Diagnosis System of Toddler Diseases Using Forward Chaining and Case-Based Reasoning

Indah Werdiningsih[1], Aisyah Shofiyyah Asma[2], Rimuljo Hendradi[3], Kartono[4], Purbandini[5], Barry Nuqoba[6], Elly Anna[7]
{indah-w@fst.unair.ac.id[1], aisyah.shofiyah-13@fst.unair.ac.id[2], rimuljo-h@fst.unair.ac.id[3]}

Study Program of Information System, Faculty of Science and Technology, Airlangga University[123456],
Study Program of Statistic, Faculty of Science and Technology, Airlangga University[7]

Abstract. Toddlers are children aged 12-36 months. This study aims to diagnose toddler diseases using forward chaining and *Case-Based Reasoning (CBR)*. There are 16 types of toddler diseases. This study consist of two steps, i.e., diagnosis using forward chaining and diagnosis using CBR. Diagnosis using forward chaining generated 18 rules. These rules were used to determine toddler diseases type, and diagnosis using CBR focused on three types of CBR calculations, i.e., Nearest Neighbour Similarity (NNS), Minkowsky Distance Similarity (MDS), and Euclidean Distance Similarity (EDS). The results of system testing using 600 data, the accuracy of diagnosis were 82% and 90%, using forward chaining and CBR respectively. Based on these results, diagnosis using CBR was better than forward chaining, because CBR justified the data that were considered wrong, to be repaired by an expert and then made them as new cases.

Keywords: case-based reasoning, forward chaining, nearest neighbour similarity, minkowsky distance similarity, euclidean distance similarity.

1 Introduction

The toddler age which ranged from 12 to 36 months was very susceptible to disease[1]. More than 400 children died every day in Indonesia; it was due to diseases that were actually easy to be prevented and treated such as pneumonia and diarrhea[2]. Toddler mortality data in Indonesia shows that toddler mortality was approximately 65 / 1,000 births[3]. This was in contrary with the statement of the World Health Organization (WHO) in 1946, which mentioned that all children had the right for health, so that they have the opportunity to become world citizens[1].

The book entitled Manajemen Terpadu Balita Sakit Berbasis Masyarakat (MTBS-M) was issued by WHO in collaboration with the Departemen Kesehatan Republik Indonesia (Depkes RI) and the Ikatan Dokter Indonesia (IDI). MTBS-M aims to improve access to sick children services in the community on areas that are difficult to be accessed by health services[4]. MTBS-M contains health assistance, toddler care at home, community training to do simple treatment for young babies and sick toddlers to reduce toddler mortality[4].

Decision support systems (DSS) are interactive computer-based systems that help decision makers by utilizing data and models to solve unstructured problems[5], [6]. Expert systems with forward chaining methods have been used in several studies including Automated Scheduling System for Thesis and Project Presentation Using Forward Chaining Method With

ICCSET 2018, October 25-26, Kudus, Indonesia

DOI 10.4108/eai.24-10-2018.2280501

Dynamic Allocation Resources[7]. Implementation of forward Chaining Method for Early Detection of Diabetes Mellitus[8], Application of Expert System for Diagnosing Infectious Diseases for Toddlers Using Forward Chaining Method[9], and expert system supporting an early prediction of the bronchol pulmonary dysplasia[10].

CBR is a problem-solving method that can be reused in similar cases to find solutions for new problems by referring to the base case[11]. The CBR method has been widely developed in the medical world. The diagnosis reasoning in the medical field using CBR method was using pattern matching type, the point was treated using case-based reasoning process based on the experience of previous patients[12]. The CBR process has four-steps, where the process made CBR easy to update its knowledge base so that it can solve complex and unstructured problems [11].

There were several studies on CBR, one of them was the CBR for Diagnosis of Heart Disease [13]. In this study, three types of classification methods were used to diagnose six types of heart disease. The classification method used was Nearest Neighbor Similarity (NNS), Minkowski Distance Similarity (MDS) and Euclidean Distance Similarity (EDS), with accuracy rates of 86.21%, 100%, and 94.83% respectively. CBR can improve the accuracy rate for diagnosing diseases.

