Detecting Major Disease in Public Hospital Using Ensemble Techniques

by Mgs Afriyan Firdaus

Submission date: 01-Apr-2023 03:00PM (UTC+0700)
Submission ID: 2052772546
File name: 2014_-_ISTMET.pdf (264.31K)
Word count: 3130
Character count: 17249

Detecting Major Disease in Public Hospital Using Ensemble Techniques

Mgs. Afriyan Firdaus[1], Rin Nadia[2], Bayu Adhi Tama[3]
Department of Information Systems, Faculty of Computer Science, Sriwijaya University
afriyan_firdaus@yahoo.com[1], puppe.meister@gmail.com[2], bayu@unsri.ac.id[3]

Abstract—**Hepatitis is chronic disease that becomes major problem in developing countries. Health experts estimate that more than 185 billion people have chronic hepatitis worldwide. This paper attempts to detect major disease such as hepatitis in public hospital using ensemble methods. Several ensemble techniques were applied to acquire knowledge from patient medical records. Afterwards, rule extraction from decision tree and neural network are summarized in order to assist experts in detecting hepatitis. Accuracy of those algorithms is also performed and from the experimental result shows that Bagging, with decision tree as base-classifier, denotes best performance among other classifiers.**

Keywords - ***ensemble methods, hepatitis, predictive accuracy, public hospital, rule extraction.***

I. Introduction

Hepatitis is a major global health problem. Currently, Indonesia is a third largest country worldwide after China and India which has large number of hepatitis patients. Experts predict about 23.9 million people in Indonesia are infected with the hepatitis type-B virus [1], whereas World Health Organization (WHO) estimates more than 240 million people worldwide have chronic infections of hepatitis. It yields about 600 thousand people die every year in the consequence of critical outcomes of hepatitis type-B virus [2].

Hepatitis has no clear clinical symptoms, so as many hepatitis patients unable to gain appropriate diagnosis and treatment. To date, symptoms of hepatitis disease is also different, depending on the type of hepatitis virus itself. Hence, it is necessary to find proper approaches as well as early detection, prevention and treatment of hepatitis disease.

Corresponds to the context of early detection and prevention, data mining could be applied as an alternative method by discovering knowledge from the medical records data. Data mining deals with the analysis and extraction of knowledge from medical data, intended for supporting diagnostic, screening, prognostic, monitoring, or overall patient management tasks [3].

In public hospital, large amount of medical data are stored, accumulated, and growing as 'data tombs' without undertaking further analysis to the data. Intelligent data analysis as well as data mining was disseminated in order to promote the creation of knowledge to assist clinicians on making decisions as for early detection of the disease. Knowledge discovery as part of data mining process has significant role to discover interesting patterns (knowledge) from huge amounts of data. Interesting patterns is commonly as non-trivial, implicit, previously unknown and potentially useful [4], so as in such way they can be put to use in areas such as decision support and prediction [5], and nowadays, in hepatitis domain, finding knowledge is an active and challenging research.

Several researches have been conducted on detection methods of hepatitis disease. A rule discovery support system for sequential medical data was firstly introduced by Ohsaki et al. They used K-means algorithm and C5.0 (latest version of C4.5) algorithm to discover rules to predict the trend of hepatitis data in the future [6]. Ho et al proposed a temporal abstraction method to find interesting pattern from hepatitis data. Temporal abstraction method was developed since many machine learning techniques could not be applied to temporal domains [7].

The latest study on data mining application in detecting hepatitis, for instance, can be found in [8]. The objective of the study is to construct a simple model to identify hepatitis patients with high-risk of developing hepatocellular carcinoma (HCC). Chronic hepatitis type-C patients were involved and analyzed by decision tree to build a predictive model for HCC development. A comprehensive literature review concerning predictive data mining in clinical medicine also can be found in [9]. They surveyed and clustered researches found in many literatures into four areas depending on the data used such as predictive data mining techniques in clinical applications, data mining with temporal approaches, feature selection in predictive modeling, and predictive medicine and 'omics' sciences.

However, most of those researches merely employed common data mining techniques that implied such other techniques i.e. ensemble methods have not been utilized. Hence, this paper aims to address ensemble techniques so as finding most influential attribute on hepatitis disease. We have collected up to 300 records of hepatitis data (all type of hepatitis are included) in one of local public hospital in Indonesia. Then, we extracted the data and constructed several classifiers. Classification accuracy at different number of features is performed and examined to find the best classifier. Finally, we extract and compare extracted rules from decision tree (J48) and neural network (with REANN algorithm).

