QS, Liang YZ, Zhang LX, Li HD (2010) The boosting: A new idea of building models. Chemom Intell Lab Syst 100: 11.
9. Park Y, Ghosh J (2014) Ensembles of (α)-trees for imbalanced classification

problems. IEEE T KNOWL DATA EN 26: 131-143.

10. Singh KP, Gupta S, Rai P (2013) Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos Environ 80: 426-437.

11. Busatlic E, Osmanović A, Jakupović A, Nuhić J, Hodžić A (2017) Using neural networks and ensemble techniques based on decision trees for skin permeability prediction. CMBEBIH 2017, Springer, Singapore.

12. Brasier AR, Ju H (2014) Analysis and predictive modeling of asthma phenotypes. Heterogeneity in Asthma. Humana Press, Boston, MA.

13. Bing G, Ordieres-Meré J, Cabrera CB (2015) Prediction models for ozone in metropolitan area of Mexico City based on artificial intelligence techniques. Int J Inform Decis Sci 7: 115-139.

14. Abuhasel KA, Iliyasu AM, Fatichah C (2015) A combined AdaBoost and NEWFM technique for medical data classification. Inform Syst Manage. Springer, Berlin, Heidelberg.

15. Ananthi S, Prathiba L (2018) An improvised technique for the diagnosis of asthma disease with the categorization of asthma disease level. Information Systems Design and Intelligent Applications. Springer, Singapore.