CP610 Data Analysis Spring 2021

Paper Critique

Data Mining Approach to Detect Heart Diseases

Vikas Chaurasia
Research Scholar, Sai Nath University,
chaurasia.vikas@gmail.com
Ranchi, Jharkhand, India

Saurabh Pal
Dept. of MCA, VBS Purvanchal University
drsaurabhpal@yahoo.co.in
Jaunpur, Uttatr Pradesh, India

Name: Dee Wu   Date: July 7, 2021

1. Introduction

This is the Paper to be discussed: Chaurasia, V. and Pal, S. (2013) Data Mining Approach to Detect Heart Disease. International Journal of Advanced Computer Science and Information Technology (IJACSIT), 2, 56-66.   [Citation Time(s):1]

The Heart Disease Data Set is composed of four data sets, processed data files, and a data set description, which are sourced from the collections of heart disease diagnosis by Cleveland Clinic Foundation, Hungarian Institute of Cardiology, University Hospital, Zurich, Switzerland, and V.A. Medical Center, Long Beach. Even the data set has 76 attributes only 11 of them are adopted in the paper and work, and the dataset only has 294 rows.

The algorithms used in the paper are Naïve Bayes, J48 Decision Tree, and Bagging. The data analysis is conducted with Weka tool and the results are compared between bagging and without bagging in 10-fold cross validation. Kappa statistic is used to evaluate the accuracy of the measuring cases, which is also prevalent in the industry to distinguish the reliability of the data being collected and its validation. However even the limitations of this exercise are notified in the paper they are not fully expanded, and the validation of the data set itself is not even mentioned.

2. Summary

Medical data analysis is intriguing, not only because the acquired dataset is decisive in the result, but also it is of an authority in the medical area. The countless data mining code/ algorithms are also problem as they need to be chosen and analyzed, as well as the fierce competition in the market of pursuing a higher accuracy in prediction with an acceptable time complexity so that able to help the health care professionals in their work.

This paper analyzed different classifiers in the diagnosis of heart diseases and evaluated the performance by the confusion matrixes, and the comparisons show the bagging algorithm is the winner of the race as its accuracy is the highest, even the time consumed to build the model is slightly longer comparatively, it is still acceptable.

As mentioned in the paper it is acceptable to produce short but accurate prediction list for the heart patients, and the model also works to identify the patients who need special attentions.

3. Critique

3.1 Strengths

The Heart Disease Date Set is from UCI Machine Learning Repository. The dataset is composed of a group of labels, in which case the classification models are more suitable. The classifiers will be discussed in the following sessions.

Another spotlight is the third-party tools are used fully in the work. Kappa statistic is used to evaluate the accuracy of the measuring cases, and Weka is used to implement the three nominated classifiers, Naïve Bayes, J48 Decision Tree, and Bagging.

3.2 Weaknesses

The paper mentioned, “some limitations on this work are noted as pointers for future research” [1], they are not listed and expanded. The reason to choose three algorithms is not clear as other classification algorithms are not mentioned in comparison, like KNNs, SVM.

Data set is the foundation of data mining especially in medical fields as the validation of the data set is critical and decisive to a good prediction result. The data set is composed of 294 records (hungarian.data) and seems falls into a smaller volume to represent the versatile and profound data mining practices.

3.3 Analysis I

The algorithm Naïve Bayes assumes that the features are independent. With respect to the class labels, practically dependencies exist among the variables, which cannot be modeled in Naïve Bayes but is to be dealt with by other algorithms, such as Bayesian belief Networks.

The probability is written as follows, in which $P(c_j)$ is the ratio of the samples with class label c_j to all the available samples.

$$\mathrm{P}(label = c_j|Y) = P(c_j) * \prod_{i=1}^{n} P(a_i = \bar{a}_\iota|c_j)$$

The experiment is conducted using Weka tool, a collection of machine learning algorithms, which also implements the three classifiers used in this paper. Kappa statistic is used to assess the accuracy of models including mean absolute error, root mean squared error, as well as the relative absolute error, and root relative squared error in the percentage format for future reference and the evaluation.

3.4 Analysis II

The recommended data mining algorithm is Bagging, which means “Bootstrap aggregation an ensemble method to classify the data with good accuracy [1].” As one of the Ensemble Methods it derives the decision trees to build the base classifiers on the bootstrap samples, with replacement from the dataset, and then a combination of all base classifiers derives the final model or decision tree. As expected, the prediction result is significantly better than a single classifier derived from the dataset, and the noise data is not considerably impacted, but overall a more robust model and a prediction.

Some limitations on this work are noticed and regarded as “pointers” for future research, while not listed in the paper, could be applying the advanced methods, such as Bayesian Belief Networks for label dependencies, neural network for weights, SVM, Frequent Patterns, KNN or any other Lazy Learner approaches, and other advanced methods, to expand the scope of the competitors, with evaluations, and achieve an even better accuracy and performance.

The credibility of the data set is not mentioned in the paper, even the owner (David Aha) of the quoted data sets mentioned that one of the data set had been messed up and the original copy of the database appears to be corrupted [2, *WARNING*].

> *The file cleveland.data has been unfortunately messed up when we lost*
> *node cip2 and loaded the file on node ics. The file processed.cleveland.data*
> *seems to be in good shape and is useable (for the 14 attributes situation).*
> *I'll clean up cleveland.data as soon as possible.*
>
> *Bad news: my original copy of the database appears to be corrupted.*
> *I'll have to go back to the donor to get a new copy.*
>
> *David Aha*

Andrew Ng, a forerunner of machine learning and AI, quoted, “AI System = Code + Data, where code means model/ algorithm.” The importance of data is never overlooked, nor emphasized. In machine learning studies, it is mistakenly believed that more data will produce more value, which resulted the big data is actually dumb data, short of good quality, and resulted less accuracy or diminished credibility.

4. Conclusion

The research work achieves the objective, that to predict accurate presence of heart disease with reduced number of attributes, by applying Naïve Bayes, J48 Decision Tree, and Bagging algorithms and the 10-fold cross validation method.

The areas for improvement could be adding more analysis and validation of the data set, as well as the directions for the left unfinished work in the medical data mining area.

5. References

[1] Chaurasia, V. and Pal, S. (2013) Data Mining Approach to Detect Heart Disease. International Journal of Advanced Computer Science and Information Technology (IJACSIT), 2, 56-66. [Citation Time(s):1]

[2] Heart Disease Data Set, UC Irvine Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/Heart+Disease. Ref on July 7, 2021.