INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

IJARnD

(Volume 4, Issue 7)
Available online at: www.ijarnd.com

CBR based kidney and urinary tract diagnosis system

Thin Thin Swe
Student, University of Computer Studies, Myitkyina, Myanmar

ABSTRACT

Nowadays, medical diagnosis reasoning is a very important application area of the computer-based systems. Case-Based Reasoning (CBR) has become a successful technique for developing medical diagnosis systems. This system presents a CBR methodology and the technical aspects of implementing a medical diagnosis system. CBR is a recent approach to problem-solving and learning that has got a lot of attention over the last few years. In the case-based reasoning system, old cases are retrieved to solve user input problems or new cases. To be a complete CBR system, case adaptation is used to revise and retain the new case for future use when no match case is found in the case base. This system is implemented as a case-based reasoning system for Kidney and Urinary Tract Diseases. In this system, the K-Nearest Neighbor classifier (KNN) is used for case retrieval. If the result of the input case is not satisfied, the adaptation process will be done. Decision Tree algorithm is used in case adaptation.

Keywords*— CBR, KNN, Decision Tree*

1. INTRODUCTION

In real-world problem solving, people usually use experience that was successful in solving previous, similar problems. In Case-Based Reasoning (CBR), experiences are modeled into a different form as concrete problems with their solutions (cases). Case-based reasoning systems were developed to help people identify previous situations that match aspects of a current problem. They were also developed to provide guidance on how to solve problems and make decisions. CBR has been used to create numerous applications in a wide range of domains including financial analysis, medical diagnosis, design, and classification of objects, help desk, and decision support system.

This system is an approach for developing case-based medical decision support systems based on the technology of Case-Based Reasoning (CBR). Case-Based Reasoning (CBR) is a method for solving problems by comparing a problem situation (a case) to previously experienced ones. When a new problem is presented to a CBR system, it first retrieves cases with similar problem descriptions from the case-base. The solutions in these retrieved cases are used to propose a solution for the new problem. It may be necessary to adapt the proposed solution to take account of differences between the new problem and the retrieved problems. In addition to returning the proposed solution as the answer to the new problem, it is common to review the new problem and its solution, and perhaps to retain this problem-solution pairs a new case in the case-base. In this system, if the user inputs the attributes of Kidney and Urinary Tract disease, the system will output the disease name of Kidney and Urinary Tract.

2. RELATED WORK

In 2011, B. Shahina and U. A. Mobyen [1] presented case-based reasoning systems in the health sciences. Medical CBR systems were reviewed in terms of their functionalities and the techniques adopted for system construction. In particular, they outlined a variety of methods and approaches that have been used for case matching and retrieval, which play a key role in these medical CBR systems. It was shown that CBR has been applied in many medical scenarios for various tasks, such as diagnosis, classification, tutoring, treatment planning, and knowledge acquisition/management.

In 2015, H. Tabatabaee and H. Fadaeiyan [2] presented the four cooperating systems that are investigated to show the methods and benefits of case-based reasoning in the medical field. Then, they investigated how these Artificial Intelligent (AI) systems do research with medical research and practice, integration of several AI and methods of computation, the impact of a small number of cases, the reason or time series data, and integration of numerical data, text data and subjective. These systems are presented care-partner, diabetes support system, recovery of hemodialysis in kidney disease and malardalen stress system.

In 2017, A. Nega and A. Kumlachew [3] presented the data mining based hybrid intelligent system for medical application. A hybrid intelligent system is a combination of artificial intelligence techniques that can be applied in healthcare to solve complex medical problems. Case-based reasoning and rule based reasoning are the two more popular AI techniques which can be easily combined. Both techniques deal with medical data and domain knowledge in diagnosing patient conditions. This system measured accuracy of 87.5% and usability of 89.2%.

3. CASE BASED REASONING (CBR)

Case Based Reasoning (CBR) is an approach to solving problems, reusing experiences from concrete problems solved in the past and gaining lessons learned for future use. CBR approach is suitable in domains for which concrete human experiences play a significant role in human problem-solving. CBR works on the basic principle that similar problems often have similar solutions.

Basically, the CBR system follows a cyclic process of four main steps; Retrieve, Reuse, Revise, and Retain. Similarity measure computes the similarity between a new case and previous cases restored in the case base. Depending on the application domain and features used for describing cases, simple or more complex measure can be applied. Solutions from past cases may not directly be reusable, in which situations they should be adapted to better fit new problem. Suggested case solution is evaluated and revised if needed. Finally, the revised case is retained to provide sustained learning [4]. Four main steps of the CBR system is shown in figure 1.

