A Development of Geriatric Internal Medicine Knowledge Base Utilizing Case Based Reasoning

Kusrini¹, Sri Hartati², Agus Harjoko³, Retantyo Wardoyo⁴
¹,²,³,⁴Computer Science Program, Gadjah Mada University, Indonesia,
¹STMIK AMIKOM, Yogyakarta, Indonesia
Email: kusrini@amikom.ac.id, shartati@ugm.ac.id, aharjoko@ugm.ac.id, rw@ugm.ac.id

Abstract — Elderly patient, well-known as geriatric, is handled by a specific group of experts named geriatric assessment team. One of the team members is an internist who carries out internal medicine assessment and performs the differential diagnosis. A computer based diagnosis tool is built in this research to help the internist performing the differential diagnosis. A case based reasoning technique is employed to construct the internal medicine knowledge.

The knowledge building is based on the determining variable and assessment data that is chosen by the internist. There are 3 kinds of data that can be used as the determining variables that is single data, multiple data, and image data. The single and multiple data are categorized text that can be in discrete or continuous model. The continuous data are transformed into discrete one with K-Means clustering method. A number of features are used to represent the image data. These are the first-order moment, entropy, energy, contrast, and homogeneity. The C4.5 algorithm was also utilized to build the knowledge.

Keyword: knowledge base, internal medicine, case based reasoning, k-means cluster

INTRODUCTION

Elderly patient that is well-known as geriatric is handled by a geriatric team in geriatric assessment. The geriatric team normally consists of geriatric consultant, internist, physiotherapist, neurologist, dentist, psychologist, nutritionist, social worker, and pharmacist.

The researchers built a web based application to support geriatric assessment. This system facilitated with internal medicine’s differential diagnosis analysis to support internist in the work.

In this paper, the researchers explain about the knowledge building process of internal medicine in geriatric patient. The process of knowledge building was utilizing case based reasoning technique, where knowledge was built by using assessment case data in internal medicine that happened in the past.

The knowledge was constructed in a decision tree by use of C4.5 algorithm. Target variable would be the diagnosis and the determinant variable would be medical record, physical check-up, and supporting check-up data in form of categorized-text and image.

Some data in determinant variable are discrete and some others are continuous. The continuous data are transformed into discrete one with K-Means clustering method.

The text based decision tree algorithm based on C4.5 was implemented to analyze cancellation of student candidate registration in STMIK AMIKOM Yogyakarta[2][3]. The same algorithm was used to classify image data[1].

THEORETICAL BACKGROUND

Case Based Reasoning

Case-based reasoning (CBR) is a problem solving technique based on previous experience knowledge [4].

Fig. 1. Case Based Reasoning Life Cycle [6]

The problem-solving life cycle in a CBR system consists essentially of the following four parts (see Fig. 1)[6]:
1. Retrieving similar previously experienced cases (e.g., problem–solution–outcome triples) which problem is judged to be similar
2. Reusing the cases by copying or integrating the solutions from the cases retrieved
3. Revising or adapting the solution(s) retrieved in an attempt to solve the new problem
4. Retaining the new solution once it has been confirmed or validated
C4.5 Algorithm

In general, steps in C4.5 algorithm to build decision tree are [5]:
- 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 [5].

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

Where \(\{S_1, \ldots, S_i, \ldots, S_n\}\) is partitions of \(S\) according to values of attribute \(A\), \(n\) is the number of attributes \(A\), \(|S_i|\) is the number of cases in the partition \(S_i\) and \(|S|\) is total number of cases in \(S\).

While the entropy is given the formula 2 [5]:

\[
\text{Entropy}(S) = \sum_{i=1}^{n} - p_i \times \log_2 p_i
\]

Where \(S\) is case set, \(n\) is the number of cases in the partition \(S\) and \(p_i\) is proportion of \(S_i\) to \(S\).

K-Means Cluster

K-Means is one of non-hierarchy data clustering method that tried to partition existing data to the form of one or more cluster/group. This method partitions data to the form of cluster/group so that data with same characteristic were grouped in same cluster and data with different characteristic were grouped to another cluster [9].

The purpose of this data clustering is to minimize objective function which was set in clustering process, which is generally try to minimize the variation in a cluster and maximize the variation between clusters [9].

