A Case Based Reasoning Framework for Prediction of Stroke Disease

Pattanapong Chantamit-o-pas[1,1] , Madhu Goyal[1]

[1] Faculty of Engineering and Information Technology
University of Technology Sydney, PO BOX 123, Broadway, NSW 2007, Australia
Pattanapong.Chantamit-o-pas@student.uts.edu.au, Madhu.Goyal-2@uts.edu.au

Abstract. Case-based reasoning (CBR) has been a popular method in healthcare sector from the last two decades. It is used for analysis, prediction, diagnosis, and recommending treatment for patients. This research purposes a conceptual CBR framework for stroke disease prediction that uses previous case-based knowledge. The outcomes of this approach not only assist in stroke disease decision-making, but also will be very useful for prevention and early treatment of patients.

Keywords: case-based reasoning, stroke disease, decision-making, prediction

1 Introduction

Stroke is the second or third most common cause of death in most countries [1, 2]. The patients who survived usually have poor quality of life because of serious illness, long-term disability and become burden to their families and health care system. There is a strong demand for the management focused on prevention and early treatment of diseases by analysing different factors. Several health conditions and lifestyle factors have been identified as risk factors for stroke. These factors have three groups that consist of the risk factors cannot be change, the risk factors can be changed (treated or controlled), and other risk factors are less well-documented. The risk factor cannot be changed are focused on demographic data such as age, heredity (Family history), race, sex (Gender), and prior stroke, Transient Ischemic Attack (TIA) or heart attack. Some patients have had some behaviour and/or other disease before stroke attack. Furthermore, they are trying to control health and behaviour (as personality behaviour, and eating behaviour) and changed a quality of life that can prevent from stroke disease, for example, hypertension, many kinds of heart diseases such as myocardial infarction (MI), diabetes mellitus (DM), valvular heart diseases and atrial fibrillation, asymptomatic carotid artery disease, blood lipids, and smoking. The other risk factors that less well-documented are geographic location, socioeconomic factor, alcohol abuse, and drug abuse [3]. Recognition of these risk factors is important to reduce the incidence of stroke, which has been increasing [4].

[1] Faculty of Engineering and Information Technology, University of Technology, Sydney.

Case-based reasoning (CBR) is a methodology for solving problem that uses a previous data or memorized problem situations called cases. The processes of CBR system proceed in four main steps such as *retrieve, reuse, revise, and retain* (Fig. 1) [5]. The new case starts at the top of stage, where an input is entered into the system. The previous case is compared to the new case and start *retrieve* step. In practical CBR system is a comparison between all the cases in the system and a new case, then result will list the ranking of similar cases.

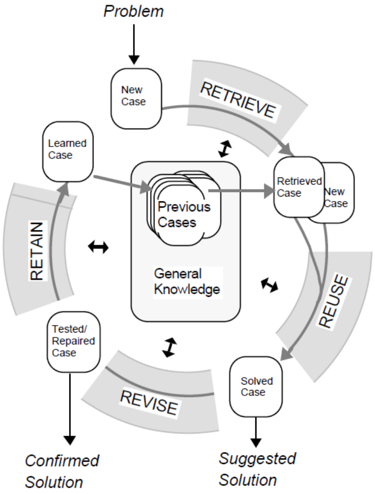


Fig. 1. The CBR cycle implemented by Aamodt and Plaza [5]

In this research, we propose a conceptual case-based reasoning framework to predict from patient risk factors and to recognize case that probably develop stroke or preparing patients to handle diseases burden outcome. This framework is comparing stroke patients in database and predicting patients who have risk factors which are related to stroke disease such smoking, high blood pressure, and so on. It would not only support to medical professionals for stroke disease decision making, but also provide suggestion and warnings to patients before they visit a hospital or goes for costly medical checkups.

The rest of this paper is organized as follows. Section 2 is the related work which reviews case-based reasoning (CBR) in healthcare sector as well as in other domains. Section 3 proposes the conceptual CBR framework for stroke disease. The conclusion and future work is presented in section 4 of this paper.