Based on the description above, this study would apply forward chaining and CBR to diagnose toddler diseases based on DSS. This study was expected to be able to appropriately diagnose cases of toddler diseases and improve the accuracy of system classification.

2 Methodology

2.1 Forward Chaining

The following were a general description of the expert system application:

1) Input, questions that appear in the expert system application.
2) Knowledge base (knowledge domain), knowledge of diseases classification based on the MTBS used as rule-based.
3) Working memory, facts entered by the user to the expert system application.
4) Inference Engine, the process of matching the facts that exist in working memory with the knowledge domain, to conclude the problem at hand.
5) Classification of diseases, the conclusion of the expert system diagnosis process

2.2 Case Based Reasoning (CBR)

CBR is a major paradigm in reasoning that can solve new problems by paying attention to the similarity of problem-solving from previous problems.

2.3 Steps of CBR

Steps of CBR were case base determination, expert weighting, local similarity, confidence level, global similarity, and selection of the highest value.

2.3.1 Case base determination

Case base was used for storage. Each stored case was divided into several main parts, i.e., toddler identity, disease symptom, and disease diagnosis.

2.3.2 Expert weighting

The purpose of parameter weighting was to determine the influence of each parameter to other parameters [12], [13]. Parameter weighting was carried out by experts and assisted by statistical calculations.

2.3.3 Local Similarity

The local similarity was the proximity between local attributes or the same attributes[13]. The local similarity was done by defining the proximity between attribute values. Local similarity calculations were calculated based on data type, the calculation for numeric data types shown in equations (1) and (2). Local similarity formula for numeric data types:

$$f(s,t) = 1 - \frac{|s-t|}{R} \tag{1}$$

Notation:
s, t: Attribute value to be compared
R: Attribute value range

Local similarity formula for boolean data types:

$$f(s,t) = \begin{cases} 1, if\ s = t \\ 0,\ others \end{cases} \tag{2}$$

Notation:
s,t∈{true,false}

2.3.4 Confidence level

The confidence level was divided into two levels, i.e. level of expert confidence, and level of confidence in a new case to the previous case. The level of expert confidence was calculated based on the disease category that was determined by experts based on the symptoms and risk factors suffered by the patient. The confidence level in a new case to the previous case was calculated by equation (3).

$$\mu_{(T,S)} = \frac{J(S,T)}{J(S)} \tag{3}$$

Notation:

$\mu_{(T,S)}$: Confidence level between T case (new case) and S (previouscase)
J(S,T) : Number of symptoms in new cases that appear in the symptoms of the previous case.
J(S) : Number of symptoms in the base case.

2.3.5 Global Similarity

Global similarity aims to look for similarity between attributes in all existing variables. The algorithms used in the global similarity process were Nearest Neighbor Similarity, Minkowski Distance Similarity, and Euclidean Distance Similarity.

1. Nearest Neighbor Similarity:

$$SimNN(T,S) = \frac{\sum_{i=1}^{n}(f_i(S_i,T_i)(w_{i,p(s)}))}{\sum_{i=1}^{n}(w_{i,p(s)})} * P(S) * \frac{J(S_i,T_{i,})}{J(T_i)} \quad (4)$$

Notation:
Simon(S,T) : Global Similarity between T case and S (source case)
N : Number of features available
$f_i(S_i,T_i)$: The same the i[th] feature from source case and target case / function of local similarity
S_i : The i[th] featurein source case
T_i : The i[th] feature in target case
$w_{i,p(s)}$: Weighting value of i[th] feature in diseases from source case
$P(S)$: Percentage of expert confidence in a case in the source case
$J(S_i,T_{i,})$: Number of features in target case that appear in the features of source case
$J(T_i)$: The number of features was in target case

2. Minkowsky Distance Similarity:

$$SimMD(S,T) = \left[\frac{\sum_{i=1}^{n}\left(w_{i,(p(s))}\right)^3 * |f_i(S_i,T_i)|^3}{\sum_{i=1}^{n}(w_{i,p(s)})^3}\right]^{\frac{1}{3}} * P(S) * \frac{J(S_i,T_{i,})}{J(T_i)} \quad (5)$$

