How to cite this paper:

Sukumar Letchmunan, Zulkefli Mansor, Nikki Lee Wan Yan, Low Kah Meng, & Nur Farhana Izwani Tahir. (2017). Predictive analytic in health care using case-based reasoning (CBR) in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference on Computing & Informatics (pp 8-15). Sintok: School of Computing.

PREDICTIVE ANALYTIC IN HEALTH CARE USING CASE-BASED REASONING (CBR)

Sukumar Letchmunan[1], Zulkefli Mansor[2], Nikki Lee Wan Yan [3], Low Kah Meng[4], and Nur Farhana Izwani Tahir[5]

[1] *Universiti Sains Malaysia, sukumar@usm.my*
[3]*Universiti Kebangsaan Malaysia, kefflee@ukm.edu.my*
[3] *Universiti Sains Malaysia, nikki.ucom13@student.usm.my*
[4] *Universiti Sains Malaysia, lkmeng.ucom13@student.usm.my*
[5] *Universiti Sains Malaysia, farhana.ucom13@student.usm.my*

ABSTRACT. Big data analytics enables useful information to be extracted in order to predict trends and behavior patterns. Predictive analytics can be applied in health care industry by using the information gained from big data analytics. There are several methods to make predictive analytics. Case-based Reasoning (CBR) is one of the methods to make prediction on patients' sickness based on previous experiences. There are several challenges when applying CBR to predictive analytics. This paper focuses on solving the number of analogies used when applying CBR. Experiments and calculations are done to compare the accuracy of the number of analogies used. The results shows one analogy has the highest accuracy as compared to two and three analogies.

Keywords: prediction, big data, health care, case based reasoning

INTRODUCTION

Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results (S.Sagiroglu, 2013). Data plays an important role in health care industry. Large amount of medical data are generated and collected daily in health care industry (B. Milovic, 2012). However, those medical data are not fully structured and utilized to help to improve health care industry. The huge amount of medical data is also too big to be analyzed by humans alone. Hence, big data analytics is needed to extract data and analyze those data to reveal the hidden patterns and behavior trends within the data so that more useful information and knowledge is produced. Big data analytics is the process of research into massive amount of data to reveal hidden patterns and secret correlation (S.Sagiroglu, 2013). It also involves data analysis from multiple perspectives to find the relations and connections between data that does not seem to be related. It enables useful information to be extracted from the medical data to predict trends and behavior patterns. Deeper knowledge between the relations and connections between the data will also be known. The information gained from big data analytics provide support for predictive models and as a result, predictive analytic in health care. Thus, predictive analytics can be applied in health care to predict sickness of patients based on previous experiences and make a huge impact to the need of public.

HEART DISEASE

Heart disease is a general name for a variety of diseases and disorders where the conditions affect the heart and blood vessels. The symptom may vary for different people depending on the type of heart disease. There are 25 different kinds of heart disease. For example, pectoris, essential hypertension and myocardial infarction. Most of the heart diseases have certain common symptoms that may occur to the patients that have heart disease. One of the most common heart diseases is congestive heart failure. Congenital heart failure referring to the problem with the heart's structure and the function that may cause abnormal heart development before birth. This type of heart failure happens when the heart does not pump blood to the other organs of the body.

EXISTING SYSTEM

Currently, there are several systems available that used data mining and supervised machine learning to make prediction in health care. There are a lot of researches regarding the number of data mining methods and predictive analytic techniques that are available and useful. Each methods and techniques also have their own pros and cons.

Collaborative Assessment and Recommendation Engine (CARE)

CARE is a patient-centred disease prediction and management (N. Chawla, 2013). CARE is suitable to serve as a data-driven computational aid for physicians assessing the disease risks facing their patients (N. Chawla, 2013). The algorithm used in CARE is based on collaborative filtering methodology used in recommendation system. It is a data mining method to predict a user's fondness towards a product based on the known preferences of other users. The basic idea for this methodology is that if users have similar preference towards certain products, there is a high chance to have similar preference towards other products as well. Collaborative filtering search for the similarity between data. This technique enables prediction of sickness to be made based on the similarity of existing medical data.

CARE is a system that leverages the similarities among the patients and constructs personalized disease risk profiles for individual. CARE considers patients' previous medical history and demographics. It makes prediction using the data from similar patients. CARE makes disease prediction by considering the individual medical history and other patient's medical history. The system will then find the similarity between those data sets. Then, collaborative filtering is applied by calculating inverse frequency, vector similarity and best match subset. After that, iterative version of CARE (ICARE) was built. ICARE is an ensemble of collaborative filtering models for each active patient (N. Chawla, 2013). Finally, a predictive analytic result is produced.