Data clustering using this K-Means method commonly used with basic algorithm below [9]:
1. Determine cluster count.
2. Allocate data into cluster randomly
3. Count the centroid/average of data in every cluster.
4. Allocate every data to closest centroid/average
5. Back to step 3, if there is still data that change its cluster or if there is centroid value change that is higher than determined threshold value of if change of value of objective function higher than determined threshold.

ANALYSIS AND DESIGN

One of member of geriatric assessment team is internal medicine specialist doctor (internist). This member task is to do internal medicine assessment. In this assessment, an internist will check the patient condition related with internal disease that possible to suffer by the patient. The output of this assessment is comparable diagnosis of the possible disease suffered by the patient.

Assessment of internal medicine is divided to 4 sub parts, there are:
1. Sub part of medical record.
2. Sub part of physical check-up
3. Sup part of supporting data
4. Sub part of diagnosis.

In the knowledge building of comparison diagnosis of internal medicine, determinant variable comes from sub part of medical record, sub part of physical check-up, and sub part of supporting data. Meanwhile, the target variable comes from sub part of diagnosis data.

From the analysis result of internal medicine assessment document, it is known that determinant variable is divided to three kinds, those are :

1. Single Data

Single data is defined as data with only one value for each assessment. Not all the data in internal medicine assessment document will be used as analysis to determine compare diagnosis. This is because there are some data that are purposed as explanation of another data. However, they still then being inserted into sub system of internal medicine assessment as internist consideration material to determine the certain diagnosis of disease that the patient suffered. As determinant variable, it is used 165. As an example of single variable used in analysis is shown in Table 1.

<table>
<thead>
<tr>
<th>Data</th>
<th>Sub Part</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking</td>
<td>Medical History</td>
</tr>
<tr>
<td>anoreksia</td>
<td>Medical History</td>
</tr>
<tr>
<td>Insomnia</td>
<td>Medical History</td>
</tr>
<tr>
<td>stomach pain</td>
<td>Medical History</td>
</tr>
<tr>
<td>paralyzed weakness</td>
<td>Medical History</td>
</tr>
<tr>
<td>Depresi depression</td>
<td>Medical History</td>
</tr>
<tr>
<td>the rate level for feeling so sad until has feeling that no one can cheer him/her up.</td>
<td>Medical History</td>
</tr>
<tr>
<td>Lie carotid</td>
<td>Physic’s check-up</td>
</tr>
<tr>
<td>Sit carotif</td>
<td>Physic’s check-up</td>
</tr>
<tr>
<td>right eye cataract</td>
<td>Physic’s check-up</td>
</tr>
<tr>
<td>left eye cataract</td>
<td>Physic’s check-up</td>
</tr>
<tr>
<td>spinal limitation movement</td>
<td>Physic’s check-up</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>Supported data</td>
</tr>
</tbody>
</table>
2. Multiple Data

Multiple data is defined as the data that have more than one value for each assessment. Multiple data that is used in internal medicine sub assessment are shown in Table 2.

Table 2: Multiple Data in sub system of internal medicine assessment

<table>
<thead>
<tr>
<th>Data</th>
<th>Sub Bagian</th>
</tr>
</thead>
<tbody>
<tr>
<td>allergic</td>
<td>Medical history</td>
</tr>
<tr>
<td>dental examination</td>
<td>Medical history</td>
</tr>
<tr>
<td>Complaint</td>
<td>Medical history</td>
</tr>
<tr>
<td>Medical examination</td>
<td>Medical history</td>
</tr>
<tr>
<td>Other examination</td>
<td>Medical history</td>
</tr>
<tr>
<td>Operating history</td>
<td>Medical history</td>
</tr>
<tr>
<td>drugs that have been used</td>
<td>Medical history</td>
</tr>
<tr>
<td>Decubitus location</td>
<td>Physic’s check-up</td>
</tr>
<tr>
<td>Support test result</td>
<td>Support data</td>
</tr>
<tr>
<td>rontgen toracs result</td>
<td>Support data</td>
</tr>
<tr>
<td>Elektrocardigraphy result</td>
<td>Support data</td>
</tr>
</tbody>
</table>

For image variable, every chosen variable will become a number of image features that will be used in analysis. In this research, there will be used 7 image features; those are first-order moment, second-order moment, third-order moment, entropy, energy, contrast, and homogeneity. The proper use of feature for diagnosis still needs further research that is not covered here.

