Journal of Computer Science 6 (2): 180-185, 2010
ISSN 1549-3636

Differential Diagnosis Knowledge Building by Using CUC-C4.5 Framework

[1,2]Kusrini, [1]Sri Hartati, [1]Retantyo Wardoyo and [1]Agus Harjoko
[1]Computer Science Program, Gadjah Mada University, Yogyakarta, Indonesia
[2]Department of Information System, STMIK AMIKOM Yogyakarta, Indonesia

Abstract: Problem statement: The Case Based Reasoning (CBR) method can be implemented in differential diagnosis analysis. C4.5 algorithm has been commonly used to help the method's knowledge building process. This process is completed by constructing decision tree from previously handled cases data. The C4.5 algorithm itself can be used with an assumption that all the cases has an exact and equal truth value thus have an exact contribution in decision tree building process. However, the decision makers sometimes not sure about the truth of the cases in the cases database, therefore the confidence value can be different for case by case. Besides that, the C4.5 algorithm can only handle cases that are stored in a flat table with data in form of categorized text or in discrete class. This algorithm has not yet explained about how is decision tree building mechanism in situation when the data are stored in relational tables. It also has not yet explained about the process of knowledge building when the data are in the form of number in continuous class. Meanwhile, the observed objects in this research, that is medical record data, are mostly stored in a complex relational database and have common form of categorized text, discrete number, continuous number and image. Therefore, the C4.5 is needed to be improved so it can handle decision building for cases database of medical record. **Approach:** We develop a knowledge building framework that can handle confidence level difference of cases in cases database. The framework we build also allows the data are stored in relational database. Moreover, our framework can process data in the form of categorized text, discrete number, continuous number and image. This framework is named CUC-C4.5, abbreviated from Complex Uncertain Case C4.5 as it is the improvement from C4.5 algorithm. **Results:** The CUC-C4.5 framework has been applied on the case of differential diagnosis knowledge building in a group decision support system to handle geriatric patient. This framework was implemented by using PHP and Javascript programming language and MySQL DBMS. **Conclusion:** The CUC-C4.5 can support differential diagnosis analysis on group decision support system for geriatric assessment.

Key words: C4.5, Framework CUC-C4.5, knowledge building, decision tree, differential diagnosis

INTRODUCTION

Case Based Reasoning (CBR) method allows making analysis or reasoning for a case based on previous experiences. With the method, every possibility of diseases for a patient (differential diagnosis) can be analyzed by using medical record data as the cases database. The analysis made by the method then can support doctors in order to determine differential diagnosis for a patient. There are many researchers found discussing the use of CBR method in medical world. Abidi (2002) has made a research about the use of CBR to manage people's personal health dynamically. He implemented it in a Java-based case based reasoning engine that applied a compositional adaptation algorithm to a personal health information package in HTML format that can be sent to user's email. Another application in clinical was implemented by Salam Khan (2003) for diet recommendation.

Beside medical, case based reasoning method can be also applied in many other disciplines. Niknafs *et al*. (2003) applied this method to generate recommendation in travel scheduling for newly tourist groups. They build a cases base from the previous tour experience, generally from the groups which ever makes a trip with their specification. After that, by using an adaptation criterion and equality function, it will result the best recommendation for the groups. In finding the similarity of a new case with its cases base, it is defined 4 factors a^1 to a^4 (a^1: Age, a^2: cost, a^3: Favorite, a^4: Origin).

Disease diagnosis' knowledge in this research is represented as a decision tree. The algorithm used in building the decision tree itself is an improvement of

Corresponding Author: Kusrini, Computer Science Program, Gadjah Mada University, Yogyakarta, Indonesia

C4.5 algorithm. The C4.5 algorithm was published by Quinland (1993) and Kohavi and Quinlan (1999). This algorithm was also used by Fan and Wen (2007) in their online test system. With this system, C4.5 algorithm was used to determine the grade of test.

C4.5 algorithm can be applied in knowledge building process with an assumption that all of the data in cases database have an equal level of confidence. As the consequent, all the data in cases database have equal contribution in knowledge building. However, in the real world, the most situations that happened are decision makers have different level of confidence about truth value of data in the cases database. Therefore, it is necessary to improve the C4.5 algorithm in order to consider the decision makers' different level of confident.