The advantage of CARE system is the system takes the users' demographics and previous medical history into account when making prediction. Other than that, CARE may also give enlightenment for doctors regarding relatively rare diseases as this system is based on data from other similar patients. Hence, doctors might be able to find patterns and behavioral trend about the disease. On the other hand, the disadvantage of this system is that diseases do not have rating or ranking system like products. The system can only know that if the user has been diagnosed with the disease before. The absence of this disease cannot be disregard as it could mean that the user may not have, yet to have or diagnosed with the disease.

Survival causes after patients out of hospital cardiac arrest prediction system

The research paper 'Predicting Survival Causes after out of Hospital Cardiac Arrest using Data Mining Method' states that the hidden relationships between the events of the sample can be found after exploring the data by using Bayesian network. The existing system is a

system that is used to predict the survival causes after patients out of hospital cardiac arrest. The system predicts the most significant aspect for patient survival. It uses Bayesian network to predict the aspect for patient survival. There are three steps in Bayesian network. The first step is learning step. Taboo order method is used to build network in the learning step. There are five variables related to the probability of being alive after cardiac arrest. The five variables are age, sex, initial cardiac rhythm, origin of heart failure and type of specialized resuscitation employed (F. L. Duff, 2004). The second step is analysis of associations. The relationships between the variables and the main node in the network which is alive or death is analyzed. The data is analyzed using hidden Markov model. This model allowed us to deduce what the variables directly related with the target node were (this model provides a simplified model with the target node and the relationship with the nodes which represent parents and sons) (F. L. Duff, 2004). Finally, the last step is do inference and prediction. The hidden Markov layer is examined and prediction is made.

This system helps to increase the survival rate among patients after a cardiac arrest. It also benefits the patients by giving guidance to the patient to change the lifestyle accordingly. The con of this system is adequate data is needed to produce a model.

Fuzzy Expert System for Determination of Coronary Heart Disease (CHD) Risk In this system, Framingham Risk Scoring is used for risk assessment to determine 10 years risk for CHD development by using the attributes such as age, total cholesterol, HDL cholesterol, systolic blood pressure, treatment for hypertension, and cigarette smoking. Among the attributes, researcher found that the main important factor that affects the Coronary Heart Disease (CHD) is cholesterol level. Low density lipoprotein (LDL) cholesterol and high density lipoprotein (HDL) cholesterol are required to provide more meaningful indicators of CHD risk.

Calculation of the number of points for each risk factor is the first stage of this system. After that, the blood pressure was taken into account at the time of assessment to determine whether the person needs to be on anti-hypertensive therapy. Extra point will be added if the person is on anti-hypertensive treatment. The sum of the points of each risk factor is known as total risk score. From the total risk score, the 10-years risk for coronary death is estimated. This system calculates step by step for the next 10 years of CHD risk.

There are three categories of risk in CHD which are (10 years risk > 20%), multiple risk factor (10 years risk 10-20% and 10 years risk < 10%), and 0-1 risk factor. After the risk score was calculated and reached the target of the classes, it will display the results and recommend three outputs which are normal living, diet and grog treatment.

CASE-BASED REASONING (CBR)

Case-based reasoning is a problem solving method that is different from other major AI approaches. This method is analogous to problem solving that compares new cases with previous indexes cases (S. Ying, 2015). In other words, it is reusing previously known cases to solve the problem. CBR is using the already stored knowledge and captures new knowledge to make it quickly available for solving the next problem. CBR provide two main functions: storage of new cases in the database through indexation module (S. Ying, 2015) and searching the indexes cases with the similarities of new cases in case retrieval module.

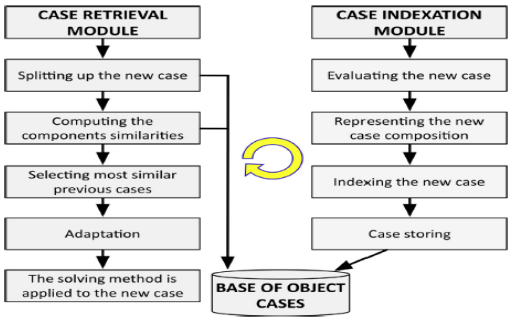

Figure 1. Case Indexation And Case Retrieval Module

General CBR cycle can be described by the four following process:

1) RETRIEVE the most similar use case to problem description.