The formula of first-order moment, second-order moment, third-order moment, entropy, energy, contrast, and homogeneity respectively are shown in formula 3, formula 4, formula 5, formula 6, formula 7, formula 8 and formula 9[7][8].

\[
\mu_c = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} p_{ij}^c \quad \text{(3)}
\]

\[
\sigma_c = \left[ \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (p_{ij}^c - \mu_c)^2 \right]^\frac{1}{2} \quad \text{(4)}
\]

\[
\theta_c = \left[ \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (p_{ij}^c - \mu_c)^3 \right]^\frac{1}{3} \quad \text{(5)}
\]

\[
\text{Entropy} = -\sum_i \sum_{i_2} p(i_1,i_2) \log p(i_1,i_2) \quad \text{(6)}
\]

\[
\text{Energy} = \sum_i \sum_{i_2} p^2(i_1,i_2) \quad \text{(7)}
\]

\[
\text{Contrast} = \sum_i \sum_{i_2} (i_1 - i_2)^2 p(i_1,i_2) \quad \text{(8)}
\]

\[
\text{Homogeneity} = \sum_i \sum_{i_2} \frac{p(i_1,i_2)}{1 + |i_1 - i_2|} \quad \text{(9)}
\]

The notation of p in formula 3, formula 4, formula 5, formula 6, formula 7, formula 8 dan formula 9 is denoted the probability, the value is in the range of 0 to 1 that is the element value in co-occurrence matrix. Meanwhile i_1 and i_2 are denoted as nearby intensity pair in the x and y direction.

To make it easier to understand, given the example of chosen determinant variable that is shown in table 4, chosen assessment is shown in table 5, then finally the case table built is shown in table 6.

3. Image Data

Image data are a kind of supporting data. Data in internal medicine assessment that are image data are shown at Table 3.

Table 3: Image data in sub system of internal medicine assessment

<table>
<thead>
<tr>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rontgen toracs image</td>
</tr>
<tr>
<td>Elektrocardigraphy</td>
</tr>
</tbody>
</table>

From the whole determinant variables that provided by the system, not all of them have to used by the internist to analyze the possibility of internal medicine compare diagnosis of the patient. Each of the internist is given the privilege to choose which variable they going to use.

Every internist will have a case table. Each of single variable chosen by the internist, will become a column in the internist’s case table. Meanwhile, for every multiple variable can be some columns in case table depending on the contain kind of those multiple variable. As an example, multiple variable allergy has 3 kind of value in assessment table (for instance. egg, dust, cold), so when an internist chooses to activate allergy variable, columns named egg_allergy, dust_allergy, and cold_allergy will be added to the internist’s case table.

Table 4: Selected Variable Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Jenis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking</td>
<td>Singgle</td>
</tr>
<tr>
<td>Anoreksia</td>
<td>Singgle</td>
</tr>
<tr>
<td>Allergic</td>
<td>Multiple</td>
</tr>
<tr>
<td>Elektrocardigraphy</td>
<td>Image</td>
</tr>
</tbody>
</table>

Table 5: Selected Assessment Data

<table>
<thead>
<tr>
<th>Asesmen</th>
<th>Smoking</th>
<th>Anoreksia</th>
<th>Allergic</th>
<th>Elektrocardigraphy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y</td>
<td>egg, fumes</td>
<td>Image</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Y</td>
<td>fumes, cold</td>
<td>Image</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Case Table
Asmen | Smoking | Anoreksia | Allergic_Egg | Allergic_Fumes | ...  
--- | --- | --- | --- | --- | ---  
1 | Y | T | Y | Y | ...  
2 | Y | Y | Y | Y | ...  

| Allergic_Cold | Elektrokardiograph | Elektrokardiograph | Elektrokardiograph | ...  
--- | --- | --- | --- | ---  
T | moment1 value | moment2 value | moment3 value | ...  
Y | moment1 value | moment2 value | moment3 value | ...  

| Elektrokardiografi | Elektrokardiografi | Elektrokardiografi | Elektrokardiografi | ...  
--- | --- | --- | --- | ---  
entropy | energy | contrast | homogeneity | ...  
entropy value | energy value | contrast value | homogeneity value | ...  

The value of moment1, moment2, moment3, energy, entropy, contrast, and homogeneity in table 6 are obtained from extraction of feature of patient image data.