II. Literature Review

In this section, ensemble methods and implementation of REANN algorithm with Java-based interface are presented. Rule extraction from J48 is not presented in detail here as it can be found in many literatures for instance in [10], [11], and [12].

978-1-4799-3704-2/14/$31.00 ©2014 IEEE

A. Ensemble Techniques

In data mining researches, ensemble techniques have spent great attention. Incorporating multiple classifiers are becoming prevalent owing to empirical outcomes that proposing them yields more robust and more accurate prediction as they are compared to the individual classifiers [13]. An ensemble contains a number of learners called base-learners. Base-learners are usually generated from training data by a base learning algorithm. Examples of these techniques include Bagging (Bag) [14], AdaBoost (Bo) [15], Random Subspace (RS) [16], Random Forest (RF) [17], and Rotation Forest (RFor) [18].

Among those approaches, AdaBoost become very popular as there were many variants. We utilized AdaBoost.M1 [19] algorithm that is already implemented in well-known open source data mining software, WEKA [12], to boost two popular base-classifiers, those are, decision tree (J48) [10] and neural network (NN) with multilayer perceptron architecture. Despite boosting, bagging, random subspace, and other ensemble methods are designed, and usually used with J48, it is also applicable to perform other base classifiers [20].

A novel ensemble algorithm called Rotation Forest was firstly introduced by Rodriguez et al [18]. This approach generates an ensemble classifier by training a base-learner on a randomly selected subspace of the input data that has been rotated using Principal Component Analysis (PCA) [21]. The latest work of ensemble technique which combined Rotation Forest and AdaBoost called RotBoost was proposed by Zhang and Zhang. It could generate ensemble classifier with significantly lower prediction error than either Rotation Forest or AdaBoost. It was also developed to perform much better than Bagging and Multi-Boost [22].

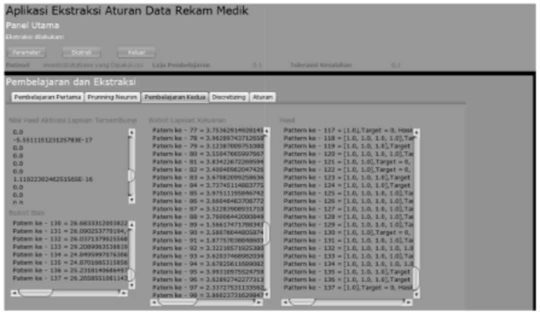
B. Rule Extraction from Neural Network

Several approaches of rule extraction from neural network have been discovered in the last decades. For instance, a method called X2R has been proposed by Liu and Tan [23]. X2R can generate concise rules from data set; however, generated rules are order-sensitive.

A hybrid method that is suitable for data set with binary attributes was proposed by Setiono et al [24]. The method was built upon neural network rule extraction (with M-of-N construct) [25] and consists of two components such as neural network and a decision tree classifier. Though the method can be performed well in general, there are some drawbacks i.e. the neural network training is slow when the data set is large in terms of the numbers of samples and/or attributes.

REANN algorithm was proposed by Kamruzzaman and Islam [26]. The objective of REANN algorithms was to find simple rules with high predictive accuracy. Some key advantages using REANN include; (1) as using constructive pruning strategy, it can determine optimal ANN architecture automatically, and (2) it can extract rules that are concise, comprehensible and highly accurate [26].

A Java-based interface for rule extraction with REANN algorithm was developed (as shown in Figure 1) in order to create simple and easy tool that can be employed by researchers, clinician, or decision makers in detecting hepatitis disease. The interface was equipped with parameter selection and rule extraction button. Training error and epochs could be selected as well to perform neural network training.



