Trend of Case Based Reasoning for Chronic Disease Diagnosis: A Review

Daniel Hartono Sutanto¹, Nanna Suryana Herman¹, Mohd. Khanapi Abd. Ghani¹
¹Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka, Melaka 76100, Malaysia

Background: Chronic disease is defined as disease that persists over a long condition which progress slowly and generally it can be controlled but not cure. Type of chronic disease likely heart disease, stroke, cancer, chronic respiratory disease, liver, and diabetes mellitus. Case-Based Reasoning (CBR) is a case-based reasoning solves problems by using or adapting solutions to old problems. The CBR system has been used for diverse purposes likely classification, diagnostic, planning, and tutoring in the field of medical. However, trend of CBR for chronic disease diagnosis need to be reviewed due to the reliable and accuracy system has to be evaluated with parameter performance.

Method: In this review, it conducts from sources, such as books, journal, conference, report, etc. The publication year from 2010-2014. The population is a chronic disease and the interventions are CBR, method, techniques, and tool.

Result: Finally, the review shown that problem, dataset, method, proposed system, parameter performance and result. The 13 primary studies found that most of them using parameter performance in accuracy rate. Otherwise, the parameter includes robustness and learning capacity.

Conclusion: The most of problem found in primary studies, which is accuracy in retrieval cases, process time and missing data. In order to evaluate, parameters performance has to measure in particular CBR method.

Keywords: Case-Based Reasoning, Chronic Disease, Parameter Performance, Trend, Review

1. INTRODUCTION

The lives of many people in the world have been curtailed by chronic disease. A chronic disease defined as disease that persists over a long condition which progress slowly and generally it can be controlled but not cure. There is types of chronic disease, such as heart disease, stroke, cancer, chronic respiratory disease, liver, and diabetes mellitus. Nowadays, total world prevalence of diabetes is around 371 million and ASEAN has 3 countries who is being the prevalence of diabetes, such as Indonesia 7,4 million, Thailand 3,4 million, Malaysia 2 million.²

Initial Case-Based Reasoning (CBR) was triggered from the work of Schank and Abelson in 1977, where The first of CBR can be traced to Yale University and the work of Schank and Abelson in 1977.³ In 1994, Aamodt and Plaza⁴ proposed a life cycle for CBR systems which is used by other CBR researchers as a framework. At the highest level of a classification and generalist, a general CBR cycle may be described by the following:

a) Retrieve the most similar and suitable case or cases
b) Reuse the information, evidence, knowledge in that case to solve the problem
c) Revise the proposed solution with the learning process
d) Retain the phase of this experience appropriately to be useful for future problem solving

Watson stated that CBR to solve problems by using or adapting solutions to old problems⁵. It has been used for diverse purposes likely classification, diagnostic, planning, and tutoring in the field of medical⁶.
A medical diagnosis is a classification process. A physician has to analyze lots of factors before diagnosing the chronic disease which makes the physician’s job difficult. In recent times, machine learning and data mining techniques have been considered to design an automatic diagnosis system for diabetes. However, there are many methods and algorithms used to medical data sets for hidden information, including Neural networks (NNs), Decision Trees (DT), Fuzzy Logic Systems, Naive Bayes, SVM (Support Vector Machine), logistic regression and other methods. The algorithms are able to decrease processing time to analyze problems and producing diagnoses, making them more accurate at the same time. Meanwhile, this literature presents a great variety of methods related to the diagnosis and classification of chronic disease.

Hence, the trend CBR in diagnose chronic disease need to be reviewed due to the reliable and accuracy system has to be evaluated with parameter performance. The aim of this review explains the state of the art of the current problem and justify the parameter performance of CBR.

2. RELATED WORK

In this section, we describe related work with CBR in diagnosing disease. The related work purposes to enrich knowledge CBR method.

Subhagata found that CBR able to diagnose Premenstrual Syndrome (PMS) using k-Nearest Neighbor (kNN) to search k similar cases based on Euclidean distance measure. The novelty of T.A prototype software tool is designed a flexible auto-set tolerance (T).