The data flow diagram level 1 from internal medicine assessment is show in figure 2.

Fig. 2. Data Flow Diagram Level 1 of Internal Medicine Assessment Sub System

The data flow diagram level 2 for process 1 (internal medicine knowledge management) is shown in figure 3, meanwhile data flow diagram level 2 for process 2 (internal medicine assessment data management) is shown in figure 4.

Fig. 3. Data Flow Diagram Level 2 of Process 1 (internal medicine knowledge management)

The building process of internal medicine knowledge is held with the steps below:

1. The internist determines the variable that will be used in knowledge construction.
2. The internist determines which cases that will be used as the references in knowledge construction.
3. The system will form case table based on the variables and cases that are chosen.
4. The system will make discrete table to discrete continuous data.
5. The system will fill the table case according to variables and cases that are chosen.
6. The system will form the knowledge based on the table case formed before.

Fig. 4. Data Flow Diagram Level 2 of Process 2 (internal medicine assessment data management)

The steps to build table case are shown below:

1. Get variable $login
2. Define the variable $nama_tabel_kasus = "internist_case" +$login
3. Drop table $nama_table_case. Create $query to get all variables chosen by the internist.
4. Check whether the result variable $query is empty.
   If not, do:
   a. if variable is single variable, then define column that will be added to table $nama_table_case equivalent with the name of that single variable itself.
   b. If the variable is multiple variable, then define the columns that will be added to table $nama_table_case as many as contain of that multiple variable.
   c. If the variable is image variable, then define the columns that will be added to table $nama_table_case as many as 7 image feature from that variable.
5. Get the result of the next $query and get back to step number 4.

   To discrete the continuous data, we used K-Means Cluster algorithm. The steps that are taken is shown below:
   1. Get the variables with also needed cluster count (n)
   2. Get highest value (h) and lowest value (l) from the variable.
   3. From $i=1$ to $n$, find the centroid cluster $i^{th}$ with formula 10. When determining the centroid, we did not use random value to make the calculation process faster.
   \[
   c_i = (i - 1)(\frac{h - l}{n}) + \frac{(h - l)}{2n} \] ................................ (10)
   4. Get all the variable values.
   5. Calculate the distance of variable value with all the centroids.
   6. Insert the variable values to closest cluster.
   7. When all the values have been entered to cluster, count the means for every cluster. If there is a cluster which have unequal centroid and means value, then change the centroid value to follow the means value and start over from step 4.

   The steps to fill the knowledge table are described below:
   1. Get all the assessment number that are chosen by the internist
   2. For each assessment, get all the variables chosen by the internist.
   3. If variable is single data and discrete, then get the variable contain from the assessment and insert it to table case.
   4. If variable is single data and continuous, get the discrete value from the variable and insert it to table case.
   5. If variable is multiple data, then get all the possibilities of data in that multiple variable. For every data, check whether the data is belonged to that assessment. If the answer is yes then fill it with “Y” and if not then fill it with “T”.
   6. If variable is image variable, then extract the image from the assessment. Insert the extraction result to table case.

   The decision tree is built by using C4.5 algorithm. The taken steps are:
   1. Choose attribute for root node. The chosen attribute for root node is one which has the highest gain. Gains themselves are calculated by using formula 1 that applied to case table.
   2. Create branch for each value of that attribute
   3. Split cases according to branches
   4. Repeat process for each branch until all cases in the branch have the same class.

RESULT AND DISCUSSION

Fig.5. Variable selection interface

The design of table case building process on internal medicine differential diagnosis knowledge building had been implemented by using web based application. The interface to chose the variable that will be used is shown in Figure 5.

Interface to insert the internal medicine assessment data had also been finished in implementation. Part of the entry process of internal medicine assessment data is shown in Figure 6.
The development process of this internal medicine knowledge building is currently in progress. When all the process will have been finished, it is expected that this application can help the internist determining geriatric differential diagnosis.

CONCLUSSION AND FURTHER WORK

This research had been succeeded to design the process of differential diagnosis knowledge building for internal medicine. The development of web based application to implement the design is currently in progress.

Testing of utilization of the knowledge built is still need to be done in order to gain the responsibility of the knowledge.

Every kind of medical images have their own characteristic so they need further research about feature that has big impact in image based diagnosis.

REFERENCES