NVEO
Natural Volatiles &
Essential Oils

Performance Analysis of Enhanced Adaboost Framework in Multifacet medical dataset

Dr. Sudharson D[1], P. Divya[2], DR.D.Palanivel Rajan[3], Ratheeshkumar A.M[4]

[1]Assistant Professor, Dept of AI&DS, Kumaraguru College of Technology, Coimbatore, India. sudharsondorai.ads@kct.ac.in
[2]ASSISTANT PROFESSOR, Dept of CSE, BANNARI AMMAN INSTITUTE OF TECHNOLOGY, divisrecme@gmail.com
[3]Professor, Dept Of Cse Cmr Engineering College,Hyderabad Palanivelrajan.D@Gmail.Com
[4]Assistant Professor Department of ITratheeshkumar@skcet.ac.in

Abstract:
Predictions that are made based on features are performed through machine learning (ML) algorithms. Machine learning allows systems to learn and develop on their own by gaining experience. In the field of artificial intelligence, machine learning is a sub-discipline. Supervised and unsupervised learning are the two prevalent categories under machine learning. Supervised ML is used for classification whereas unsupervised ML is used for clustering. Currently, machine learning is being employed in a plethora of fields. Biometric recognition, handwriting recognition, and medical diagnosis are some of the use cases of ML. A significant role is played by machine learning in the medical field: identify diseases based on a patient's characteristics. Software applications based on ML algorithms are helping doctors in diagnosing various diseases like cancer, cardiac arrest, etc. We employed an ensemble learning strategy to predict heart problems in this paper. Through the comparison of different evaluation parameters namely ROC, F-measure, recall, precision and accuracy, our paper describes the performance of ML algorithms. The study used a mix of machine learning classifiers to predict heart problems, including Naive Bayes (NB), Decision Tree (DT), Random Forest (RF) and Support Vector Machine (SVM) algorithms. It was observed that implementing Paretto Distribution enabled adaboost resulted in 98.61% accuracy. NB, DT, RF and SVM models were also trained and tested separately.

Introduction

Oversampling strategies may be utilized to reproduce these outcomes for a more equal proportion of positive results in training when one class of data is the disadvantaged class in the data sample. When the volume of data is inadequate, oversampling is used. SMOTE (Synthetic Minority Over-sampling Technique) is a renowned over-sampling technique that produces synthetic samples by dynamically sampling characteristics from events in the minority class [1].

In contrast, if a data class is highly represented by the majority, under-sampling may be used to manage it with the minority class. When the data accumulated is adequate, under-sampling can be used. Under-sampling methods widely used include cluster centroids and Tomek links, both of which aim to minimize the proportion of dominant data by targeting potential intersecting features inside the retrieved data sets.

Simple data duplication is seldom recommended in both over and under-sampling. In fact, oversampling is desirable to undersampling because undersampling can lead to the loss of valuable data [4]. When the information captured is significantly larger than ideal, under-sampling is recommended to assist data mining tools to stay inside the boundaries of what they can proficiently perform [6].

Literature survey

A lifelong condition, diabetes arises when the pancreas is incapable of secreting insulin or when the body is unable to effectively use the insulin produced [4,6]. Glucose from meals enters the bloodstream through the pancreas, where it is converted into energy. Insulin is a hormone produced by the pancreas. Everything that contains carbohydrates is turned into glucose in the blood. As a result, glucose is more easily absorbed by our cells when insulin is present. Type 1 diabetes, type 2 diabetes, and gestational diabetes are the three types of diabetes, amongst which type 2 is the most prevalent. It affects adults and accounts for roughly 90% of all people diagnosed. While we have type 2 diabetes, our bodies do not properly utilize the insulin that we produce. Type 2 diabetes treatment relies on a healthy lifestyle, frequent physical activity, and an adequate diet. Most patients with type 2 diabetes, on the other hand, will however need oral medicines and/or insulin to keep their blood sugar levels in check.