Under research Intelligent Forensic Autopsy Report System (I-AuReSys), Yeow proposed system based on a CBR technique with a Naive Bayes learner for feature-weights learning.

In another case, Xiong stated hybrid CBR genetic-based fuzzy rule learning using Iris, Wine, and Cleveland (heart disease) data sets and finally the result shown that accuracy reaches to 93.25%.

Jha stated that CBR can be applied in Diabetes Detection and Care uses 4 Diabetes Support System TM (4DSS). The Intelligence System has aimed to: (a) automatically Detect problems in BG control; (b) propose solutions to detect problems; and (c) remember which solutions are effective or ineffective for individuals.

Other research on case liver disease, Lin used model Neural Network – Back Propagation Network and CBR to determine the type of liver disease. The best accuracy rate is 98.04%, while its using learning rate 0.1, momentum rate 39-9-1 and the architecture 39-9-13.

Hsu purposed hypertension detection used Genetic Algorithm and CBR. The dataset has false positive 172 subjects, whereas 44 false negative cases and accurate rate 94.8% accuracy rate.

Otherwise, Gimenez shown CBR system, kNN algorithm is combined with various information obtained from a Logistic Regression (LR) model using 1137 patients and prediction rate 92.7% after adding 50 random attributes.

Guessoun proposed RespiDiag using CBR to diagnose of chronic obstructive pulmonary disease. The different approaches for addressing the issue of missing data, it solved with optimistic, pessimistic or medium (offline or online) approach.

Nowadays, the CBR tools have been published such as jColibrí, eXiT*CBR.v2 and eXiTDSS. The researcher able to adopt CBR tool to diagnose many cases.

3. METHODOLOGY

This systematic literature review arranged based on the original guidelines as proposed by Kitchenham. The ways to do a systematic literature review method are explained at below.

3.1. Research Questions

To focus on review, research questions were specified. They were arranged with the PICOC criteria.

<table>
<thead>
<tr>
<th>ID</th>
<th>Research Question</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1</td>
<td>What is the trend CBR application for chronic disease diagnosis?</td>
<td>Identify the trend CBR application for chronic disease diagnosis.</td>
</tr>
<tr>
<td>RQ2</td>
<td>What are parameters to evaluate CBR performance?</td>
<td>Identify parameters were used for evaluating CBR performance.</td>
</tr>
</tbody>
</table>

- Population: chronic disease
- Intervention: case-based reasoning, method, techniques, tool.
- Comparison: n/a.
- Context: literature review in CBR method.

3.2. Study Selection

During the systematic review, we included study selection in diagnosing chronic disease using CBR methods. For the selection of primary studies, the following inclusion and exclusion criteria were used:

- Publication year: 2009-2014
- Significant: CBR method
- Object: chronic disease
- Source: book, report, journal, conference, etc.