Dataset ...ments\Database yang Dipakai.csv Laju Pembelajaran 0.1 Toleransi Kesalahan 0.1

```
0.0
-5.551115123125783E-17
0.0
0.0
0.0
0.0
1.1102230246251565E-16
0.0
0.0
0.0
```

```
Patern ke - 130 = 26.6833312883822
Patern ke - 131 = 26.090253770194,
Patern ke - 132 = 26.0371379025568
Patern ke - 133 = 29.2089063538819
Patern ke - 134 = 24.8495997076306
Patern ke - 135 = 24.0701665315856
Patern ke - 136 = 25.2318140686497
Patern ke - 137 = 26.2658551081143
```

```
Patern ke - 77 = 3.75362914928145
Patern ke - 78 = 3.96289743712658
Patern ke - 79 = 3.12387009751080
Patern ke - 80 = 3.55047665997667
Patern ke - 81 = 3.83422672269594
Patern ke - 82 = 3.40040962047426
Patern ke - 83 = 3.67082099258636
Patern ke - 84 = 3.73745114803775
Patern ke - 85 = 3.97511195846742
Patern ke - 86 = 3.66046483708772
Patern ke - 87 = 3.52283908931710
Patern ke - 88 = 3.76006442080849
Patern ke - 89 = 1.56617471788343
Patern ke - 90 = 3.58076044805874
Patern ke - 91 = 1.87757038848603
Patern ke - 92 = 3.32210571925380
Patern ke - 93 = 3.62037468982034
Patern ke - 94 = 3.67825611689082
Patern ke - 95 = 3.99310975524759
Patern ke - 96 = 3.62892742277313
Patern ke - 97 = 2.33727531133562
Patern ke - 98 = 3.86023731629847
```

```
Pattern ke - 117 = [1.0],Target = 0, Hasil
Pattern ke - 118 = [1.0, 1.0, 1.0, 1.0],Tar
Pattern ke - 119 = [1.0, 1.0, 1.0],Target
Pattern ke - 120 = [1.0, 1.0, 1.0, 1.0],Ta
Pattern ke - 121 = [1.0, 1.0],Target = 0,
Pattern ke - 122 = [1.0, 1.0],Target = 0,
Pattern ke - 123 = [1.0, 1.0, 1.0],Target
Pattern ke - 124 = [1.0, 1.0, 1.0],Target
Pattern ke - 125 = [1.0, 1.0, 1.0, 1.0],Ta
Pattern ke - 126 = [1.0, 1.0, 1.0, 1.0],Ta
Pattern ke - 127 = [1.0, 1.0, 1.0, 1.0],Ta
Pattern ke - 128 = [1.0, 1.0, 1.0, 1.0],Ta
Pattern ke - 129 = [1.0, 1.0, 1.0, 1.0],Ta
Pattern ke - 130 = [1.0, 1.0],Target = 0,
Pattern ke - 131 = [1.0, 1.0, 1.0, 1.0],Ta
Pattern ke - 132 = [1.0, 1.0, 1.0, 1.0, 1.0
Pattern ke - 133 = [1.0, 1.0, 1.0, 1.0],Ta
Pattern ke - 134 = [1.0, 1.0, 1.0, 1.0, 1.0
Pattern ke - 135 = [1.0, 1.0, 1.0],Target
Pattern ke - 136 = [1.0, 1.0, 1.0],Target
Pattern ke - 137 = [1.0],Target = 0, Hasil
```

Figure 1. An Interface for Detecting Hepatitis Disease with REANN Algorithm.

III. Experimental Design

A. Data Collection and Pre-processing

Raw data was obtained from patient medical records in one of the public hospital in Indonesia from 2009 to 2013. We included all type of hepatitis disease. After performing data pre-processing task such as cleaning, transformation and integration, the data set comprises 206 records and 12 features, where 51.46% (106) cases in hepatitis type-A (Class A), 27.18% (56) cases in hepatitis type-B (Class B), and 21.36% (44) cases in hepatitis type-C (Class C). The top-12 significant features used in this study in descending order by information gain are: (1) Myalgia, (2) Hives, (3) Age, (4) Gender, (5) Abdominal pain, (6) Cough, (7) Heartburn, (8) Fever, (9) Headache, (10) Jaundice, (11) Fatigue, and (12) Appetite.

B. Experimental Setup

We conducted an empirical study using our real dataset described previously. Firstly, we applied all ensemble methods with decision tree and neural network as base learners. Hereafter, we compared and examined their accuracy through different feature number.

We used 10-folds cross validation to evaluate all classifiers. Standard cross-validation is 10-cross validation as extensive experiments have shown that this is the optimal number to get an accurate estimate.

Several interesting patterns (rules) that were successfully extracted from J48 and REANN are also presented and discussed in the next section. Rules are beneficial for early detection of hepatitis in the public hospital.

IV. Result and Discussion

A. Classification Analysis

We carried out and compared algorithms as single-classifiers (J48, NN, and RF) and all ensemble techniques (Bag, Bo, RS, RF, and RFor) for both base-classifiers (NN and decision tree) by assessing them with different feature number. A total of 77 experiments were conducted for classification analysis. Table 1 shows the classification accuracy as the number of input variables increase, we can find out that the higher classification accuracy was acquired.

In term of average performance as shown in Figure 2, bagging J48 performed best accuracy among other methods. When applying J48 in several ensemble techniques as well as bagging, boosting, and random subspace; their accuracy were slightly worse compared to individual classifier, except for bagging and random forest. In other side, when applying neural network in several ensemble methods such as bagging and boosting, their accuracy were worse compared to neural network as individual classifier.