A big data-applicable Convolution Neural Network-based multimodal disease risk prediction (CNN-MDRP) algorithm was introduced by Kai Hwang et al. The diabetes risk model is created by combining structured and unstructured features and its accuracy is evaluated. It outperformed the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm in terms of accuracy. The data used in the study came from a Chinese hospital and included EHR, medical image data, and gene data. The data focuses on inpatient department data, which is predominantly made up of structured and unstructured text data.

In order to train the parameter in the CNN-MDRP algorithm, the stochastic gradient descent algorithm is used which is mainly employed in big data applications. Big data analytics, powered by the Hadoop framework, has revolutionized how healthcare professionals use advanced technologies to gain insights from clinical datasets and draw conclusions. By predicting the needs and demands of people, effective data-driven services can be delivered to them, as a result of which effective healthcare management can be attained. In healthcare, big data analytics is described as gathering, storing, processing, and analyzing a huge amount of data surrounding health in multiple forms and offering relevant information to users, allowing them to quickly uncover business value and insights. Data mining, machine learning, statistical analysis, and visualization are several techniques used for big data analytics.

A semi-automated framework which is machine learning-based was introduced by Ya Zhang and Tao Zheng, that makes use of a big data EHR database [8]. In China, data from 15 local EHR systems were instantly stored in a centralized repository every 24 hours. A supervised learning algorithm served as the foundation for the framework. Feature Engineering was required to properly frame the frequently unstructured and sparse raw EHR. A total of 16 features were extracted and built to be utilized in the machine learning framework. The machine learning algorithms Random Forest, Logistic Regression, etc were analyzed against enhanced Adaboost were used, and the results were better. The algorithms further enhance the filtering characteristics to improve recollection while minimizing false-positive rates.

An automatically analyzing machine learning prediction results automatically was introduced by Gang Luo. Predictive modeling is a process that estimates results using data mining and probability [9]. Each model consists of a number of predictors, which are variables that are likely to impact potential results. A statistical model is developed after data for relevant predictors has been gathered. A simple linear equation may be used in the model, or it may be a complex Neural Network that has been plotted out by advanced tools. The statistical analysis method is verified or modified as new data becomes available. Predictive analytics can help with healthcare, profitability, and better outcomes throughout the value-based care continuum. Rather than simply presenting consumers with information about past events, predictive analytics forecasts the possibility of a positive outcome based on trends in past information. The electronic health record source data from the Practice Fusion diabetes classification competition, which included medical files from all 50 states in the United States, was used in this work, which demonstrated a model for estimating type 2 diabetes diagnosis relatively soon. To make the prediction, two models were used: one for prediction, and the other to explain it. The first model is mostly used to make predictions and is largely concerned with accuracy rather than interpretability. It can be any machine learning model, no matter how complex. The latter is a rule-based associative classifier that is merely used to explain the first model's outcomes with less or no regard for its accuracy.

Enhanced AdaBoost in prediction

Although the AdaBoost framework's algorithms vary, AdaBoost.M1 is the one that is chosen to be discussed here. The purpose of this research is to diagnose a certain illness using HR data using an issue of binary classification. The identical algorithm is produced by AdaBoost.M1 and AdaBoost.M2, but the first one is less complicated when compared to the second.

Pseudoframework1 narrates the phenomenon of AdaBoost.

1. All instance boosting weights, B1,n(n = 1,, N), are initialised as 1/N..
2. In the further steps, weak classifier learning is repeated after initialization.
3. In next step the $(qx)^{th}$ weak classifier v_{qx} is trained in order to minimise the following objective function K_{qx}: which is of $K_{qx} = \Sigma n = 1NB_{qx}, nI(hqx, n \neq yn)$