Notation:
SimMD(S,T) : Global Similarity between T (target case) and S (source case)
R : Minkowski factor (positive integer) (r=3)

3. Euclidean Distance Similarity:

$$SimED(S,T) = \left[\frac{\sum_{i=1}^{n}\left(w_{i,(p(s))}\right)^2 * |f_i(S_i,T_i)|^2}{\sum_{i=1}^{n}(w_{i,p(s)})^2}\right]^{\frac{1}{2}} * P(S) * \frac{J(S_i,T_{i,})}{J(T_i)} \quad (6)$$

Notation:
SimED (S,T) : Global Similarity between T (target case) and S (source case)

2.3.6 Selection of The Highest Value

Selection of the highest value was the process of selecting cases that had been stored in case base to be selected as a solution, where the data had the highest level of similarity[13].

3 Result

3.1 Data and Information Collection

Data and information collection were done by interview, literature study, and data collection. Interview was done with one of the medical staff in a hospital in Surabaya. Interview result was variables that influence toddler diseases. These variables were given weight to assess severity of the diseases. Literature study was done by studying matters related to the method used, and finding out the variables that influence the diagnosis of toddlers.

Data collection was done by reading the patient's medical record data. The number of collected data was 600 data. The data was divided into training and testing data. 450 training data and 150 testing data.

3.2 Data and Information Management

Data and information obtained were analyzed to build a system that was matching with the user needs. Data and information management consist of two steps, i.e. forward chaining and CBR.

3.2.1 Management of Forward Chaining

The results of literature study were the steps of forward chaining method. Code of complaints were K1 = cough, K2 = Diarrhea, and K3= Fever. Data and information about symptoms fact, diseases and complaints experienced by toddler were shown in Table 1 and Table 2.

1. Rule-based generation
   Based on the facts obtained, the rule was generated. Eighteen rules were generated based on diseases. The rules generated could be seen in Table 3.

Table 1. Symptoms Variable.

Code	Symptoms	Code	Symptoms
G1	Child cannot drink or suckle	G14	Diarrhea of 14 days or more
G2	Vomited	G15	Blood in the feces
G3	Seizures	G16	Fever
G4	Children appear unconscious	G17	Child can not bow down until chin reaches chest
G5	Quick breathing	G18	Rash on skin
G6	chest wall pulled inside	G19	Cough, cold or red eyes
G7	Stridor	G20	Turbidity of the cornea of the eyes
G8	liquid or mushy defecation	G21	Wounds in the mouth

G9	Hollowed eyes	G22	Purulent eyes
G10	Puncture of abdominal skin slowly returned	G23	Fever of 2 - 7 days
G11	Fussy / irritable	G24	Continuous Fever with suddenly high temperature
G12	Thirsty	G25	Diarrhea
G13	Gently drinking	G26	Cough/cold

Table 2.Diseases Table.

Code	Classification of disease	Code	Classification of disease
P1	Common risk sign	P10	Dysentery
P2	Cough	P11	Fever
P3	Pneumonia	P12	Fever with common risk sign
P4	Severe Pneumonia	P13	Measles
P5	Diarrhea	P14	Measles with severe complication
P6	Mild Dehydration Diarrhea	P15	Measles with complication
P7	Severe Dehydration Diarrhea	P16	fever may be DBD
P8	Persistent diarrhea	P17	DBD
P9	Severe persistent diarrhea	P18	Fever is not DBD

Table 3. Generated Rules

Rule	IF	THEN	Rule	IF	THEN
1	G1 OR G2 OR G3 OR G4	P1	10	P5 AND G15	P10
2	K1 AND G5	P2	11	K3 AND G16	P11
3	K1 AND G6	P3	12	P1 AND P11 OR G17	P12
4	K1 AND P1 OR G7	P4	13	P11 AND G18 AND G19 OR G21	P13
5	K2 AND G8	P5	14	P13 AND P1 AND G21 OR G20	P14
6	P5 AND G10 AND G11 OR G12 OR G13	P6	15	P13 AND G21 OR G22	P15
7	P5 AND G10 OR G12 OR G13	P7	16	P11 AND G23 AND G24	P16
8	P5 AND G14	P8	17	P11 AND G23 AND G24 OR G25 OR G11	P17
9	P8 AND P6 OR P7	P9	18	P11 AND G16 OR G26	P18

2. Application of Forward Chaining Method

   The expert system used forward chaining; if the premise (if) is true then the conclusion will also be true. Here were the search steps with forward chaining:

   Step 1: Ask questions to user

   Step 2: Receive input from user as known fact in short-term memory stored in each variable of the question asked

   Step 3: Check the rule based on facts on short-term memory using forward chaining method.