TABLE I. ACCURACY COMPARISONS OF EACH MINING TECHNIQUE (%)

Methods	Feature Number						
	3	4	6	7	8	10	12
J48	44.2	60.2	75.2	68.9	62.6	70.4	70.9
NN	44.2	64.1	71.8	62.1	61.7	64.1	68.0
RF	39.3	60.2	71.8	59.7	63.1	62.1	67.0
Bag-J48	44.2	61.7	75.2	66.5	66.5	67.0	72.3
Bo-J48	44.2	58.7	72.3	65.5	63.1	62.1	69.4
RS-J48	44.2	64.1	75.2	64.6	57.3	66.0	68.9
RFor-J48	42.7	60.2	75.2	67.0	63.6	67.0	75.2
Bag- NN	40.3	59.2	72.3	61.7	62.1	65.5	69.0
Bo-NN	44.2	62.6	72.3	58.7	65.0	62.1	64.6
RS-NN	44.2	64.1	75.2	64.6	57.8	64.6	71.4
RFor-NN	41.8	63.1	72.8	64.1	66.0	65.0	68.9

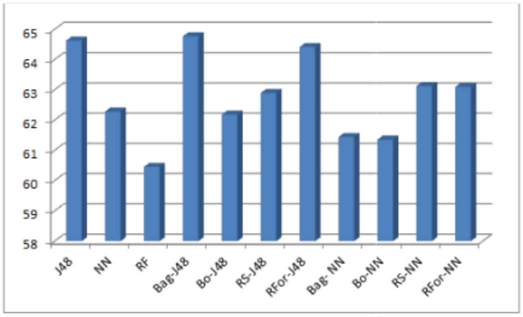


Figure 2. Average Performance of Classifiers

Furthermore, to our experimental study carried out for real data set, ensemble classifiers generally outperform single classifiers as the conclusion of empirical study of bagging [14], which stated that bagging improve performance of single classifiers by reducing bias and variance. Meanwhile, our results for random forest were poor in comparison to other methods opposes to previous study [17], which stated that random forest give results competitive with bagging and boosting.

B. Rule Extraction

Decision tree (J48) uses information gain measure to choose among the candidate attributes at each step while producing the tree [11]. From our experimental result, J48 generated a total of 9 rules. Three rules classify samples as hepatitis type-A, the 2 rules classify samples as hepatitis type-B, and the 4 rules classify samples as hepatitis type-C. The accuracy rate of these rules on the training sets is 78.64%, whilst accuracy rate of these rules on validation sets is 71.8%. The most significant rules for each type of hepatitis are presented as follows:

R1 : IF hives = false and myalgia = false THEN
Class A
R2 : IF hives = false AND abdominal pain = true AND
myalgia = true AND fatigue = true AND
appetite = true THEN Class C
R3 : IF hives = true AND myalgia = false THEN
Class B

Otherwise, total of 2 extracted rules were obtained from REANN. The accuracy rates of these rules on the training and validation sets are 79.15 and 70.60%, respectively. The rules are presented as follows:

R1 : IF jaundice = true AND myalgia = false THEN
Class A
R2 : IF abdominal pain = true AND
hives = true AND appetite = true THEN Class B

From the rule extraction experiment, it is shown that training set accuracy rates from the two methods are quite similar, the accuracy rates of the rules extracted from REANN on the validation set are higher by up to 0.51% than the accuracy of the rules from J48. However, of the total 9 significant rules that had been obtained from J48, we think that those three rules are the most significant because it can directly address the question as for early detection method of hepatitis in public hospital.

Furthermore, two rules extracted from REANN have close relation to rules obtained from J48. Two rules are only for hepatitis type-A and type-B, while hepatitis type-C could not be successfully extracted by REANN. All significant rules have been confirmed by clinicians and they agreed to accept the rules as second opinion concerning detection hepatitis.

V. CONCLUSION

We employed and examined several ensemble techniques in detecting hepatitis in public hospital. Several significant rules from decision tree and neural network have been successfully obtained. From the experimental result, bagging methods with decision tree as base classifier yielded best accuracy among other methods. This research might have some limitations. There is direction for further research that could be taken. It would be interesting to perform cross-sectional research by comparing the characteristics of patient with other type of hospitals. Finally, it would be useful if more medical data records could be acquired later.

References

[1] J. Michael Hall, "Indonesia's Hepatitis B Crisis," Hepatitis B Foundation, PA, Newsletter 2013.