2 Related work

This section reviews the research done on case-based reasoning in various domains and also case-based reasoning in healthcare sector.

2.1 Case-based reasoning in healthcare

The case-based reasoning systems have many application areas in healthcare sector which have provided solutions for diagnosis and treatment of diseases based on past experiences. For example, the mixture of experts for case-based reasoning (MOE4CBR) [6] is an application for high-dimensional biological domains to prediction to disease. The data sets are used in ovarian mass spectrometry, leukemia and lung microarray data sets. The biomedical domains are complex, but also a system is unsuitable method for this research. They used data-mining and a logistic regression method applied in a system and also improved the classification performance. A case is defined by logistic regression approach that supports to filter the important feature in CBR. Similar cases are also grouped by data-mining technique. Thus, the system also supports for the "dimensionality" problem in this domain. For complex medical diagnosis, if patients have a complex disease, more medical domains have to be used for this. For example, the Premenstrual syndrome (PMS) is related among gynaecology and psychiatry and also need complex algorithm for diagnosis. The CBR-based expert system used the k-nearest neighbor (k-NN) algorithm to search k similar case that focusing on the Euclidean distance measure [7]. A CBR in treatment and management of diabetes is also represented in an application. It solved problem by using patient health record as demographic data, laboratory result, and physical examination. Those are compared with previous case by using k-NN algorithm [8]. For a complex data, a CBR has applied by using machine-learning and data-mining technique that based on gene expression profiles. This method used k-NN with weighted-feature based technique to retrieved and compared among previous cases and new cases. The herein-proposed methodology used on several data sets in this framework. The results shown that how many percentage of gene expression profile of a new patients are similarity previous cases and to help predict at risk of disease [9].

Moreover, Sharaf-el-deen, Moawad and Khalifa [10] is introduced the automated adaptation process, which applies the adaptation rules for solving the new case. To evaluate the approach, the researchers develop the prototype for diagnosing breast cancer and thyroid diseases. They proposed a hybrid based medical diagnosis approach in order to enhance the performance of the CBR retrieval system. The main idea of the proposed approach is to combine both case-based and rule-based reasoning. In addition, Ahmed et al. [11] apply various data processing and feature extraction techniques by considering time and frequency domains for disease prediction. Given input data, the CBR system discovers the relevant cases and then creates an alarm based on the output. To evaluate the proposed system, the researchers is compared it with the classification results from experts.

Furthermore, Amin, Agarwal and Beg [12] proposed clinical decision support systems for disease prediction and diagnosis. These approaches are able to extract hidden pattern and relationships among medical data. This leads the proposed approaches to be efficient for designing the decision support systems.

2.2 Case-based reasoning in other domains

The CBR had been applied in other sectors; such as information technology, educational technology, bankruptcy prediction modelling, and so on. Jonassen and Hernandez-Serrano [13] stated that problem solving on organization is complex that can solve the program from previous case or similar case. Normally, organization had been applied the lessons learned from their old stories to the new problems that has significance to decision-making and to justify the use of previous case as instruction support. In addition, Bryant [14] proposed a CBR to bankruptcy prediction modeling. He stated that financial company has risk to many factors as stakeholders, customers, investors, managers, and employees. This model used various factors from financial statement in 500 firms from 1990-1994 in nonbankrupt and 14 firms between 1990-1994 in bankruptcy case and also used clustering and decision tree technique for analyse and prediction to bankruptcy in organization.

Moreover, a CBR is integrated with a fuzzy decision tree (FDT), and genetic algorithm (GAs), called “The hybrid classification model” [15]. The approach aimed to develop a decision-making system for solving classification problems among various databases. The case-based approach is used for clustering data into small cases and then the genetic algorithms are applied for enhancing the fuzzy decision tree.

3 Case-based Reasoning Framework for Stroke Patients

The overview of framework proposed for prediction method is shown in Fig 2. The framework consists of 6 processes for prediction of stroke patient: clustering process, retrieval process, reusing process, prediction process, retain and review process, and store process respectively. This framework has two processes; clustering process and prediction process (which are shown in dot line) which differ from the original framework.

