Research Article

Effective Diagnosis of Coronary Artery Disease using Case-based Reasoning

Yong-Gyu Jung[1], Bumsu Kim[2], Hojin Nam[3], Minseo Rhee[4], Jeung-Sun Lee[*5]
[1]Dept. of Medical IT, Eulji University, Korea, ygjung@eulji.ac.kr,
[2]Div. of Customer &Media, Korea Telecomm, ben_kim@kt.com
[3]CEO, Purium Co., ltd, ceo@purium.kr
[4]VI FORM, Berkshire School/ Purium Co., ltd, michelle6224@purium.kr
[*5]Dept. of Mortuary Science, Eulji University, Korea, jslee@eulji.ac.kr,

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: With the advent of big data, data mining is more increasingly utilized in various decision-making fields by extracting hidden and meaningful information from large amounts of data. Even as exponential increase of the request of unrevealing the hidden meaning behind data, it becomes more and more important to decide to select which data mining algorithm and how to use it. There are several mainly used data mining algorithms in biology and clinics highlighted; Logistic regression, Neural networks, Support vector machine, and variety of statistical techniques. Among them Case-based reasoning (CBR) is relatively seems to be simplistic but very powerful to disclose unseeable problems in complex environments with only simplistic use of the above single technique for prediction of nonlinear models. On the other hand, quantities of the human momentum and activities are more diminished, whereas lifestyle of drinking, smoking and western eating habits are changing, and thus such as the unrevealed risks caused by heart attack or angina are growing up more and more. Therefore according to the increase of patients suffering from heart disease, a number of data mining studies are undergoing to assist medical doctors by prediction of whether to perform coronary angiography which requiring much resources in cost and procedures. Our study uses the same datasets on heart disease patients, that made use of multiple datasets collected from Cleveland, Hungary, Long Beach and Switzerland. Unlike the approach of , we observed that the experimental dataset is composed of multiple populations. And they are similar in use of same kinds of disease patients but different in the time and area of investigation. Through the experimental results, CBR made better performance than the techniques proposed from the original study for the disease prediction. Consequently we conclude effective diagnosis prediction must accompany with selection of the data mining technique considering the characteristics of samples and data collection.

Keywords: CAD, Coronary artery disease, CBR, Logistic, Bayesian network, Heart disease, k-NN, Discriminant function, Cleveland, Hungary, Switzerland, Long Beach

1. Introduction

As human momentum and activities are diminished and changed the lifestyle of drinking, smoking and western eating habits in recent days, unrevealed risks caused by heart attack or angina are growing. It is known that three main risk factors causing the pathogenesis of CAD are high blood pressure, heavy smoking and high cholesterol. CAD is classified as angina pectoris and myocardial infarction referred both collectively Ischemic heart disease [2]. Main symptoms of those diseases are chest pain, discomfort in the chest, and easy feeling of tiredness. In order to diagnose severity and status of heart disease, electrocardiogram (ECG), angiocardiography (ACG) and coronary angiography contrast agent are checked with administered degree of narrowing of the coronary arteries. 3D Coronary artery CT and nuclear medicine tests are conducted to assess the function of the heart muscle, and high specificity test are also used for diagnosis. But it is almost impossible to manage data generated exponential growth quantity and quality of information. Thus it became difficult simply to use statistical techniques or queries alone for huge data search and only with a considerable technical effort to find useful information. Data mining is an important process required in decision-making fields that moves vast amounts of data currently stored in large database system into uncovering hidden data meaning. There are several mainly used data mining algorithms in biology and clinics highlighted; logistic regression, neural networks, support vector machine and other variety of algorithms derived from statistics. Among them, Case-based reasoning (CBR) is relatively seem to be simplistic but very powerful to solve unseeable problems in complex environments only with simplistic use of the above single algorithm for prediction of nonlinear models.

