A Narrative Summary

A Random Forest based predictor for medical data classification using feature ranking

Alam et al., Bangladesh Department of CSE, 2019
DOI: 10.1016/j.imu.2019.100180

INTRODUCTION

As artificial intelligence (AI) becomes more prevalent in our lives, its effect on medical data analysis is also growing. It's currenly being utilized to assist and help doctors and healthcare professionals in diagnosing diseases [1]. The end goal isn't to replace them, but to support their decisions. That's where decision support systems (DSS) come in. Machine learning (ML) fits into this by treating diagnosis like a classification problem, where the model tries to guess what's wrong with the patient based on existing data.

However , this isnt a simple process. Real-world medical data is quite messy, inconsistent, and hard to normalize [2]. So far, majority of the studies have been about specific cases [1-3], which limits their usefulness. This paper, however, takes does the opposite. Instead of one dataset, it tries to explore a general approach that can be applied to all (or most) medical problems.

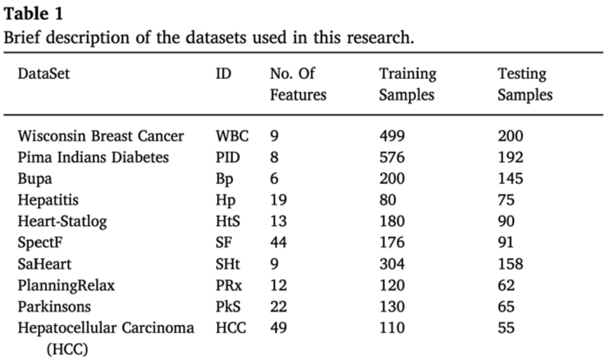

Table 1
Brief description of the datasets used in this research.

DataSet	ID	No. Of Features	Training Samples	Testing Samples
Wisconsin Breast Cancer	WBC	9	499	200
Pima Indians Diabetes	PID	8	576	192
Bupa	Bp	6	200	145
Hepatitis	Hp	19	80	75
Heart-Statlog	HtS	13	180	90
SpectF	SF	44	176	91
SaHeart	SHt	9	304	158
PlanningRelax	PRx	12	120	62
Parkinsons	PkS	22	130	65
Hepatocellular Carcinoma (HCC)	HCC	49	110	55

To achieve this, the authors worked with 10 individual datasets. They applied a feature ranking method to pick out the most relevant data points and used Random Forest as the main classification algorithm. Not only did the model perform well, but the study also highlighted the most important features in each dataset. This could actually help medical professionals understand the data better and make more informed decisions.

METHODS AND MATERIALS

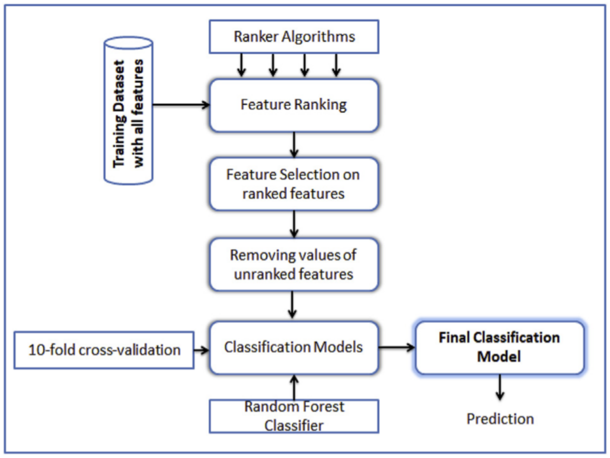
Ranker Algorithms
Training Dataset with all features
Feature Ranking
Feature Selection on ranked features
Removing values of unranked features
10-fold cross-validation
Classification Models
Final Classification Model
Random Forest Classifier
Prediction

Fig. 1. Model construction overview.

The data sets of this paper were provided by the UCI Machine Learning Repository [4]. To properly test the model, each disease dataset was randomly splitted into training and testing sets. This way, the model will not be overfitted to the data [2]. Before training, the researchers checked which features were more useful for our model by utilizing feature ranking [5-7].