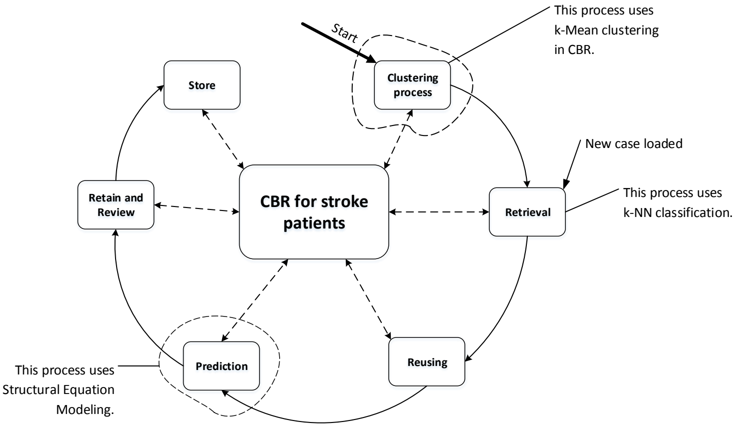

Start
Clustering process
This process uses k-Mean clustering in CBR.
Store
New case loaded
Retain and Review
CBR for stroke patients
Retrieval
This process uses k-NN classification.
Prediction
Reusing
This process uses Structural Equation Modeling.

Fig.2. An overview of the case-based reasoning framework for prediction of stroke disease.

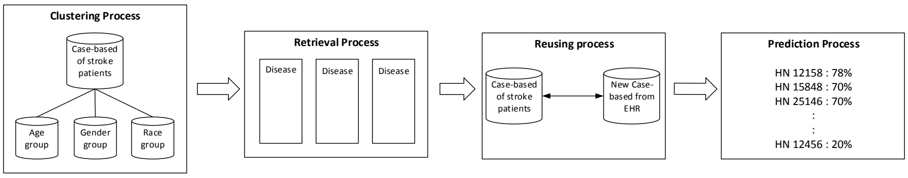


Fig. 3. Flowchart of the case-based reasoning for prediction of stroke patients.

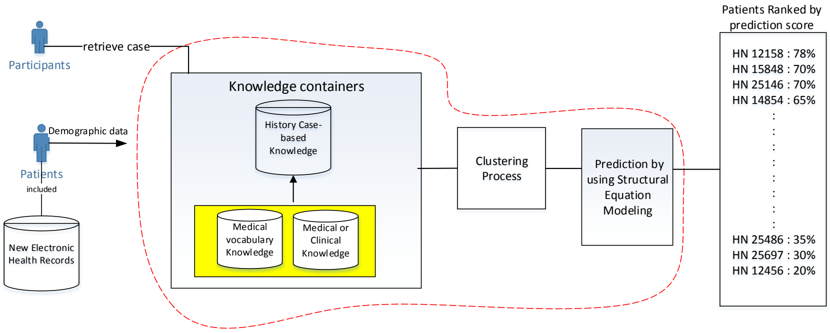
Flowchart of the proposed framework is presented in Fig 3. The detail of processes describe in below as follows:

a) Clustering process –this process aims to cluster stroke patient records, based on age, gender, and race of patients. Those clusters are important factors to predict stroke disease. K-Mean clustering technique is applied for finding groups to partition n observations into k clusters. The basic algorithm is given by equation (1) [16]:

$$j = \sum_{i=1}^{k} \sum_{x \in S_i} \|x_i - c_i\|^2 \qquad (1)$$

We assume that (x_1, x_2, x_3,…., x_n) is a collection of observations; where x_i is the i^{th} dimensional real vector. The observations are partitioned into k groups; s = {s_1, s_2, s_3,…, s_k}, and c_j is mean of s_j.

b) Retrieval process – this process is retrieval in which an electronic healthcare records (EHR) is compared with information stored in "*knowledge containers*" [17]. A CBR system for stroke patients includes a cased-based knowledge, two medical knowledge databases (medical vocabulary knowledge and medical or clinical knowledge), and EHR (Fig. 4). The medical vocabulary knowledge contains stroke vocabularies and related diseases. Knowledge from experts is represented in clinical knowledge for a hospital with an acute stroke unit. Given output from the previous process, this process uses k-nearest neighbor (k-NN) approach based on medical and vocabulary knowledge to classify each patient group into risk factors.