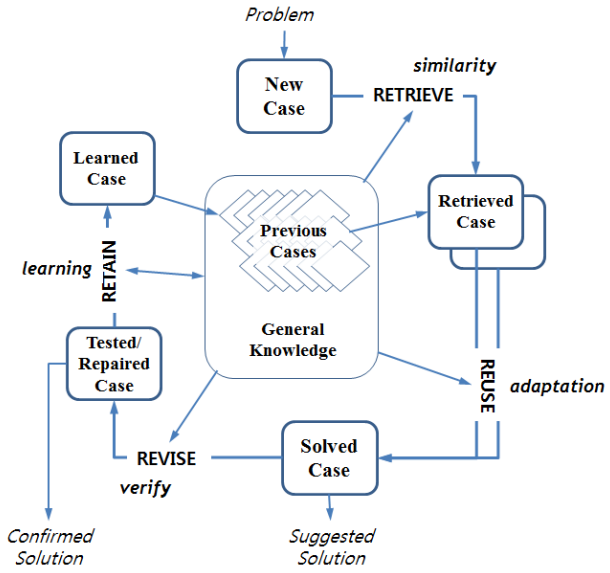
In this paper, we try to find alternative computer based classification models which remedy its shortcomings and provide more high performance of prediction by using the characteristic of data about CAD and analyse by comparison with the models of original study. Our paper is consisted with 6 sections; in the second section we explain about our motivated algorithms to this study; Material and Methods treat the datasets we used and our strategy for experiments. In this section, we introduce about a modified approach from the filtering method used in the original study for clinician's practical needs. In section 4 and 5, we describe the experiments conducted for comparative advantage analysis and discuss about the performance evaluation by comparison of the results. In the last section, we conclude with descriptions about why our study has significance and how can we support the clinician's decision for diagnosing the CAD patients.

2. Literature Review

2.1 Case-Based Reasoning (CBR)

There has been argued that case-based reasoning is not only a powerful method for computerized inference but also a pervasive behavior for human problem solving in every day. Even more radically, all reasoning are based on past cases personally experienced [3]. In CBR, training examples are stored and accessed to be used to solve a new problem [4]. To make a prediction for a new example, those past cases that are similar, or close to the new example are used to predict the target value. Generally CBR is used for classification and regression. In addition, it can be applied when the cases are entangled with complicated cases and where the cases are previous solutions to new complex problems. CBR has advantages as the following; First of all it is intuitive. Secondly, it is not the knowledge elicited to create rules or methods. Hence this makes development easier. Thirdly, it only learns by acquiring new cases by use. Then this makes maintenance easy and the precedent case or rule is accepted as a method for justifying the decision. On the contrary, it has been argued that CBR has some issues in that how many cases are needed, how to remove overlapping cases, how to search efficiently, what features to use for indexing and how to weight the features. Through the above problems, there are several disadvantages in CBR [5]; Can we take a large space for all the cases? Can we take large processing time to find similar cases in case-base? Are there some needs such as case-base, case selection algorithm and case-adaptation algorithm. CBR can be evaluated and justified by confidence. In this time, confidence level is based on number of cases matched, similarity of matched cases to new problem or similarity of matched cases to each other. Following Figure 1 shows Aamodt& Plaza's (1994) classic model of the problem solving cycle in CBR [6].The individual tasks in the CBR cycle (i.e., retrieve, reuse, revise, and retain) have come to be known as the "4 Res" [7].

Figure 1. CBR cycle[6]



2.2 Logistic Analysis

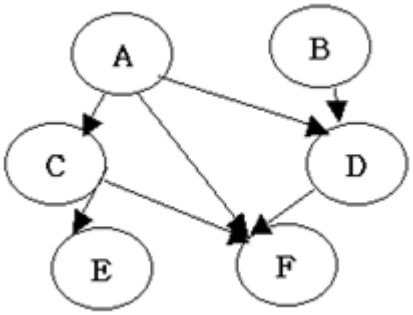
In general, whereas simple logistic regression analysis is to generate a model for the purpose with only one independent variable to predict the dependent variable, multiple linear regression analysis makes a regression model to predict the dependent variable with the multiple independent variables [8]. Multiple linear regression analysis let us know how much a certain descriptive variable effect on the dependent variable, thus by using these descriptive variables to make accurate prediction relatively for the dependent variable. Logistic analysis is, as one of the multiple linear regression analysis, when the samples to be classified are divided into two or more than two populations used to predict individual observed values where to be classified. Unlike discriminant analysis, logistic analysis has strengths that it is possible to use categorical variables for the explanatory variables and also apply in case that the dependent variable is not displayed as a quantitative measure but a qualitative one.