   Step 4: if rulewas found then conclusion was accommodated in short-term memory, if there is new fact then step one up to step four are repeated. If the rule was not found, give the default output.

   Step 5: provide solution

3.2.2 Data Analysis using Case Based Reasoning

Data analysis consist of 8 steps, i.e. building case representation, weighting by experts, local similarity calculation, confidence level calculation, global similarity calculation, highest value selection, revising and storing data in database.

a. Building Case Representation
   Case representation was used to identify variables needed. Variables used were weight, height, gender, age, body temperature, and 26 symptoms. Symptoms could be seen in Table 1. Variables were grouped into two kinds, i.e. generic and specific variables. Generic variables were variables that have general properties, such as body height, body weight, body temperature, age, and gender. Specific variables were variables existed only on specific disease.
b. Weighting
   Weighting was done by an expert. Weighting is a process of giving weight to each variable. Weighting was calculated using statistical analysis, started by calculating number of variables appeared in target case on training data. Weighting was done by series of iteration. Weighting calculation process using statistical analysis had two prerequisites that must be satisfied. The first was the highest global similarity value should be the comparison value between a testing data and training data matched with doctor's diagnosis data. If the highest global similarity value obtained did not match with doctor's diagnosis, weighting value of system diagnosis should be changed in order to get lower similarity value than global similarity value resulted from doctor's diagnosis. The process of weight changing was adjusted with variable similarity from testing and training data. If variable similarity value was 1 and the weight was high, then the weight must be reduced. If variable similarity value is 0, then the weight must be added. The second was doing weighting normalization by maintaining the total weight value of all disease to have the same value. In this case, total weight value of all diseases was set to 290.
c. Local Similarity Calculation
   Local similarity calculation was a process of comparing values of symptoms variables between new cases and previous cases. Local similarity calculation for numerical data type was done using equation (1). Local similarity calculation for Boolean data type was done using equation (2).
d. Confidence Value Calculation
   Process done in confidence value calculation was calculating the comparison between numbers of variables in new cases and previous cases. The result of confidence value calculation was used to decide whether the new case can continue to the next step of calculation, or must do the reverse step instead. Confidence value of all data having value < 0.75 (threshold) would be put into reverse process.
e. Global Similarity Calculation
   The algorithm used in global similarity process was Nearest Neighbor Similarity using Equation (4), Minkowski Distance Similarity using Equation (5), and Euclidean Distance Similarity using Equation (6).

f. The Highest Value Selection
The results of the previous step were then processed to get the highest value. The highest value was obtained from the results of global similarity calculation for data 1 and 2. The highest value obtained showing the diagnosis in new cases. Comparation of similarity values using global Nearest Neighbor Similarity, Minkowski Distance Similarity and Euclidean Distance Similarity in data 1 and 2 could be seen in Table 4.

Table 4. The Highest Value Selection.

No	CBR	Data 1	Data 2	Diagnosis
1	NNS	0,49	0	Mild Dehydration Diarrhea
2	MDS	0.1	0	Mild Dehydration Diarrhea
3	EDS	0.06	0	Mild Dehydration Diarrhea

g. Revision Process
The revision process was carried out only if the confidence level calculation process has a value of < 0.75. This process was carried out by asking experts for cases that the system cannot diagnose. Determining the system whether can or cannot answer the diagnosis was from the level of confidence. A level of confidence that had a value of <0.75 meant that the value did not resemble all data contained in the database.

h. Storing Data in Database
If a new case had been calculated with global similarity and received a diagnosis, the new case data will be entered into the database. The data was then considered to be old data that will be compared with other new data cases to diagnose.