[2] World Health Organization (WHO). (2013, January) Media Center. [Online]. http://www.who.int/mediacentre/factsheets/fs164/en/index.html

[3] N. Lavrac and E., Zupan, B. Keravnou, "Intelligent Data Analysis in Medicine," in *Encyclopedia of Computer Science and Technology*. New York: Dekker, 2000.

[4] J. William Frawley, G. Piatetsky-Shapiro, and Christopher J. Matheus, "Knowledge Discovery in Databases: An Overview," *AI Magazine*, vol. 13, no. 3, pp. 57-70, 1992.

[5] Yue Yue Huang, Paul McCullagh, Norman Black, and Roy Harper, "Feature Selection and Classification Model Construction on Type 2 Diabetic Patient's Data," in *Lecture Notes in Artificial Intelligence*, Petra Perne, Ed.: Springer-Verlag, 2005, ch. 3275, pp. 153-162.

[6] Miho Ohsaki, Yoshinori Sato, Hideto Yokoi, and Takahira Yamaguchi, "A Rule Discovery Support System for Sequential Medical Data- In the Case Study of a Chronic Hepatitis Dataset," in *IEEE Int'l Conf. on Data Mining (ICDM'02)*, 2002, pp. 97--102.

[7] Tu Bao Ho et al., "Mining Hepatitis Data with Temporal Abstraction," in *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 2003, pp. 369-377.

[8] Masayuki Kurosaki et al., "Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis C," *Journal of Hepatology*, vol. 56, no. 3, pp. 602–608, March 2012.

[9] Riccardo Bellazzi, Fulvia Ferrazzi, and Lucia Sacchi, "Predictive data mining in clinical medicine: a focus on selected methods and applications," *WIREs Data Mining and Knowledge Discovery*, vol. 1, pp. 416-430, Spetember 2011.

[10] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kauffman Publishers, 1993.

[11] Tom M. Mitchell, *Machine Learning*. New York: McGraw-Hill Science, 1997.

[12] Ian H. Witten, Eibe Frank, and Mark A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition*. San Francisco: Morgan Kaufmann, 2011.

[13] D. Opitz and R. Maclin, "Popular Ensemble Methods: An Empirical Study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[14] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[15] Y. Freund and R.E. Schapire, "A Desicion-Theoretic Generalization of On-line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[16] T.K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions onPattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832 - 844, 1998.

[17] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[18] J.J. Rodriguez, L.I. Kuncheva, and C.J. Alonso, "Rotation Forest: A New Classifier Ensemble Method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619 - 1630, 2006.

[19] Y. Freund and R.E. Schapire, "Experiments with a New Boosting Algorithm," in *Thirteenth International Conference on Machine Learning*, San Francisco, 1996, pp. 148–156.

[20] Marina Skurichina and Robert P. W. Duin, "Bagging, Boosting and the Random Subspace Method for Linear Classifiers," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 121-135, 2002.

[21] Mark Hall et al., "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10-18, 2009.

[22] Chun-Xia Zhang and Jiang-She Zhang, "RotBoost: A Technique for Combining Rotation Forest and AdaBoost," *Pattern Recognition Letters*, vol. 29, pp. 1524–1536, 2008.

[23] Huan Liu and Sun Teck Tan, "X2R: A Fast Rule Generator," in *IEEE International Conference on Systems, Man and Cybernetics*, vol. II, Vancouver, 1995, pp. 1631 - 1635.

[24] R. Setiono, S. -L. Pan, M.-H. Hsieh, and A. Azcarraga, "Automatic Knowledge Extraction from Survey Data: Learning M-of-N Constructs Using a Hybrid Approach," *Journal of the Operational Research Society*, vol. 56, no. 1, pp. 3-14, January 2005.

[25] Rudy Setiono, "Extracting M-of-N Rules from Trained Neural Networks ," *IEEE Transaction on Neural Networks*, vol. 11, no. 2, pp. 512-519, March 2000.

[26] S. M. Kamruzzaman and Monirul Md. Islam, "An Algorithm to Extract Rules from Artificial Neural Networks for Medical Diagnosis Problems," *International Journal of Information Technology*, vol. 12, no. 8, pp. 41-59, 2006.

Detecting Major Disease in Public Hospital Using Ensemble Techniques

ORIGINALITY REPORT

97%	22%	96%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Mgs. Afriyan Firdaus, Rin Nadia, Bayu Adhi Tama. "Detecting major disease in public hospital using ensemble techniques", 2014 International Symposium on Technology Management and Emerging Technologies, 2014 Publication	96%
2	Submitted to Telkom University Student Paper	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		