2.3 Bayesian network Analysis

Bayesian network has a graph structure with connected nodes by representing observed target objects as nodes and describing ordered or relational meaning as links [9]. Bayesian network, also called Bayesian Belief network, is DAG (Directed Acyclic Graph) with conditional probabilities for each node. In Bayesian network, each node represents random variables in a problem domain and each arc conditional dependence relationship among these variables [10,19]. Each node contains a conditional probability table that contains probabilities of the node being a specific value given from the values of its parents. The direction of links specifies the conditional dependency relations of nodes or variables, and non-descendants in the graph, which have no links connecting each other

conditionally independent. Therefore Bayesian network is known that it is suitable to reflect the relevance or the relationships of causes and effects between nodes as well as easy for knowledge representation and possibility of inference. Each node is described by a table containing local conditional probabilities of that specific variable in association with other attributes as parameters. As a result Bayesian network provides the advantage of possibility to be calculated more easily than conventional statistics by beginning from the assumption that all things interested statistically such as population parameters, missing values and predicted values are uncertainty but the amount of information is described as the probability.

Figure 2. Bayesian network [10]



3. Material and Method

3.1 Experimental Data

The data used were downloaded from UCI repository of the heart disease. The data wereobtainedfrom303 clinical and non-invasive test results of 4citiesin three countries separated by Cleveland, Hungary, Switzerland, and Long Beach. The reference group used to derive the model consisted of 303 consecutive patients referred for coronary angiography at the Cleveland Clinic of Ohio. The patients had heart tests except myocardial infarction and valvular heart disease and consists of 200 samples at Veterans Administration Medical Center in Long Beach. Hungary dataset comprises 294 samples, which is subtracted the samples having myocardial infarction or valvular heart disease from the patients who took examination of myocardial infarction at the Heart Institute of Cardiology. And Switzerland dataset gives 123 samples from the collected patients who had heart disease test at the university hospitals in Zurich and Basel. All data set was composed of the attributes 76but only 14 attributes were used for our experiments according to the published data set [11].

Table 1. Data Definition and Types

Attributes	Type
Age	numeric
Sex	1:male, 0:female
Chest pain	1 : typical angina 2 : atypical angina 3 : non-anginal pain 4 : asymptomatic
Resting blood pressure	numeric
Serum cholesterol	numeric
Fasting glucose	1:true, 2:false
Resting electrocardiogram results	0:normal 1:having ST-T 2:ventricle problem
Maximum heart rate	numeric
Reduced exercise angina	1:yes, 2:no
Reduced by relaxation exercises associated with ST-segment	numeric
The slope of the peak exercise ST	1:up sloping 2:flat 3:down sloping
The number of major blood vessels	numeric
Thallium defects	3:normal 6:fixed defect 7:reversible defect
The result of heart disease diagnosis	0:absence, 1:presence