3.3 System Testing

System testing was made to be able to determine the level of system accuracy. The results were tested using 600 patient data consisting of 450 training data and 150 testing data. System testing was performed on classifications with forward chaining and classification with CBR. System testing results were described in Table 5.

Table 5. Recapitulation of Test Results.

Method	Result of Correct Data
Forward Chaining	123
NNS	120
EDS	129
MDS	135

4 Discussion

Evaluation of system testing results was done by calculating accuracy, which matched the diagnosis of medical personnel with the results of system diagnosis. Diagnosis using forward chaining was obtained accuracy of 82%, CBR with NNS was obtained accuracy of 80%, EDS of 86%, and MDS of 90%. Optimal accuracy was obtained by diagnosis using CBR with global

similarity used by MDS. Diagnosis of toddlers using CBR had better accuracy than forward chaining. This was because CBR had a revision process, which justified data that was considered wrong to be repaired by experts and then made a new case[12].

5 Conclusion

Based on research and system testing results, it could be concluded that:

1. Diagnosis of toddlers using forward chaining resulted in 18 rules.
2. The accuracy obtained was 82% for forward chaining, diagnoses of toddlers using CBR obtained 80% accuracy for NNS, 86% for EDS, and 90% for MDS.The most optimal accuracy was accuracy using CBRwith global similarity used by MDS of 90%, this is because CBR has a revision process.

References

[1] N. H. Perlroth and C. W. Castelo Branco, "Current knowledge of environmental exposure in children during the sensitive developmental periods," *J. Pediatr. (Rio. J).*, vol. 93, no. 1, pp. 17–27, 2017.

[2] M. Garenne, C. Ronsmans, and H. Campbell, "The magnitude of mortality from acute respiratory infections in children under 5 years in developing countries.," *World Health Stat. Q.*, vol. 45, no. 2–3, pp. 180–191, 1992.

[3] Ministry of National Development, *Report on The Achievement of The Millennium Development Goals Indonesia 2010*. 2010.

[4] K. K. R. Indonesia, *Pedoman Penyelenggaraan Manajemen Trpadau Balita Sakit Berbasis Masyarakat (MTBS-M)*. 2014.

[5] I. Subakti, "Sistem Pendukung Keputusan Jurusan Teknik Informatika Fakultas Teknologi Informasi Institut Teknologi Sepuluh Nopember Surabaya," *J. Ilm. Teknol. Inf. Inst. Teknol. Sepuluh Nop.*, vol. 4, no. Management Support System, pp. 5–9, 2013.

[6] J. P. Jiawei Han, Micheline Kamber, *Data Mining Concepts and Techniques*. 2012.

[7] C. Fiarni, A. S. Gunawan, Ricky, H. Maharani, and H. Kurniawan, "Automated Scheduling System for Thesis and Project Presentation Using Forward Chaining Method with Dynamic Allocation Resources," *Procedia Comput. Sci.*, vol. 72, pp. 209–216, 2015.

[8] D. S. Pinurbo and E. Ariyanto, "Implementasi Metode Forward Chaining Untuk Analisa Pendeteksian Dini Penyakit Diabetes Mellitus," *Inst. Teknol. Telkom Bandung*, 2012.

[9] Y. K. P. Tjumoko, A. Sukmaaji, and J. Lemantara, "Aplikasi Sistem Pakar Diagnosis Penyakit Menular Pada Balita Dengan Metode Forward Chaining," no. 1, pp. 1–8, 2008.

[10] M. Ochab and W. Wajs, "Expert system supporting an early prediction of the bronchopulmonary dysplasia," *Comput. Biol. Med.*, vol. 69, pp. 236–244, 2016.

[11] H. Ahn and K. jae Kim, "Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach," *Appl. Soft Comput. J.*, vol. 9, no. 2, pp. 599–607, 2009.

[12] S. C. K. S. Sankar K. Pal, *Foundations of Soft Case-Based Reasoning*, vol. 7, no. 1. 1993.

[13] E. Wahyudi and S. Hartati, "Case-Based Reasoning untuk Diagnosis Penyakit Jantung," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 11, no. 1, p. 1, 2017.