Paper Review DSC 540

Paripon Thanthong

1927018

Ozcift (2011)

Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithm.

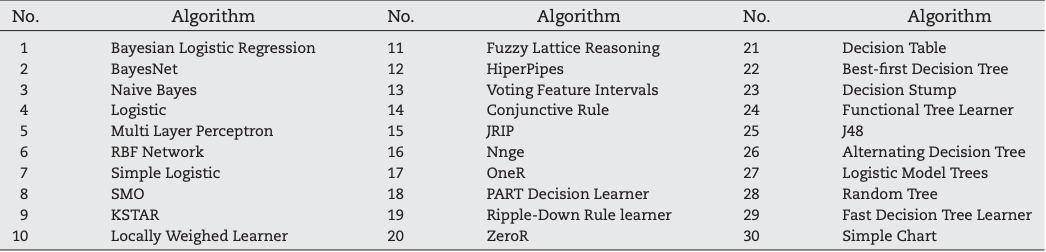
**Introduction**

In Ozcift’s paper, the authors utilized three medical datasets which are Parkinson’s, diabetes and heart diseases by using machine learning to observe the change of accuracies of the diagnosis. In the research the authors applied Rotation Forest (RF) ensemble classifier with 30 machine learning algorithms to analyze performance of the model. In addition, the authors also applied correlation based feature selection (CFS) algorithm as a feature selection method and using 10 fold cross validation (leave one out) to validate model performance. Thus, three metrics; classification accuracy (ACC), kappa error (KE), and area under the receiver operating characteristic (ROC) and curve (AUC) were used to evaluate performance. As a result, the authors compared the evaluation from based classifier method and RF method which showed the performance from RF is better from every metrics.

To begin with, the authors mainly focused on the improvement of the diagnosis predictive accuracy using machine learning which pointed out to two strategies; feature selection method and construction of classifier ensembles. Accuracy was the focusing point of this research, the authors indicate the problem that too many features were selected to the model that can cause overfitting problem. CFS, which was multivariate filter approach, was applied to pick by evaluated the most strength feature into the model. Then, the authors apply 30 algorithm correspond with RF ensemble method to increase the performance by utilized these methods to Parkinson’s , Diabetes and Heart disease datasets.

The Datasets are from UCI machine learning repository which included 768 data sample and 8 features from diabetes dataset , 303 data samples and 13 features from Cleveland Heart diseases and the last dataset had 31 data samples and 21 features from Parkinson’s dataset. Moreover, all three datasets had 2 classes to identify each patient( disease or not having disease). With the CFS algorithm application, the authors obtained 4, 7 and 10 features from diabetes, heart disease and Parkinson’s dataset respectively without any evaluation on the feature’s selected.

Machine Learning methods and RF classifier ensemble were applied to assess performance. The authors select 30 machine learning algorithms including in Table 1. All algorithm was used the default method such Principal Component Analysis(PCA).



*Table 1 : List of Machine Learning algorithms(Ozcift, Gulten , 2011, p. 447 )*

Also, at least 10 number of classifiers were used as another parameter to monitor the effect of classification. During the CFS process, the authors stated that they do not evaluate the feature’s performance but only reduced the size of overall number of feature. The authors applied 30 algorithms to selected features with CFS method through three datasets twice. One did not correspond with RF and another does. They evaluated all performance of both times they run the algorithm by comparing three score metrics which are ACC, KE and AUC.

As a result, the authors claimed that overall of their score from every dataset was better with RF ensemble classifier. They generated ACC score to show the different from both based classifier and RF, and use KE and ROC as a supporter to those score. The author found 2.97% , 2.32% and 2.7% increase in accuracy from Diabetes, Cleveland Heart disease and Parkinson’s datasets respectively. Moreover, KE and ROC score were also increase with ensemble classifier which they used these score to support the overall increase of ACC. Thus, they concluded that RF ensemble classifier provide a better performance.

In my opinion, first, the result was based on a CFS feature selection without any feature evaluation before running the model which I disagree with this approach. This automatically pick feature can cause wrong interpretation and also affect the validation of the model. There was also no evidence to support the features that they selected to support their experiment is applicable.

In addition, the author selected algorithm from a software that automatically pick algorithm which the authors believed it would be efficient for their experiment. There also no clear support to these machine learning algorithm that why this selection was suitable for this type of dataset and some models just showed the score that could cause overfitting. Moreover, the main discussion on this article was focused on the improvement of average accuracy combining from all 30 machine learning algorithms by using ACC score. The authors computed KE and AOC mainly to support the result of ACC by observed the improvement of both score. However, the author did not give clear statement that why KE and AOC performance were efficient enough to use.

In the conclusion section, the authors stated several issues from this article which were based classifier the were being used and classification structure . They said, there might be a problem about the ACC when analyzed too many based classifiers. Because of this , the authors came up with the example of one algorithm which is ‘J48’ to validate the performance of RF. It is quite confused why they pick J48 and how they used it to compare the model performance. There was no explanation on how this J48 algorithm works in any part of this paper that I could agree with the authors that RF had high performance classification method. Second, the authors also stated the issue of the structure of class in the dataset that if it was the multiclass it could have given a better performance. I think it could be better or it could not because the nature of the dataset does not have many class other than having disease or not. It might have the diseases datasets that have multiclass which I believed the prediction percentage would drop and It may not be able to use to predict the result.

For the visualization, I found some of figures and tables showed fault interpretation and wrong information between the figure and context. The authors stated in the context that the figures was very obvious that RF improved the accuracy without plotting all algorithm in the graph and the number of algorithm that plot on the graph was not accurate.

In conclusion, I rejected this article for several reasons. First, the features of the model should be evaluated before using in any algorithm. Without assessing the features, some algorithm that the authors selected may not be useful. Second, the machine learning algorithms that were used in the article did not have any explanation on why it was suitable to use in the type of dataset and this set of features. Finally, figures and tables were inaccurate and lead to a fault information for the reader.

Reference,

A. Ozcift , A. Gulten (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithm.

One of your critiques is about the visualizations. Next time I would recommend adding an example to the paper and going a little more in-depth on the example.

What did you think the author did well in his/her research?  Critiques should address both the good and the bad.