As the cases data, the C4.5 algorithm uses data in the form of categorized text and number in discrete class. Meanwhile, medical record data as the cases database in this research, contained data in some other forms that cannot be processed by the C4.5 algorithm. Categorized text, discrete number, continuous number, or even image, can be seen frequently in the medical record data. Furthermore, the C4.5 algorithm can work only on the data saved in flat table model (Quinlan, 1993) when the medical record mostly saved in complex relational database. These facts become our purposes to improve the classic C4.5 algorithm to handle problems related to the form of medical record data as the observed object.

C4.5 algorithm improvement, that is termed as CUC-C4.5, was implemented in knowledge building in group decision support system to handle geriatric patient. Geriatric patient is an old person patient that is handled by a geriatric team in a geriatric assessment (Darmojo, 2002). Geriatric team in this research's object hospital included geriatric consultant, internist, occupational therapist, neurologist, dentist, psychiatrist, nutritionist, social worker and pharmacist.

MATERIALS AND METHODS

The objects to be processed into knowledge base building machine are the data of medical records, selected variables and selected cases. Selected variable data are defined as variable data that are selected by decision makers as determinant variable. Selected cases data are defined as medical record data that are selected by the user as the cases base in the knowledge building. Medical record data can be whether in the form of singular data, unlimited plural data, or limited plural data (image). Singular data can be whether a discrete number, continuous number, or categorized text data.

In this research, we improved the decision tree building process by considering a weighting for each case representing users level of confidence for the case. The determinant variable that used in the decision tree building is in the form of categorized text, number in discrete class, number in continuous class and image. The target variable from this knowledge building process is the diagnosis and the determinant variables are medical record data. The determinant variables are grouped to three forms: Single variable, unlimited plural variable and limited plural variable. This grouping meets the existing mixture in medical record data.

The framework of this decision building model is named CUC-C4.5. The framework is shown in Fig. 1.

Singular variable defined in Fig. 1 as a variable that has single value in the assessment. For example: age, sex and smoking status. The data in singular variable can be in form of categorized text, continue or discrete data. Continuous data are defined as numerical data that are not sliced precisely, hence make the data variation is unlimited. For example in the temperature data, these data can be valued 30, 30.2, 30.25, 30.257, 30.2577 and 30.25772°C and so on. In contrast with the continuous data before, discrete data are numerical data that is sliced precisely and the values are only exist in particular points, for example children number (0, 1, 2, 3, 4,…). Categorized text data are text data that have been categorized wee like sex, religion, occupation and so on.

Limited plural variable is defined as variable that can have more than one value for each assessment, for example symptom, allergy and operation history. Unlimited plural variable is defined as variable that is specifically referred to image data. Data in this form will be extracted to 7 visual features, those are color moment order 1, color moment order 2, color moment order 3, entropy, energy, contrast and homogeneity. Limited plural and unlimited plural features are continuous data.

In this research, continuous data and discrete data with high variation of value will be mapped into discrete form that has number variation predefined by a categorization process. After all the data mapped to discrete class the next step will be cases table building and filling. These data in cases table become references in decision tree building.

The clustering process in this framework is using k-means clustering algorithm. K-means clustering is a popular algorithm in data clustering process activities. The use of this algorithm was published by MacQueen (1967) and Berry and Gordon (2004). This algorithm was also applied by Moon Kim *et al*. (2007) for clustering music in a musical website.