4. RESULTS

4.1. Primary Studies

In doing a literature review, we collected 13 primary studies which are related to the adaptive CBR method in diagnosing chronic disease. An empirical study investigating a specific research question and its shown in Table 2.
<table>
<thead>
<tr>
<th>(Author) Ref.</th>
<th>Problem</th>
<th>Dataset</th>
<th>Application Domain</th>
<th>Proposed System</th>
<th>Parameter Performance</th>
<th>Result (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Chang et al., 2010)$^8$</td>
<td>The high dimensionality and non-stationary variations within the larger historical data</td>
<td>5 datasets (UCI Machine Learning)</td>
<td>Liver Disorder, Breast Cancer</td>
<td>Hybrid</td>
<td>-Accuracy</td>
<td>98</td>
</tr>
<tr>
<td>(Colantonio et al., 2010)$^{22}$</td>
<td>Unbalanced case-based classifier, ProtoClass, based on different medical datasets</td>
<td>699 cases, Wisconsin dataset (UCI Machine Learning)</td>
<td>Breast Cancer</td>
<td>ProtoClass</td>
<td>-Accuracy</td>
<td>96.48</td>
</tr>
<tr>
<td>(Tomar et al., 2011)$^{23}$</td>
<td>Knowledge acquisition, remembering, robust and maintenance</td>
<td>127 cases were collected for 14 occupational chronic lung diseases</td>
<td>Chronic Lung</td>
<td>-</td>
<td>-Accuracy</td>
<td>95.3</td>
</tr>
<tr>
<td>(Fan et al., 2011)$^9$</td>
<td>Multiclass problem needs to provide diagnosis aid accurately</td>
<td>Liver disorders dataset, Breast Cancer Wisconsin (Diagnosis)</td>
<td>Liver Disorder, Breast Cancer</td>
<td>CBFDT</td>
<td>-Accuracy</td>
<td>98.4 81.6</td>
</tr>
<tr>
<td>(Chuang, 2011)$^{24}$</td>
<td>Minimize possible bias, reduce problems false diagnosis.</td>
<td>166 outpatient visitors at a medical center in Taiwan from 2006 to 2008</td>
<td>Liver Disorder</td>
<td>Hybrid</td>
<td>-Accuracy -Sensitivity -Specificity -AUC</td>
<td>95</td>
</tr>
<tr>
<td>(Hsu et al., 2011)$^{15}$</td>
<td>Intolerant of noise and of irrelevant features</td>
<td>1152 subjects from Department of Health Examination of CGMC from July 2000 to July 2001</td>
<td>Hypertension</td>
<td>UWCBR</td>
<td>-Accuracy</td>
<td>94.8</td>
</tr>
<tr>
<td>(Huang et al., 2012)$^{25}$</td>
<td>Compare classifier system for reducing prediction errors</td>
<td>UCI Machine Learning dataset</td>
<td>Breast Cancer</td>
<td>-</td>
<td>-AROC</td>
<td>83.6</td>
</tr>
<tr>
<td>(Nnamoko et al., 2013)$^{26}$</td>
<td>Reduce the classification complexities of similar patterns</td>
<td>Diabetes UK 2012</td>
<td>Type 2 Diabetes Mellitus (DM)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(Kumari, R. Chitra, 2013)$^7$</td>
<td>Classifier for the detection of Diabetes disease with optimal cost and better performance</td>
<td>768 records of female patients Pima Indian diabetic -UCI machine learning laboratory</td>
<td>DM</td>
<td>-</td>
<td>-Accuracy -Sensitivity -Specificity -AUC</td>
<td>78 80 76.5</td>
</tr>
<tr>
<td>(Jha et al., 2013)$^{13}$</td>
<td>Actual problem a notion of similarity is used to retrieve similar cases from case bases</td>
<td>-</td>
<td>DM</td>
<td>DDCS</td>
<td>-Accuracy</td>
<td>65.4</td>
</tr>
<tr>
<td>(Cabrera, Marré, 2013)$^{27}$</td>
<td>Retrieval time, depth of the generated trees, and breadth-distribution of the nodes of the trees</td>
<td>8 datasets (UCI Machine Learning)</td>
<td>DM</td>
<td>NIAR k-d Trees -Time -Depth of generating tree -Nodes of the trees</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>(Guessoum, et al., 2014)$^{28}$</td>
<td>Missing data when in retrieval case</td>
<td>-</td>
<td>Chronic Obstructive Pulmonary</td>
<td>RespiDiag</td>
<td>-Accuracy</td>
<td>85.72</td>
</tr>
<tr>
<td>(Sharaf-El-Deen et al., 2014)$^{29}$</td>
<td>The accuracy of the retrieval- only CBR systems</td>
<td>2 datasets (UCI machine learning)</td>
<td>Breast Cancer, Thyroid</td>
<td>Hybrid</td>
<td>-Accuracy -AUC</td>
<td>99.33</td>
</tr>
</tbody>
</table>
4.2. RQ1: What is the trend CBR method for diagnosing chronic disease?
In Table 1, it shown author and year, problem, dataset, object, proposed system, performance parameter, and result. On other hand, this review summarize trend of hybrid CBR method, it shown on Diagram 1.