Participants
retrieve case
Demographic data
Patients
included
New Electronic Health Records
Knowledge containers
History Case-based Knowledge
Medical vocabulary Knowledge
Medical or Clinical Knowledge
Clustering Process
Prediction by using Structural Equation Modeling
Patients Ranked by prediction score
HN 12158 : 78%
HN 15848 : 70%
HN 25146 : 70%
HN 14854 : 65%
:
HN 25486 : 35%
HN 25697 : 30%
HN 12456 : 20%

Fig. 4. The detail of the case-based reasoning for prediction of stroke patients.

To calculate the distance between p and q in k-NN algorithm, Equation (2) is applied [18].

$$\mathrm{dist} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad (2)$$

; where p = (p_1, p_2, p_3,…, p_n) , q = (q_1, q_2, q_3,…, q_n) and n is the number of dimensions.

c) Reusing process – this process aims to match cases that are relevant to the given risk factors from the previous process. As we mentioned above, cases are collected from the real cases in the hospital and stored in the knowledge container. In this paper, we use those cases for stroke prediction in the next process.

d) Prediction process - data mining and statistical methods are well-known for dealing with medical data analysis and prediction. To properly select tools and develop prediction models, general and incomplex guidelines are necessary and required [19]. This process uses structural equation modeling (SEM) because the data type has shown to multiple groups in current patient records and risk factors if they got it such as diabetes data set, heart disease data set, behaviour data set, and so on, which is a data set that depend on stroke and other diseases. The SEM supports multiple values to prediction and can be described "The basic statistic of SEM is the covariance, which is defined for two continuous observed variables X and Y. where r_{XY} is the Pearson correlation and SD_X and SD_Y are their standard deviations. A covariance represents the strength of the association between X and Y and their variablities, although with a single number. Because the covariance is an unstandardized statistic, its value has no upper or lower bound. [20]" The formula is given by (3):

$$COV_{xy} = r_{xy}\, SD_x\, SD_y \qquad (3)$$

In term of stroke disease, there exist various risk factors that are useful for effectively predicting disease. The analysis process identifies variables. The age values are independence variables (called "primary variables") and risk factors are dependence variables, then algorithm analyses and predicts between previous cases and current patient records. It processes case-by-case with other disease groups that relate to risk factors. After that, the output presents in terms of stroke risk estimation. For data sets of stroke that use in three main groups; such as the risk factor cannot be changed, the risk factor can be changed, treated or controlled, and other risk factors are less well-documented. The first group is demographic data such as age, race, gender, and Prior stroke. The second group consists of behaviour and historical disease from EHR such as Hypertension, Heart Disease, Atrial Fibrillation, Peripheral Artery Disease, Carotid, Diabetes Mellitus, Obesity, High Blood Cholesterol, Sickle Cell Disease, First Stroke, Alcohol Abuse, Poor diet, Physical Inactivity, Drug Abuse, and Smoking. The last group includes hometown of patient, socioeconomic factor, alcohol abuse, and drug abuse. These are loaded from current patient records. For stroke patients, incidence of stroke is required and loaded from historical records.

e) Retain and Review process - After that, the output will be verified and sent to participants or nurses. The result shows percentage of stroke for individual patient.

f) Store process - The prediction results of patients who have risk factors in stroke disease will be stored in CBR system for reuse in the future. This information can help in decision-making for participants in order to make a suggestion and warnings to patients as care plan, life style, quality of life, and behaviour and so on. Finally, the outputs are updated in historical case-based knowledge.