After ranking the features, only the top ones were selected for the final model. Then, the Random Forest model was used to build and train the classifier. Multiple models were built for each daataset to see which ones perform better. Feature ranking was an important step, identifying and seperating useful data for making accurate predictions. Features with negative correlation coefficients were found to hurt model performance, so they were excluded.

RESULTS

In each dataset, researchers first trained a basic and beginner model using all features, then trained three more models using only the top features. After the comparison, the best-performing model was chosen. Then, only the base model and the best one were used for final testing, instead of testing all four. In almost every case, the models using ranked features gave better results. This explain the reasoning behind removing unimportant features and how it leads to improving both training and testing accuracy.

Compared to other methods, the models outperformed most of the existing approaches on many datasets . In the breast cancer dataset, a deep learning model from another study did slightly better, however small [8]. In the heart disease dataset, one other method had higher accuracy, but again, the difference was small [9]. Overall, the model in this study performed quiet well.

CONCLUSION

Essentially, classifying medical data is quite a complex achievement. To remedy that, a general method that focuses on singling out important features before model training was proposed by the authors. 10 separate databases were tested and, in each case, the results were consistently better when using feature selection instead of doing it will all features. The aim was to improve the overall accuracy while simplifying the data input process. As tested, SVM, Bayes Network, and Random Forest gave the best results regarding all the datasets. While the paper doesn't introduce new studies or theories, the consistent results makes it a strong and impressive contribution to the field.

REFERECNES

[1] Chabat F, Hansell DM, Yang G-Z. Computerized decision support in medical imaging. IEEE Eng Med Biol Mag 2000;19(5):89–96.

[2] Mohapatra P, Chakravarty S, Dash P. An improved cuckoo search based extreme learning machine for medical data classification. Swarm and Evolutionary Computation, vol. 24. 2015. p. [25]–49]. https://doi.org/10.1016/j.swevo.2015.05.003https://www.sciencedirect.com/science/article/pii/S2210650215000413.

[3] Duda RO, Hart PE. Pattern classification and scene analysis. NY, USA: A Wiley- Interscience Publication, John Wiley & Sons; 1973.

[8] Karthik S, Perumal RS, C.Mouli PVSSR. Breast cancer classification using deep neural networks. Knowledge Computing and Its Applications 2018. p. 227–41.

https://doi.org/10.1007/978\-981\-10\-6680\-1_12https://link.springer.com/chapter/10.1007/978/discretionary{-}{ }{ }981/discretionary{-}{ }{ }10/discretionary{-}{ }{ }6680/discretionary{-}{ }{ }1_12.

[9] Eshtay M, Faris H, Obeid N. Improving extreme learning machine by competitive swarm optimization and its application for medical diagnosis problems. Expert Syst Appl 2018;104:134152. https://doi.org/10.1016/j.eswa.2018.03.024https://ww. sciencedirect.com/science/article/pii/S0957417418301696.

[4] University of California at irvine (uci) machine learning repository. https://archive. ics.uci.edu/ml/datasets.html.

[5] Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997;1(1–4): [131]–56]. https://doi.org/10.1016/S1088-467X(97)00008-5http://www.lsi.us. es/~riquelme/publicaciones/kes03.pdf.

[6] Ruiz R, Riquelme JC, Aguilar-Ruiz JS. Fast feature ranking algorithm. In: Palade V, Howlett RJ, Jain LC, editors. KES 2003 (LNAI 2773). 2003. p. 325331http://www.
lsi.us.es/~riquelme/publicaciones/kes03.pdf.

[7] J. Novakovi, P. Strbac, D. Bulatovi, Toward optimal feature selection using ranking methods and classification algorithms, Yugosl J Oper Res ([21]). doi:10.2298/ YJOR1101119N.URL
http://elib.mi.sanu.ac.rs/files/journals/yjor/41/yujorn
41p119/discretionary{-}{ }{ }135.pdf.