4. where $I(h_{n,q} \neq y_n)$ is an indicator function that returns 1 if $h_{n,qx} \neq y_n$ is true and 0 if it is false.
5. In future steps, the error aε_{qx} is calculated.
6. The the above iterations are used to change a parameter β_{qx} and the increasing weights as $B_{qx,n}$:$hqx, n = yn\ otherwise\ \ B_q + 1, n = B_{qx}, nZ_{qx} \times \{\beta_{qx}1if$ where the normalising constant is Z_{qx}.
7. The final classifier H(x) is built as a weighted vote of the Q weak classifiers after Q iterations, as $H(x) = arg\ maxy\epsilon Y\Sigma q{:}\, hq = ylog(\frac{1}{\beta_q})$

Framework Implementation

A prominent and very well-known supervised ML algorithm is the Decision Tree. Each node signifies a characteristic, the decision rule is represented by each link or branch, and the leaf signifies the actual classified class in a hierarchical structure. For classification problems, decision trees are commonly used. A decision tree can be constructed using well-used algorithms such as ID3, C4.5, CART, and J48. In this approach, there are two key steps, first is to build a tree, and next training and testing the tree to the dataset. A supervised ML model that may be utilized for classification and regression is the Support Vector Machine (SVM). To perform classification, SVM builds a hyperplane in between data points. The hyperplane is designed in a way that it fits between two classes as much as possible. Support vectors are the data points closest to the hyperplane. Margin is the distance between the support vectors and the hyperplane. The primary goal of SVM is to get the best-fit hyperplane in N-dimensional space. The easiest way to categorise two classes is to use a hyperplane with the most margin. Naive Bayes (NB) is a well-known supervised machine learning method that is widely used for classification. The Bayes theorem is used in this method of classification. According to the NB theorem, the likelihood of an event occurring may be calculated using the probability of another event occurring.

To begin processing the dataset, we must first clean it. The missing and unnecessary elements of the data are initially detected during data cleansing. If there are multiple missing values in certain rows of a dataset, that row from the dataset is deleted. To fill up missing data, we may use either the average of current values or the mean squared values. The removal of noisy data is another task in data cleaning.

A noise is said to be when there is a different data type, for example, age can be considered as an aspect whose intended value is of the type integer, but instead, there is a string value present in its place. Various methods, like the mean of all values in a column or the use of bounding parameters, can be used to replace noisy data. Data integration is the second stage in data preparation. This phase combines scattered data into one sequential unit in order to signify data in a similar range. Since our data isn't scattered, we skip this phase. High computation resources are required to process large amounts of data. When data is large, analysis becomes a lengthy process. When we have a large amount of data, data reduction is a crucial component. Data reduction is the third stage in data preparation dealing with large datasets. Since our study is based on a small dataset, we omit this stage in data preparation.

Data transformation is the process of converting raw data into a dataset that meets the needs of the project. Smoothing, aggregation, and generalisation are some of the techniques used in data transformation. We employed data transformation to train our model in our work. We divided the dataset into 80 % training and 20 % testing after preprocessing. We separated our work into two phases in this study. On the medical dataset, we trained several machine learning algorithms like Naive Bayes, SVM, The decision tree, and Random forest. We examined each model's performance (accuracy, F-measure, ROC, and precision). Decision tree's accuracy was 70%, SVM's accuracy was 77.4074 %, and Random forest's accuracy was 79.2593 %. Then, using several combinations such as NB with DT, NB with SVM, NB with RF, DT with RF, DT with SVM, and RF with SVM, we create ensemble models. All of the assessment criteria showed that NB with DT performed better. The following diagram depicts the overall procedure:

Conclusion

A confusion matrix was used to determine the accuracy of the classifiers created using various ML algorithms. The matrix's diagonal elements shown below signify the number of points where the prediction is correct and is labeled as a "true label" (TP/TN), whereas the off-diagonal elements of the matrix signify the number of points where the classifier is incorrect and is labeled as a "false label" (FP/FN). True positives (TPs) are cases in which we predicted true(they have the disease) and they do. True negatives (TN) occur when the classifier predicts no and the individual does not have the disease. False positives (FP) occur when the classifier predicted yes but the person does not have the disease. False negatives (FN) occur when the classifier predicts no but the person has the disease. The greater the number of diagonals in the confusion matrix, the greater the model's denoting power to form accurate predictions [15]. The accuracy rate is the classifier's overall performance, i.e. how often is the classifier correct? It is calculated as (TP+TN)/total, or it can be found in the classification report as weighted avg. The misclassification/error rate is the overall frequency with which the classifier is incorrect. It is calculated as follows: (FP+FN)/total OR 1-accuracy rate. Table 1 presents a schematic review of different parameters taken from the confusion matrix (TP, TN, error rate) as well as the classification report (accuracy or weighted average) for the various combinations used, which aids in determining which model combination should be selected as the prediction model. It obviously demonstrates that the ensembled model's parameters are the right choice.

ML Model	**TP**	**TN**	**Accuracy rate (%)**	**Error rate (%)**
Logistic Regression + K-Fold CV	300	210	96.43	3.57
KNN + GridSearch	280	210	96.93	3.07
Decision Tree + GridSearch	280	200	97.18	2.82
Random Forest	280	210	96.89	3.11
SVM + GridSearch	280	200	96.89	3.11
Adaboost	280	210	98.86	1.14

References

1. Sudharson, Dorai; Prabha, D; "A novel machine learning approach for software reliability growth modelling with pareto distribution function" Soft Computing Vol.23, issue.18, pp.8379-8387 2019, Springer Berlin Heidelberg.
2. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
3. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
4. Sudharson, D Prabha, D "Improved EM algorithm in software reliability growth models" International Journal of Powertrains, Vol.9 Issue.3, pp.186-199, 2020,Inderscience Publishers (IEL).
5. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
6. Ratheesh Kumar A.M, Dr. Sudharson D, P.Divya, Dr. Sakthi Govindaraju, "A NOVEL AI AND RF TUTORED STUDENT LOCATING SYSTEM VIA UNSUPERVISED DATASET", Turkish Journal of Physiotherapy and Rehabilitation, Vol.32, Issue.2, pp.882-887 2021.
7. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
8. Dr.B.Arunkumar, D.Sudharson, "A Novel Approach for Boundary Line Detection using IOT During Tennis Matches", Advancement of Electrical, Information and Communication Technologies for Life Application, Volume.13, Issue.4, pp.243-246 2020.
9. Qian Yuexia, Gu Weijie 2012, “The Research on Reliability Optimization of Software System Based on Niche Genetic Algorithm” AASRI Procedia, Vol. 1, Pp. 404-409
10. Rana Özakıncı, Ayça Tarhan 2018, “Early software defect prediction: A systematic map and review” Journal of Systems and Software, Vol. 144, Pp. 216-239

11. Ramakanta Mohanty, V. Ravi, M. R. Patra 2013, “Hybrid intelligent systems for predicting software reliability” Applied Soft Computing, Vol. 13, No. 1, Pp. 189-200
12. Sangeetha M, C. Arumugam, K. M. Senthil Kumar, S. Hari Shankar 2015, “An Effective Approach to Support Multi-objective Optimization in Software Reliability Allocation for Improving Quality” Procedia Computer Science, Vol. 47, Pp. 118-127
13. Syed Wajahat Abbas Rizvi, Vivek Kumar Singh, Raees Ahmad Khan 2016, “Fuzzy Logic Based Software Reliability Quantification Framework: Early Stage Perspective (FLSRQF)” Procedia Computer Science, Vol. 89, Pp. 359-368.
14. Vahid Garousi, Mika V. Mäntylä 2016, “A systematic literature review of literature reviews in software testing “ Information and Software Technology, Vol. 80, Pp. 195-216.
15. Yongqiang Lian, Yincai Tang, Yijun Wang 2017, ”Objective Bayesian analysis of JM model in software reliability” Computational Statistics & Data Analysis, Vol. 109, Pp. 199-214.