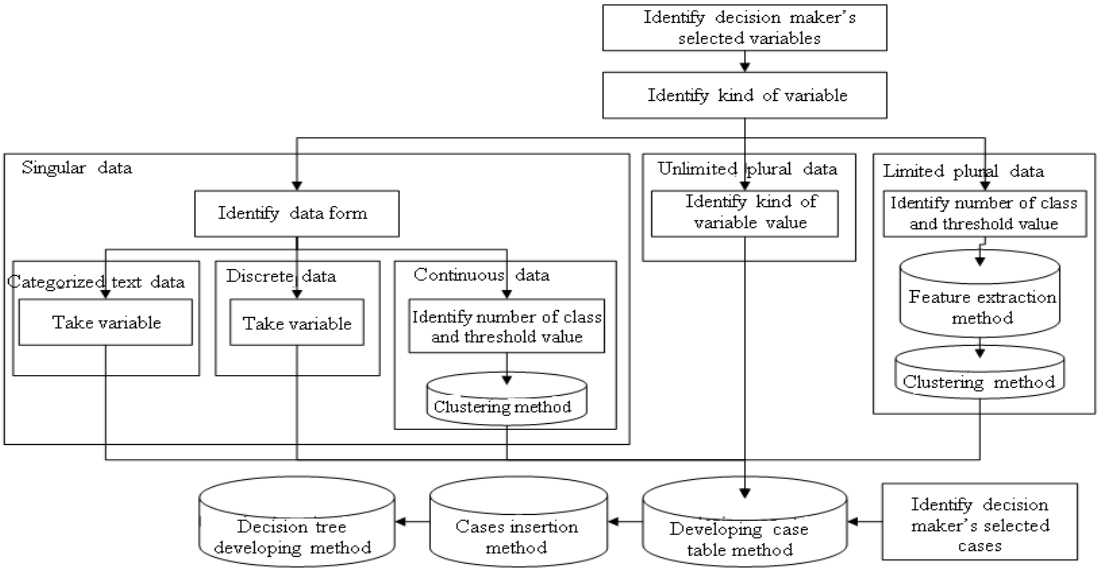

Fig. 1: CUC-C4.5 framework

They extracted music properties from music's sound wave. They could give music recommendation that is best-meet the users music needs. Meanwhile, Gobert and Hiroshi (2007) used k-means clustering to classify unlabelled MRI data.

The knowledge building process uses the C4.5 algorithm. In general, steps in C4.5 algorithm to build decision tree are (Craw, 2005):

- Choose attribute for root node
- Create branch for each value of that attribute
- Split cases according to branches
- Repeat process for each branch until all cases in the branch have the same class
- Choosing which attribute to be used as a root, is based on highest gain of each attribute, the gain is counted using Formula (1) (Quinlan, 1993; Craw, 2005):

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i) \qquad (1)$$

Where:
S = Case set
S_i = A partition of S according to the value of attribute A
n = The number of attributes A
$|S_i|$ = The number of cases in the i^{th} partition
|S| = Total number of cases in S

While the entropy is given the Formula 2 (Quinlan, 1993; Craw, 2005):

$$\text{Entropy}(X) = \sum_{i=1}^{n} -p_i \times \log_2 p_i \qquad (2)$$

Where:
X = Case set,
n = The number of partition in S
p_i = Proportion of S_i to S

In this research, we make an improvement to the classic C4.5 algorithm. With the new algorithm, a user is allowed to give a level of confidence for each case involved in knowledge building. The level of confidence value is ranged from 0-10. A 0 value means that a user does not believe the cases truthness at all, therefore the case does not give any contribution to decision building. A 10 value means that a user fully confident with the case and therefore the case has a full contribution in decision making.

By considering the degree of confidence of old cases, then the gain formula in Formula (1) should be modified. The $|S_i|$ value that previously was the number of cases in i^{th} partition, is becomes the number of case's degree of confidences in i^{th} partition (S_i). The value |S| that was the count of cases in S, becoming the number of case's degree of confidence in S.

RESULTS

The CUC-C4.5 framework has been implemented in a Group Decision Support System (GDSS) for geriatric assessment in knowledge management module.