3.2 Method

Result In original study, experimental policy was planned and preceded by considering the fact that the most relevant probability thresholds for making decisions concerning angiography or therapy laid between 0.2 and 0.8 for subjects with the chest pain syndrome [1]. It means that subjects whose clinical and test data are concordant will generally have very high or very low probability estimates in other test groups from any algorithm. Based on this like idea, we focused onto selecting the attributes of data set for leaving the problems of dividing which ranges of the subjects into concordant or discordant corresponding to each heterogeneous test group to choose an effective prediction algorithm. Chest pain syndrome is an essential factor in diagnosis of CAD according to domain knowledge. It is separated into 4 types which are typical angina, atypical angina, non-anginal pain and asymptomatic shown in Table 1. Patients with a history of typical angina but negative exercise electrocardiography represent a subgroup with an intermediate likelihood of having coronary artery disease and future cardiac events [12]. Thus such patients can be said to be relatively in stable status. But patients with the term “atypical chest pain” are led to physicians for investigating coronary angiography [13]. In addition, there are many presumptive signs of non-anginal chest pain such as localization with one finger, radiation to the nuchal area, an inframammary primary site, a pain that reaches maximum at the onset, or relief within a few seconds of swallowing food [13]. Therefore these latter two kinds of patients with atypical or non-anginal chest pain are in somewhat dangerous status having possibilities to be easily progressive to coronary disease. As for asymptomatic, there is a report that one problem in defining prognosis in totally asymptomatic patients is the relatively small number of such patients who undergo coronary arteriography and there is a death of such statistics [14]. It means that asymptomatic patients have lower possibility to progress to coronary disease but coronary disease patients with asymptomatic may be dead at high rates. So we recognized that the odds ratio of chest pain syndrome was included in the highest group among all the attributes of Cleveland data set through our preliminary study. And we divided the values of chest pain into 2 types, which are explicit and inexplicit. Explicit types are defined as the patients with typical angina and asymptomatic. In explicit types include the patients with atypical angina and non-anginal pain. In this time, we assumed the explicit types as a self-evident group to be normal or patient and guessed that it might not be his interests who wanted to diagnose and predict patients for therapy. This is due to the fact that the approach that the results are apparently to be predicted may be uniformly applied to all the data sets, so it will not be appropriate for the goal of data mining study as like the above method of original study. Since the angiography test is expected to be obviously useless when the value of chest pain is the explicit types, samples with the inexplicit types only need to be performed. For investigation of it, we tried to find how close the relationship between the chest pain and the classis.

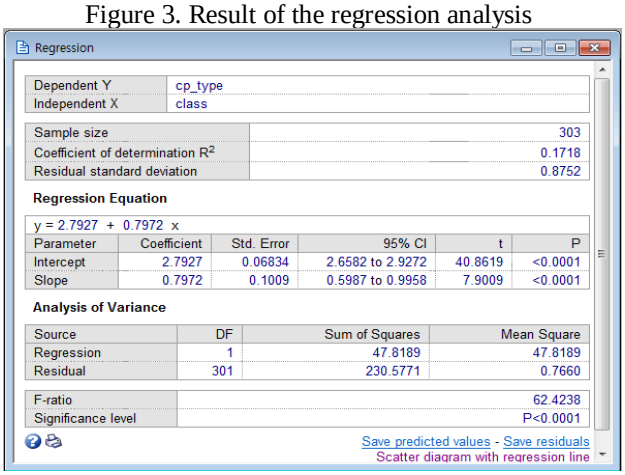
Generally regression analysis helps us understand how the typical value of the dependent variable changes when any one of the independent variables varies, while the other independent variables are held fixed [15].In the result of regression test, we could find that the chest pain syndrome had a major effect on decision of the class value which showing the prediction result across all the data sets (Figure 3).Also according to [16], McNemar test is a test on a 2x2 classification table when the two classification factors are dependent, or when we want to test the difference between paired proportions. Thus we designed a 2x2 matrix by using the explicit and the explicit types to each case with which the class value is 1 or 0 for the test of paired proportions. After the test, the expected result by us has not the same direction with the original study, so we designed several plans for experimental analysis with the chest pain or not like Table 2.

Whereas considering that Logistic and Bayesian network techniques were used to establish patterned rules and predict heart disease patients with good performance in the original study, we selected to use k-NN as our classification method which was expected to make the better performance by recognizing the characteristic of integrated datasets than the prior techniques. This is due to the assumption that because our data sets were not collected from one place but obtained in several countries and hospitals, though they are common in that all the constituents are the patients data suffering heart disease with differences in region and time difficult to find the uniformed rules.

4. Experiments

Result of obtaining probabilistic estimates for the impact of chest pain syndrome X to the class variable Y using the Cleveland data set is presented as Figure 3.