4 Conclusion and Future Work

A case-based reasoning has been applied for diagnosis diseases such as diabetes, leukemia and lung, premenstrual syndrome, and breast cancer and thyroid. In this paper, we have purposed the CBR framework for stroke disease. There are two processes which differ from the original case-based framework (clustering process and prediction process). The result of CBR framework is quite significant decision-making for patients. Specially, it can give suggestions and warnings to patient in spite of the fact that stroke do not have warning signs. Consequently, the proposed framework is beneficial for stoke disease management. In future, we will compare our framework with other prediction techniques and implement an e-stroke application.

References

1. Langhorne, P., Bernhardt, J., Kwakkel, G.: Stroke rehabilitation. The Lancet 377, 1693-1702 (2011)
2. Khosla, A., Cao, Y., Lin, C.C.-Y., Chiu, H.-K., Hu, J., Lee, H.: An integrated machine learning approach to stroke prediction. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 183-192. ACM, Washington, DC, USA (2010)
3. The American Heart Association, http://www.strokeassociation.org/STROKEORG/AboutStroke/UnderstandingRisk/Understanding-Stroke-Risk_UCM_308539_SubHomePage.jsp
4. Gorelick, P.B., Sacco, R.L., Smith, D.B., Alberts, M., Mustone-Alexander, L., Rader, D., Ross, J.L., Raps, E., Ozer, M.N., Brass, L.M.: Prevention of a first stroke: a review of guidelines and a multidisciplinary consensus statement from the National Stroke Association. Jama 281, 1112-1120 (1999)
5. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI communications 7, 39-59 (1994)

6. Arshadi, N., Jurisica, I.: Data mining for case-based reasoning in high-dimensional biological domains. IEEE Transactions on Knowledge and Data Engineering 17, 1127-1137 (2005)
7. Chattopadhyay, S., Banerjee, S., Rabhi, F.A., Acharya, U.R.: A Case-Based Reasoning system for complex medical diagnosis. Expert Systems 30, 12-20 (2013)
8. Kiragu, M.K., Waiganjo, P.W.: Case based Reasoning for Treatment and Management of Diabetes. Diabetes 145, (2016)
9. Anaissi, A., Goyal, M., Catchpoole, D.R., Braytee, A., Kennedy, P.J.: Case-Based Retrieval Framework for Gene Expression Data. Cancer Informatics 14, 21-31 (2015)
10. Sharaf-el-deen, D.A., Moawad, I.F., Khalifa, M.E.: A New Hybrid Case-Based Reasoning Approach for Medical Diagnosis Systems. J Med Syst 38, 1-9 (2014)
11. Ahmed, M.U., Banaee, H., Loutfi, A.: Health monitoring for elderly: An application using case-based reasoning and cluster analysis. ISRN Artificial Intelligence 2013, (2013)
12. Amin, S.U., Agarwal, K., Beg, R.: Genetic neural network based data mining in prediction of heart disease using risk factors. In: Information & Communication Technologies (ICT), 2013 IEEE Conference on, pp. 1227-1231. (Year)
13. Jonassen, D.H., Hernandez-Serrano, J.: Case-based reasoning and instructional design: Using stories to support problem solving. Educational Technology Research and Development 50, 65-77 (2002)
14. Bryant, S.M.: A case-based reasoning approach to bankruptcy prediction modeling. Intelligent Systems in Accounting, Finance & Management 6, 195-214 (1997)
15. Chang, P.-C., Fan, C.-Y., Dzan, W.-Y.: A CBR-based fuzzy decision tree approach for database classification. Expert Systems with Applications 37, 214-225 (2010)
16. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297. University of California Press, (1967)
17. Richter, M.M., Weber, R.: Case-Based Reasoning: A Textbook. Springer Science & Business Media (2013)
18. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier (2011)
19. Bellazzi, R., Zupan, B.: Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics 77, 81-97 (2008)
20. Kline, R.B.: Principles and practice of structural equation modeling (2015)