2) REUSE the information and solution in the case to solve the problem.

3) REVISE the proposed solution to fit the new problem.

4) RETAIN the solution that is useful once it has been confirm for future problem.

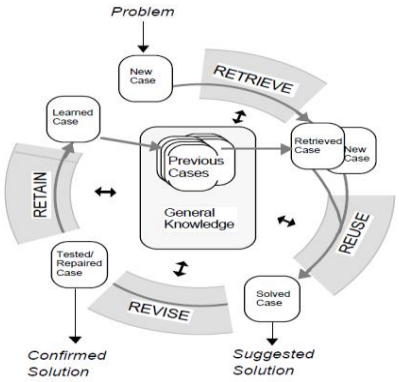

Figure 2. Process Of CBR Cycle (A. Aamodt, 1994)

CHALLENGES IN CBR

There are several problems when CBR is applied to predictive analysis. The problems that most researchers encounter are in the following categories (Shepperd, 2001):

a) Selecting feature subset.

It is not very obvious to know that which features are useful. The dataset might have a lot of features. However, not all of the features are applicable to make a prediction. The features may have error data or redundant. Therefore, decision to remove useless features is known as feature subset selection problem.

b) Scaling

In scaling, attributes values are transformed according to defined rules in order for the attributes to be measured using the same unit.

c) Similarity measure

A distance measure in CBR is the degree of similarity between two projects in terms of their effort drivers (A. Kofod-Petersen, 2006). Categorical data usually assign the value 1 for data that matches and 0 for data that does not match for similarity measure.

d) The number of analogies

It means how many alike cases are used for the prediction is consider as suitable. There is no fix numbers of analogies when applying CBR in prediction. Usually, the number of analogies used is from the range of one to three analogies. Last but not least, analogy adaptation is another problem when applying CBR for prediction.

e) Analogy adaption

It involves the generation of estimation after the analogies are retrieved. Nearest neighbor or the mean of analogies are the examples of methods that can be used in analogy adaption.

In this research paper, the focus is on the number of analogies to be used in the prediction. Experiment will be done for one to three analogies in order to get a most accurate prediction result. The further details will be explained in the following sections about the number of analogies.

RESEARCH METHODOLOGY

UCI Machine Learning Repository is a collection of databases, domain theories and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms (M.Lichman, 2013). Heart disease data set from UCI Machine Learning Repository is being used for this research. There are four databases in this data set namely Hungarian database, Long Beach database, Switzerland database and Cleveland database. The four databases are compared and analyzed. Cleveland database is chose as the main database for the research as Cleveland database has the least amount of missing values in contrast to the other databases. The data in the Cleveland database is gathered from the V.A Medical Center, Long Beach and Cleveland Clinic Foundation by Robert Detrano, M.D, Ph.D. Patients' name and social security number are removed from the database. This database has 14 attributes. The list of attributes and their descriptions is as shown in Table 1.

Table 1. Table Of Attributes

No.	Attributes	Description	Values
1	age	Age	Value in years
2	sex	Sex	• Value 1 = male • Value 0 = female
3	cp	Chest pain type	• Value 1: typical angina • Value 2: atypical angina • Value 3: non-angina pain • Value 4: asymptomatic
4	trestbps	Resting blood pressure	Value in mm Hg on admission to the hospital
5	chol	Serum cholesterol	Value in mg/dl
6	fbs	Fasting blood sugar	• Value 1 : If values > 120 mg/dl • Value 0 : If values ≤ 120 mg/dl
7	restecg	Resting electrocardiographic results	• Value 0: normal • Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) • Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	thalach	Maximum heart rate achieved	Value in beat per minute
9	exang	Exercise induced angina	• Value 1 = yes • Value 0 = no
10	oldpeak	ST depression induced by exercise relative to rest	Continuous value
11	slope	Slope of the peak exercise ST segment	• Value 1: up sloping • Value 2: flat • Value 3: down sloping
12	ca	Number of major vessels coloured by fluoroscopy	0-3
13	thal	Defect type	• Value 3 = normal • Value 6 = fixed defect • Value 7 = reversible defect
14	num	Presence/Absence of heart disease	• Value 1 = absence • Value 2,3,4,5 = presence

The 14th attribute, num represent if the patients have heart disease. The attribute ranges from 1 to 5 to discriminate between absence and presence of heart disease. 1 represents ab-

sence of heart disease while 2, 3, 4 and 5 represents presence of heart disease. However, in the original database, the 14th attribute, num is represented from the value 0 to 4. For the research purposes, the values is increased by 1 in order to calculate the magnitude of relative error (MRE) and mean magnitude of relative error (MMRE) in the next section.