Figure 3. Result of the regression analysis



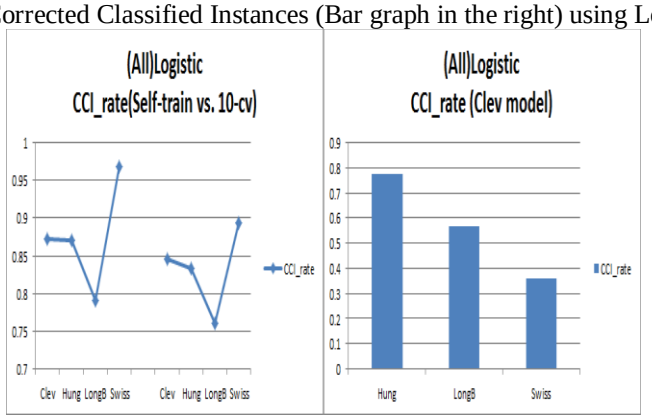
For experiments various classification algorithms are used in WEKA. Especially CBR is implemented as named IB*k*, which is described in Section 2 in WEKA [17]. Each experiment was repeated 10 times of 10-foldscross validation to provide a mean value of experimental results in principles. But there were some cases that result values are unchanged despite of repeated experimentations. For performance evaluation, several measures of positive classified instances and precision / recall and AUC (area of ROC curve) were tested [18].

Table 2. Experimental methods and data algorithms

#Attribute	Algorithm	Dataset			
		Cleveland	Hungary	LongBeach	Switzerland
(14)	Logistic	(1)	(2)	(3)	(4)
(13)	Logistic		(5)	(6)	(7)
	Bayes Network.		(8)	(9)	(10)
	IB*k*		(11)	(12)	(13)

Table 2 shows our trials of experimentation conducted for evaluation of the difference between algorithms to 4 data sets. The cases from (1) to (4) mean testing steps to find self-predicted result from each dataset through the process of self-training and test with 14 attributes. On the other hand, the cases from (5) to (13), which were reduced to 13 features, describe steps that predict each Hungarian, Long Beach and Switzerland data respectively by using Cleveland data as a training one.

Figure 4. Comparison analysis based on Self-train vs. 10-cv(Broken line graph in the left) vs. Cleveland modelin the rate of Corrected Classified Instances (Bar graph in the right) using Logistic algorithm



It shows the logistic prediction result of (1) - (4) of the experiments (refer Table 2) describing the CCI rate in left side of Figure 4. Among two graphs, left one is showing the result run by self-training and prediction, whereas the right one is showing that of using 10 cross-validations. Apparently the self-predicted result is shown to be higher than that of 10 cross-validations. Even the case of Swiss dataset, it can be found in both models that the results appear to be much higher than others relatively. This means Swiss dataset may be overestimated because of the reason for imbalanced data samples. The right side shows the result with the Cleveland model by using each data set as a test data in logistic prediction with 10-folds cross validations.

Figure 5. Comparison analysis of CCI and ICI on the Cleveland model

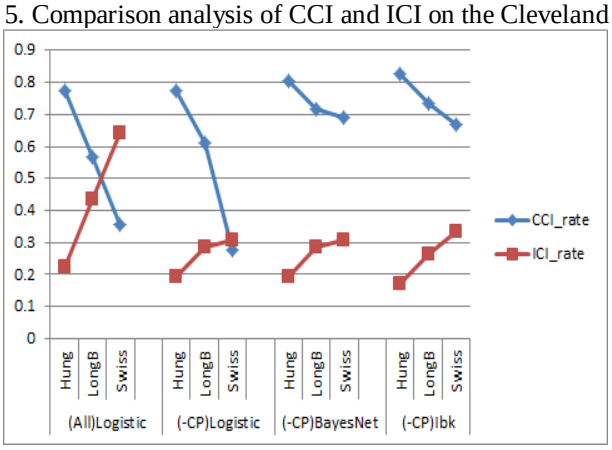


Figure 5~8represent the prediction results to three separated datasets run by the Cleveland model like the way of showing the right side of Figure 4. In Figure 5, CCI rates having13 attributes, which mean the case where the chest pain is deleted, are entirely higher than that with 14, and ICI with 13 attributes are far lower than that with 14 except Hungarian dataset in both cases. None the less, it shows that there isn't any identified difference corresponding to each algorithm.