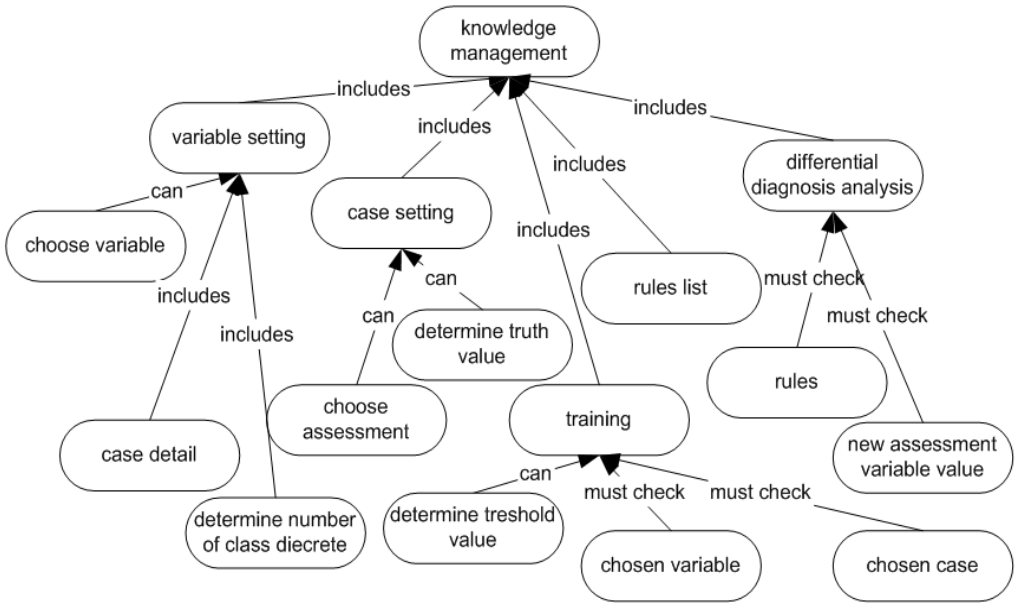

Fig. 2: Knowledge management concept map

Table 1: Test result

Test	Result
Framework model test	The knowledge generated by system is equal to that generated manually
Performance test	3.08

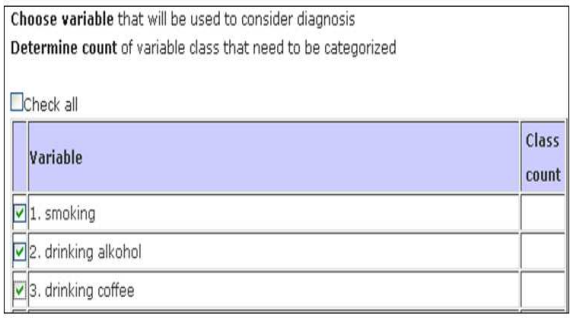

Fig. 3: Variable setting interface

This module contains facilities of setting variable, case setting, case detail, training, rule list, and diagnosis analysis. The facilities contained in the knowledge management module are illustrated as a concept map shown in Fig. 2.

The interfaces of variable setting, cases setting, rules list and differential diagnosis analysis facilities are shown in Fig. 3-6 respectively.

The test process of this framework is divided into two activities, framework model test and performance test. The process' result is shown in Table 1.

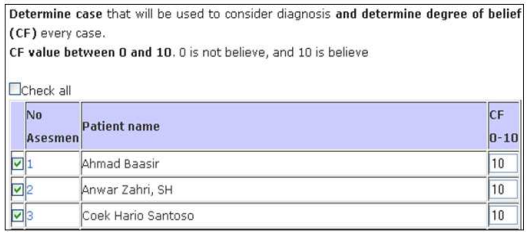

Fig. 4: Case setting interface

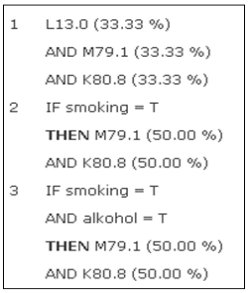

Fig. 5: Rules list interface

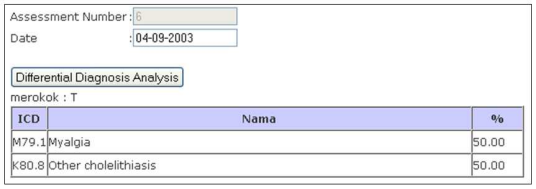

Fig. 6: Differential diagnosis analysis

DISCUSSION

Knowledge management module in GDSS for geriatric has been intended for internist, neurologist, dentist and psychiatrist, in the process of knowledge building. Variable setting facility is intended to determine the selected variables that will be used in the knowledge building process for each user. Cases setting facility is intended to determine assessments that will be used in the knowledge building process for each user. The cases detail facilitated is intended to browse the variable values in a assessment. The training facility is intended for user to trigger the knowledge building process. The rule list facility is intended to visualize the knowledge into production rule form. Differential diagnosis analysis facility is intended for user to do differential diagnosis analysis for a patient. These interfaces were built with PHP and Javascript Programming Language and MySQL DBMS.