Diagram 1. Trend of Hybrid CBR Method

Most of problem found in primary studies, which is the accuracy in classification and prediction chronic disease. The accuracy be affected by similarity index in kNN method. Meanwhile, noisy data impact accuracy in retrieval case of CBR cycle. Noisy data two types of noise can be foremost in a given dataset noisy attribute and class attribute. Attribute noise is corruptions in the values of one or more attributes, examples erroneous attribute values, missing or unknown attribute values, and incomplete attributes or “do not care” values. To obtain high accuracy, we able to handle noisy data with filtering the noise. However, the robustness of the classifier or predict is able to produce correct predictions given noisy data or data with missing values. One of data mining technique is neural networks where are robust to noisy data as long as too many epochs are not considered since they do not over fit the training data.

UCI contributes to provide chronic disease dataset such as diabetes, breast cancer Wisconsin, heart disease, liver disorders, lung cancer, mammography, Pima Indian dataset, ILPD (Indian Liver Patient Dataset), and Thyroid. One of the trend CBR is diagnosing chronic disease with multi dataset, the result shown the improvement of retrieval time and DT capacity.

From primary studies, there is a chance to improve accuracy in diagnosing chronic disease such as diabetes mellitus, liver disorder and hypertension.

4.3. RQ2: What are parameters to evaluate CBR performance?
To analyze the performance of classification, the accuracy and AUC measures are adopted. Four cases are considered as the result of classifier.

- TP (True Positive): the number of examples correctly classified in that class.
- TN (True Negative): the number of examples correctly rejected from that class.
- FP (False Positive): the number of examples incorrectly rejected from that class.
- FN (False Negative): the number of examples incorrectly classified in that class.

The level of effectiveness of the classification model is calculated with the number of correct and incorrect classifications in each possible values of the variables being classified. From the results obtained the following equations are used to measure the accuracy, sensitivity, and specificity. In diagnostic systems divided two classes of events, essentially signals and noise. To do the analysis in term of the Relative Operating Characteristic (ROC) of signal detection theory provides a precise and valid measure of diagnostic accuracy. Receiver Operating Characteristic (ROC) curve is plotted between false positive rate and true positive rate and it describes the measure of positively predicting the disease. ROC curve is a technique for summarizing a classifier’s performance over a range by considering the tradeoffs between TP rate and FP rate, shown in Table 3.

### Table 3. Formula Parameter of CBR Performance

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>( \frac{TP}{TP + TN + FP + FN} )</td>
</tr>
<tr>
<td>Sensitivity (TP Rate)</td>
<td>( \frac{TP}{TP + FN} )</td>
</tr>
<tr>
<td>Specificity (FP Rate)</td>
<td>( \frac{FP}{FP + TN} )</td>
</tr>
<tr>
<td>Positive Predictive Value</td>
<td>( \frac{TP}{TP + FP} )</td>
</tr>
<tr>
<td>Negative Predictive Value</td>
<td>( \frac{TN + FN}{TN + FP} )</td>
</tr>
</tbody>
</table>

Sensitivity and specificity are statistical measures that describe how well the classifier discriminates between a case with positive and with negative class. Sensitivity is the spotting of disease rate that needs to be maximized and specificity is the false alarm rate that is to be minimized for accurate diagnosis.

Meanwhile, the CBR method is evaluated by robustness and learning capacity. Most of primary study has not evaluated in robustness and learning capacity yet, hence in order to evaluate the CBR performance the researchers have to calculate the accuracy, robustness and learning capacity.

5. CONCLUSIONS
The main objective of this systematic review was to find and assess primary studies validating the trend CBR method in diagnosing chronic disease. An extensive search was performed in order to find primary studies, which included searching seven digital libraries, snowball sampling and contacting the authors directly.

One characteristic of medical dataset is missing and noisy data. The benefit of neural networks where high
tolerance of noisy data as well as their ability to classify patterns on which they have not been trained.

Researchers who are new to the area can find the complete list of relevant papers in the field. They will also find the most frequently used the trend CBR in diagnosing chronic disease, with references to the studies in which they are used, including problem, method, dataset, parameter performance, proposed system and result.

The parameter performance of CBR is accuracy, robustness and learning capacity. The accuracy related parameter performa

REFERENCES