The Cleveland database is then being filtered for missing values. The records which have missing values are removed from the database in order to have a more accurate prediction in the research. Initially, there are 303 records in the database. However, record number 88, 167, 193, 267, 288 and 303 are being removed due to missing value. Record number 88 and 267 have missing value for the 13th attribute which is thal, the defect type. Meanwhile, the other records have missing value for the 12th attribute, ca that represents the number of major vessels colored by fluoroscopy. As a result, there are 297 records left in the database.

In order to predict whether a patient has heart disease or not, case-based reasoning method is used. As mentioned previously, case-based reasoning compares the new cases to the cases which already in the database. Hence, the 297 records in the database will be used to compare and contrast with the new cases. K-nearest neighbor algorithm is used to measure the distances between the new cases and the cases in the database. Distance need to be measured to get to know who the nearest neighbors to the new cases are. In other words, distance is measure in order to know which cases in the database are the most similar to the new cases. Since all the attributes in the database are in terms of numerical data, therefore k-nearest neighbor algorithm is very suitable to be applied as the algorithm work the best with numerical data. The distance between two cases, x_i and x_j are calculated using the Euclidean distance. The distance between two cases is the square root of sum of square of the differences between two corresponding attributes. There are 13 attributes to be compared. Thus, the formula is (E. Mendes 2002):

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^{13} [x_{ni} - x_{nj}]^2}$$

After the distances between the new cases with all the cases in the database are calculated, the shortest distance among all the distances will be selected. The corresponding record with the shortest distance will be used as the results of the prediction. The 14th attribute which is the presence or absence of heart attack of that record is the results of the heart disease prediction.

However, there is no fix number of analogies to be used when applying CBR in doing prediction. Number of analogies refers to the number of similar cases is used for the prediction. Hence, accuracy of number of analogies used is going to be compared in this research. The number of analogies ranges from one to three analogies as it is common numbers of analogies used in CBR. A system is developed to test the accuracy of the number of analogies.

Jack-knife method or also known as leave one out cross validation is applied in this experiment (S. Letchmunan, 2010). 16 random cases were picked for the experiment. First, one case is removed from the database and the information from the case is used to make a prediction using the remaining cases in the database for one, two and three numbers of analogies. Then, the case is returned to the database. The steps are repeated for all the others 15 cases.

RESULTS

In order to compare the accuracy of the predicted results using different numbers of analogies, calculations using magnitude of relative error (MRE) and mean magnitude of relative error (MMRE) were done. The formula of MRE is as following (E. Mendes 2002):

$$MRE = \left|\frac{Result_{Actual} - Result_{Predicted}}{Result_{Actual}}\right|$$

Meanwhile the formula for MMRE is as below :

$$MMRE = \frac{1}{n}\sum_{i=1}^{i=n}\left|\frac{Result_{Actual} - Result_{Predicted}}{Result_{Actual}}\right|$$

The results of the predicted results, MRE and MMRE according to the number of analogies are shown in Table 2.

Table 2: Results of experiment

Case	Actual Result	Number of Analogies					
		1		2		3	
		Predicted Result	MRE	Predicted Result	MRE	Predicted Result	MRE
1	1	1	0	2	1	2	1
2	3	2	0.3333333	3	0	3.6666667	0.22222
3	2	3	0.5	2.5	0.25	3	0.5
4	1	1	0	1	0	1	0
5	1	1	0	1	0	1	0
6	1	1	0	1.5	0.5	1.3333333	0.33333
7	4	4	0	2.5	0.375	2	0.5
8	1	1	0	1	0	1.3333333	0.33333
9	3	2	0.3333333	2.5	0.1667	2.3333333	0.22222
10	2	1	0.5	1.5	0.25	1.3333333	0.33333
11	1	3	2	2.5	1.5	2	1
12	1	1	0	1	0	1.3333333	0.33333
13	3	2	0.3333333	2	0.3333	2	0.33333
14	1	1	0	1	0	1	0
15	1	1	0	2	1	1.6666667	0.66667
16	5	4	0.2	2.5	0.5	2.3333333	0.53333
MMRE		0.2625		0.3671875		0.394444444	