Figure 6. Comparison of TP and FP rates based on Cleveland Model

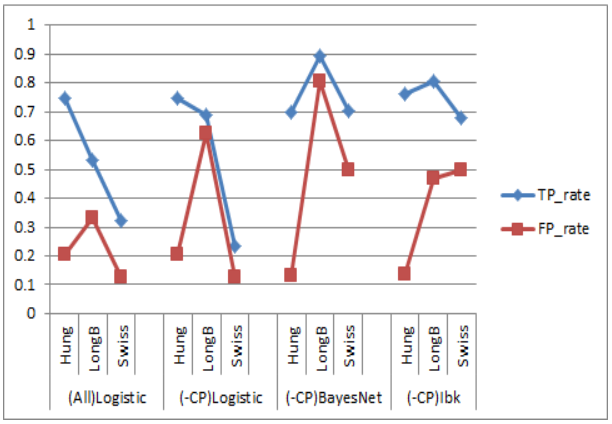


Figure 6 shows TP rates of Bayesian Network and IB*k* (*k-NN*) are significantly higher to appear than the other two cases. And we can find that FP rate of logistic with all 14 attributes is shown to be the lowest on the whole.

Figure 7. Comparison of Precision and Recall across algorithms

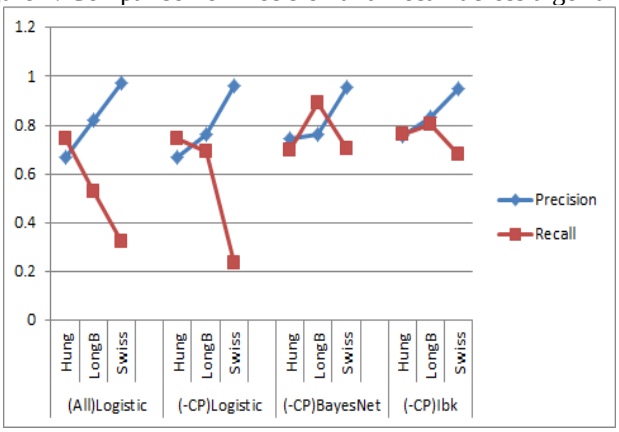
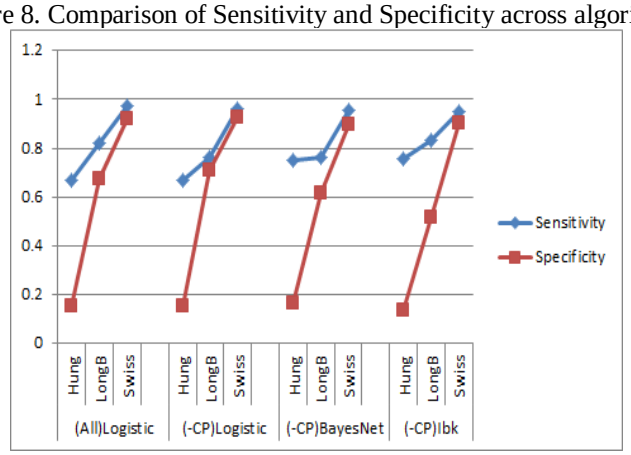


Figure 7 shows that the precision of IB*k* is little higher than other three cases and the recall is also higher together with Bayesian network. There are two interesting things that Hungry data set has not big variations corresponding to each algorithm in recall, but also Long Beach is appeared to be the highest on using Bayesian network. Whereas, there are little differences in sensitivity and specificity to all algorithms (see Figure 8).

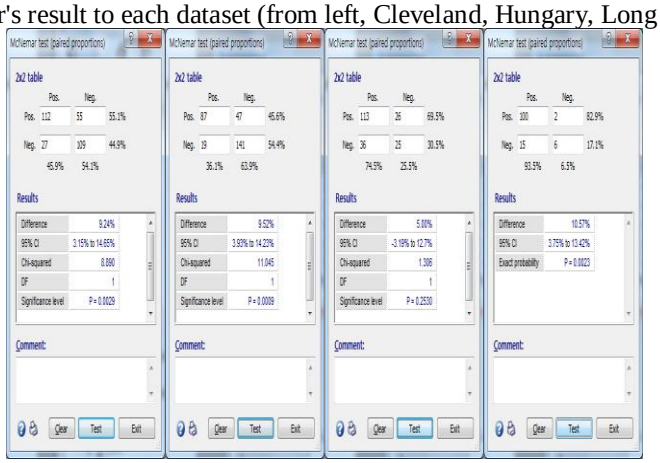
Figure 8. Comparison of Sensitivity and Specificity across algorithms



5. Discussion

Through McNemar's test to all the data sets respectively, there were significant differences between the two proportions in Cleveland, Hungary and Swiss data set, but insignificant in Long Beach data set (Figure 9). Because of discordant results of our McNemar test, we have found there is discordance in the result between the original study and ours.