The framework model testing was made by comparing the result of knowledge building by manual and the one that was resulted by system. Meanwhile the performance testing was made by giving questioners to user candidates after operating the built system.

The performance of knowledge building process of differential diagnosis was tested by doctors of member of geriatric assessment team in the "dr Sardjito" Hospital in Yogyakarta Indonesia. Team members that doing the knowledge building process testing were ones whose given the privilege to make diagnosis that were internist, neurologist, dentist and psychiatrist. The test was carried out by giving questioner to geriatric team members regarding to the completeness of facilities in knowledge building management sub system, information in differential diagnosis analysis, the easiness to do variable setting, to do cases setting, to do knowledge training process, to see training process result, to evaluate analysis result, to access differential diagnosis analysis result to support in determining the exact diagnosis and to evaluate analysis result and the easiness of operating, the accuracy of differential diagnosis analysis result and the speed of analysis process. The grades in a questioner were scaled from 1-4 with 1 = very bad, 2 = bad, 3 = good and 4 = very good. The test result in Table 1 showed that the knowledge building facilities had a good performance.

CONCLUSION

The CUC-C4.5 framework can be applied to build differential diagnosis knowledge with the medical record data as the cases database. This framework also allows decision maker to determine level of confidence value for each case in cases database.

Having improvement of the classic C4.5 algorithm, the CUC-C4.5 framework is capable of being implemented with the medical-record's relational database model. The framework accept many data input types (now working with categorized text, discrete number, continuous number and image) and also meets mostly medical record's type of data.

The CUC-C4.5 framework has been successfully implemented in a group decision support system for geriatric assessment.

REFERENCES

Abidi, S.S.R, 2002. Designing Adaptive Hypermedia for Internet Portals: A Personalization Strategy Featuring Case Base Reasoning with Compositional Adaptation, Springer Berlin/Heidelberg, ISBN: 978-3-540-00131-7, pp: 60-69.

Berry, M.J.A. and L.S. Gordon, 2004. Data Mining Techniques for Marketing, Sales and Customer Relationship Management. 2nd Edn., Wiley Publishing, Inc., Indianapolis, Indiana, ISBN: 0-471-47064-3, pp: 354-358.

Craw, S., 2005. Case based reasoning: Lecture 3: CBR Case-Base Indexing. http://www.comp.rgu.ac.uk/staff/smc/teaching/cm3016/Lecture-3-cbr-indexing.ppt

Darmojo, B., 2002. Penatalaksanaan penderita lanjut usia secara terpadu. http://www.tempointeraktif.com/medika/arsip/012002/sek-2.htm

Fan, J. and P. Dan Wen, 2007. Application of C4.5 algorithm in web-based learning assessment system. Proceeding of the International Conference on Machine Learning and Cybernetics, Aug. 19-22, IEEE Xplore Press, Hong Kong, pp: 4139-4143. DOI: 10.1109/ICMLC.2007.4370871

Kim, D.M., K. Kim, K.H. Park, J.H. Lee and K.M. Lee, 2007. A Music Recommendation System with a Dynamic K-means Clustering Algorithm. Proceeding of the 6th International Conference on Machine Learning and Applications, Dec. 13, IEEE Xplore Press, Cincinnati, OH., DOI: 10.1109/ICMLA.2007.97

Kohavi, R. and R. Quinlan, 1999. Decision tree discovery. DOI: 10.1.1.4.5353

Lee, G.N. and H. Fujita, 2007. K-means Clustering for Classifying Unlabelled MRI Data. Proceeding of the Digital Image Computing Techniques and Applications, Dec. 3-5, IEEE Xlpore Press, USA., pp: 92-98. DOI: 10.1109/DICTA.2007.4426781

Niknafs, A.A., M.E. Shiri and M.M. Javidi, 2005. An intelligent knowledge sharing strategy featuring item-based collaborative filtering and case based reasoning. Proceeding of the Intelligent Systems Design and Applications, Sept. 8-10, IEEE Xlpore Press, USA., pp: 67-72. DOI: 10.1109/ISDA.2005.22

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. 2nd Edn., Morgan Kaufmann Publishers, Inc., USA., ISBN: 1-55860-238-0, pp: 20-26.