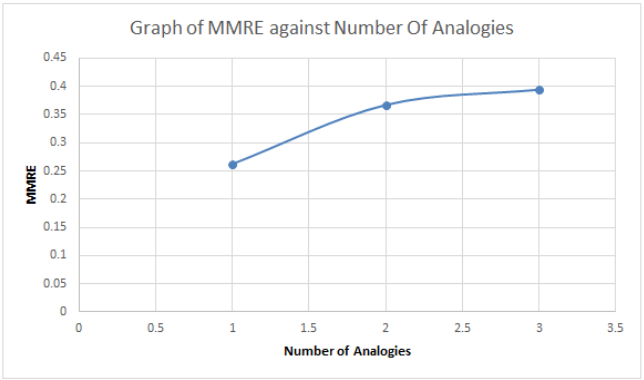

Figure 3: Graph of MMRE against Number of Analogies

Figure 3 shows the graph of MMRE against number of analogies. As shown in the graph, the MMRE for one number of analogies is the least as compared to two and three numbers of analogies. This shows that the difference between the predicted results and the actual results is the least for one analogy. The MMRE increases as the number of analogies increases. This shown that as the number of analogies increases, the predicted results differs further from the actual results. Therefore, prediction using only one analogy produced more accurate results in contrast to two and three analogies.

CONCLUSIONS AND FUTURE WORKS

Big data analytics helps to reveal hidden patterns and secret correlation from large amount of data. Useful information can be extracted through big data analytics and that information can be used in predictive models. Hence, predictive analytics are able to be applied in the health care industry.

CBR are used in predictive models as CBR compares the new cases with the previously known cases. CBR retrieve the most similar case to the new cases and reuse the information and solution in the case to solve the problem of the new cases. Despite that, there are a few challenges in applying CBR to predictive analytics. One of the problems that were being focus is the number of analogies. Number of analogies refers to the number of alike cases is used for the prediction. The number of analogies experimented were 1, 2 and 3.

The main discovery that can be found from this research is that prediction using case-based reasoning is most accurate when one analogy is used. The value of MMRE calculated for one analogy is the smallest as compared to the MMRE for two and three number of analogies. This also means the predicted results and the actual results have the least difference for one analogy in contrast to two and three number of analogies.

However, in the experiment only 16 cases are being tested for the MRE and MMRE values. Hence, in the future, more random cases can be populated to visualize the patterns in order to have better understanding on the number of analogies.

REFERENCES

A. Aamodt, E. Plaza,(March 1994), Case-Based Reasoning: *Foundational Issues,Methodological Variations, and System Approaches, AI Communications*, Vol. 7 Nr. 1, pp 39-59

A.Kofod-Petersen,(2006),Challenges in Case-Based Reasoning for Context Awareness in Ambient Intelligent Systems, *8th European Conference on Case Based Reasoning*, Workshop Proceedings, Ölüdeniz.

B. Milovic, M. Milovic,(December 2012), *Prediction and Decision Making in Health Care using Data Mining, International Journal of Public Health Science (IJPHS)* Vol. 1, No. 2, pp. 69~78.

E. Mendes, N. Mosley,(2002), Further Investigation into the Use of CBR and Stepwise Regression to Predict Development Effort for Web Hypermedia Applications, *Empirical Software Engineering*.

F. L. Duff, C. Muntean, M. Cuggia, P. Mabo.(2004) Predicting Survival Causes after out of Hospital Cardiac Arrest using Data Mining Method. *Stud Health Technol Inform*.

Lichman, M. (2013). *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. [Online] Available at: http://archive.ics.uci.edu/ml

N. Chawla, D. Davis.(2013 Sep),Bringing, Big Data to Personalized Healthcare: A Patient Centered Framework, *J Gen Intern Med*.

S. Letchmunan, M. Roper, M. Wood,(2010), Investigating Effort Prediction of Web-based Applications using CBR on the ISBSG Dataset, *14th International Conference on Evaluation and Assessment in Software Engineering (EASE)*.

S. Ying , C. Jo, J. Armelle , L. Kai,(2015 Aug). Emerging medical informatics with case-based reasoning for aiding clinical decision in multi-agent system, *J Biomed Inform*.

S.Sagiroglu, D. Sinanc.(2013), Big Data: A Review, *Collaboration Technologies and Systems (CTS)*.

Shepperd, M. and Kadoda, G.(2001), "Using Simulation to Evaluate Prediction Techniques," *Seventh International Software Metrics Symposium (METRICS'01)*.