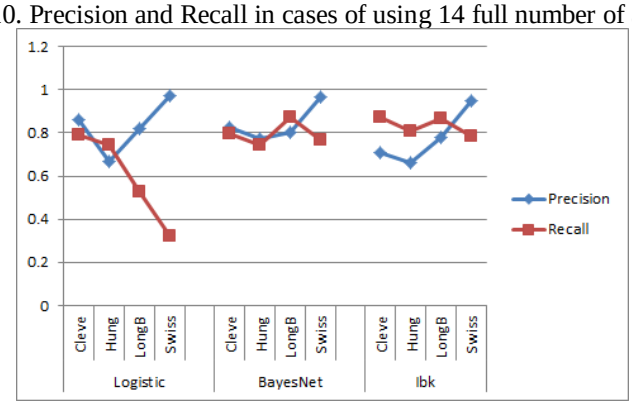
Figure 9. McNemar's result to each dataset (from left, Cleveland, Hungary, Long Beach and Swiss)



In Figure 4, we can see that in all datasets the experimental performance of self-trained samples is more excellent than that of cross-validated samples in comparison of CCI rates. Especially in prediction test using self-trained samples, except for Long Beach, as considering that each data set showing relatively high accuracy is lowered to less than 85% in 10-cv test except for Swiss data set, we can guess the distribution of samples in each data set is uneven. Even in the case of Long Beach, it might be more biased than others. When we consider that the right side is a predicted accuracy based on the Cleveland model, it can be seen that the difference of number of samples in the patient and the normal group is spreading increasingly. This means, despite of every data set commonly consisting of the patients' information suffering from the heart disease around the same time, it is somewhat unsound results are drawn.

According to [1], the Cleveland discriminant function developed based on logistic algorithm had excellent performance rather than Bayesian network of the CADENZA in the rate of overestimation and CCI, but Bayesian network and IB*k* being proposed in this study made better performance in CCI (see Figure 5). Moreover in the aspect of marking the highest value in TP and the second lowest in FP, IB*k* seems to take precedence over logistic and Bayesian network (Figure 6). This like superiority of the CBR is observed in comparison between Figure 7 and 10, thus we can reach the conclusion that CBR is appeared to be the most superior than others in precision without significant degrees of reduction in recall than the results using other techniques when we run a prediction test except for the attribute of chest pain having a great impact on the dependent variable. Also that of CBR is observed in comparison of the sensitivity and specificity (Figure 8).

Figure 10. Precision and Recall in cases of using 14 full number of attributes



6. Conclusion

Data mining is widely used to obtain the useful information from each sector in company's customer management, the bank's personal and business credit score calculation, risk management, health care for the treatment of patients in clinical trials and DNA sequencing analysis in biotechnology. Also it is used in decision-making using variety of techniques for the diagnosis of patient's disease. It this paper we described our approaches and the results tried for getting much higher performance of prediction as consideration of the characteristic of integrated dataset collected in different hospitals, even across several different countries. Despite of being collected in different data sources, all the patients are commonly suffering from heart disease around the same time for identifying whether to run he final angiogram test on the basis of the result of previous study.

As a result, we could find that IBk was, as one of the CBR techniques, the most applicable method to our data set which was composed of several heart disease data sets collected in different hospitals, even across several different countries in several types of experiment for the discovery of effective data mining algorithm considering the characteristics of samples and data collection. And by comparison the results in precision and recall, we could utilize the effect of exclusion of the independent variable if we tried to find unknown effects of that with a great impact on the dependent variable.

Although we have obtained two significant results like the above, our study has disadvantages in several issues. Firstly we only used the algorithms with default options provided from WEKA. Normally the better we control the options of data mining algorithm, the better we can get the results. Secondly, because we mainly depended on the original study with not considering other studies about heart diseases, we did not include other impact factors having influence on causing the heart diseases. Thirdly, our study has not any novel idea in the fields of healthcare informatics but we convince that it will be helpful for medical doctorswho want to use or select effective features in decision of whom belongs to the normal or the diseased. Because it is increasing the varieties of newly occurred unknown diseases, the occurrence of heart diseases and the needs from medical experts for satisfying the customer’s

demand as more and more the societies are being complicated and the technologies are being developed.

The significance of this study resides in reproducibility of the previous study faithfully with the same datasets and clarifying that our proposed CBR algorithm has the higher applicability by considering the characteristic of integrated dataset utilized in the same way.

References

1. Dentrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., and Froelicher, V. (1989). International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease. *American Journal of Cardiology*, Vol. 64, 304-310.
2. Jespersen L., Hvelplund A. andAbildstrom S.Z. (2012). Stable angina pectoris with no obstructive coronary artery disease is associated with increased risks of major adverse cardiovascular events”,Eur Heart J, Vol. 33, 734-744.
3. http://en.wikipedia.org/wiki/Case-based_reasoning.Accessed on 2013.10.5.
4. David Poole, Alan Mackworth. (2010). Artificial Intelligence: Foundations of Computational Agents, *Cambridge University Press*.
5. Cunningham, P. (1998). CBR: Strengths and Weaknesses. *Proceedings of 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, LNAI 1416, Vol. 2, 517-523, Springer.
6. AgnarA.,EnricP. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches, AI Communications, 7(1), 39-59.

7. Ramon M., David M., Derek B., David L., Barry S., Susan C.,Boi F., Mary L. M., Michael T. C., Kenneth F., Mark K., AgnarA.andIan W. (2005), Retrieval, reuse, revision, and retention in case-based reasoning, *Knowledge Engineering Review*, 20(3), 215-240, 2005.
8. Joseph M. Hilbe,Logistic Regression Models. *Chapman & Hall/CRC Press*, 2009.
9. Vladimir P., Ashutosh G., James M. R. and Thomas S. H. (2000). Multimodal speaker detection using error feedback dynamic Bayesian networks. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Vol. 234-41.
10. Kim I.C., Jung Y.G. (2003). Using Bayesian Network to analyze Medical Data. LNAI2734, *Springer Berlin Heidelberg*, pp.317-327.
11. http://archive.ics.uci.edu/ml/datasets/Heart+Disease.Accessed on 2013.10.5.
12. Bairey C. N., Rozanski A., Maddahi J., Resser K. J., Berman DS. (1989). Exercise thallium-201 scintigraphy and prognosis in typical angina pectoris and negative exercise electrocardiography. *Am J Cardiol*, 64(5), 282-287.
13. Davis T, Bluhm J, Burke R, Iqbal Q, Kim K, Kokoszka M, Larson T, Puppala V, Setterlund L, Vuong K and Zwank M. (2012). Diagnosis and Treatment of Chest Pain and Acute, *Institute for Clinical Systems Improvement, Coronary Syndrome (ACS)*, http://bit.ly.ACS1112, Updated Nov 2012.
14. Peter F. C. (1983). Prognosis and treatment of asymptomatic coronary artery disease. *Journal of the American College of Cardiology*, 1(3), pp. 959-964.
15. http://en.wikipedia.org/wiki/Regression_analysis. Accessed on 2014.3.6.
16. http://download1.medcalc.org/medcalcmanual.pdf. Accessed on 2014.3.6.
17. Mark H., Eibe F., Geoffrey H., Bernhard P., Peter R. and Ian H.W. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).
18. Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes and Ian H. Witten. (2004). Data Mining in Bioinformatics using Weka. *Department of Computer Science, University of Waikato*, Bioinformatics Advance Access, 2004.
19. Bonal, J. R., Lorenzo Calvo, A., & Jiménez Saiz, S. L. (2019). Key Factors on Talent Development of Expertise Basketball Players in China. Revista de psicología del deporte, 28(3), 0009-16.