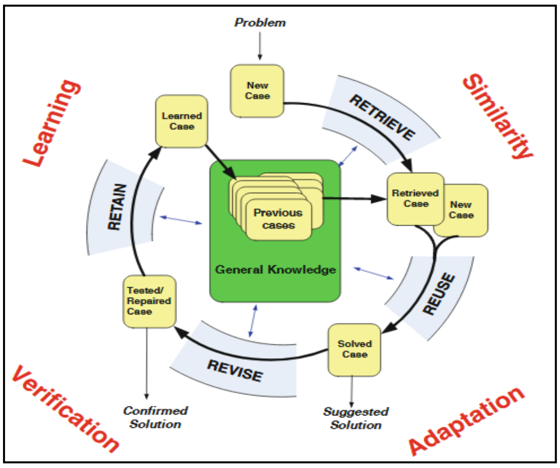

Fig. 1: Four Main Steps of CBR

4. K-NEAREST NEIGHBOR (KNN) CLASSIFIER

K-Nearest Neighbor classifiers are based on learning by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by *n* attributes. Each tuple represents a point in *n*-dimensional pattern space. When given an unknown tuple, a k-nearest neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple.

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}\left(x_i - y_i\right)^2} \quad (1)$$

These k training tuples are the k "nearest neighbors" of the unknown tuple. "Closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, $X(x_1, x_2 ... x_n)$ and $Y(y_1, y_2 ... y_n)$ is shown in equation 1 [5].

4.1 KNN Classifier Algorithm

In retrieving the case, this system uses the KNN algorithm. KNN algorithm is as follows [6]:

Algorithm: K-Nearest Neighbor
Step 1: begin
Step 2: **For** each row in D rows
Step 3: Do
Step 4: distance = calcdistance (R, row);
Step 5: RecordIfCloset (row, Index, distance);
Step 6: endo;
Step 7: **EndFor**;
Step 8: **If** isExistIdentical() **Then**
Step 9: Display identical record;
Step 10: **Else**
Step 11: Display nearest neighbors.
Step 12: **EndIf**;
Step 13: **end**.

5. DECISION TREE CLASSIFIER

A decision tree is a flow-chart-like tree structure that employs a top down. To select the test attribute at each node in the tree, the information gain measure is used. The decision tree is built by using the attribute selection measure equation and decision tree algorithm [7].

Algorithm: Generate_decision_tree.
Step 1: create a node N;
Step 2: **if** samples are all of the same class, C **then**
Step 3: return N as a leaf node labeled with the class C;
Step 4: **if** attribute-list is empty **then**
Step 5: return N as a leaf node labeled with the most common class in samples;
Step 6: select test-attribute, the attribute among attribute-list with the highest information gain;
Step 7: label node N with test-attribute;
Step 8: **for each** known value a_i **of** test-attribute
Step 9: grow a branch from node N for the condition test attribute=a_i;
Step 10: let s_i be the set of samples in samples for which test-attribute=a_i;
Step 11: **if** s_i is empty **then**
Step 12: attach a leaf labeled with the most common class in samples;
Step 13: **else** attach node returned by decision tree.

5.1 Attribute Selection Measure

Information gain measure is used to select the test feature at each node in the tree. Let S be a set consisting of s data samples. Suppose the class label attribute has m distinct values defining m distinct classes, Ci (for i=1... m). Let si be the number of samples of S in class Ci. The expected information needed to classify a given sample is given by

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log(p_i) \quad (2)$$

Where p_i is the probability that an arbitrary sample belongs to C_i and is estimated by s_i/s. Let attribute A have v distinct values, $\{a_1, a_2 \ldots a_v\}$. Attribute A can be used to partition S into v subsets, $\{S_1, S_2 \ldots S_v\}$, where S_j contains those samples in S that have value a_j of A. Let s_{ij} be the number of samples of class C_i in a subset S_j. The entropy, or expected information based on the partitioning into subsets by A, is given by

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j}, \ldots, s_{mj}}{s} I(s_{1j}, \ldots, s_{mj}) \quad (3)$$

The term $\frac{s_{1j}, \ldots, s_{mj}}{s}$ acts as the weight of the j^{th} subset and is the number of samples in the subset divided by the total number of samples in S. For a given subset S_i,

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log(p_{ij}) \quad (4)$$

Where p_{ij} is the probability that a sample in S_j belongs to class C_i. The encoding information that would be gained by branching on A is

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A) \quad (5)$$

The feature with the highest information gain is chosen as the test feature for the given set S. A node is created and labeled with the feature, branches are created for each value of the feature, and the samples are partitioned accordingly [7].

6. PROPOSED SYSTEM DESIGN

The proposed system is implemented for testing kidney and urinary tract diseases by using Case-Based Reasoning (CBR). When user inputs new case to the system, these old cases are retrieved from the database and reused for the process of finding the most similar cases. This process is done by using case retrieving of K-Nearest Neighbor Algorithm. The system finds the old cases from the database that are similar to the new case. Similar case finding is accomplished by matching the attribute values of the new case with each of the existing cases.

If the system can retrieve the exact match case from the database, it will show the result solution to the user. If the system not found any match case, it can retrieve the solution and retain the modified solution as a case into the database by using a decision tree algorithm. This solution can be reused for future problem-solving. Finally, this system calculates accuracy for a new case. Proposed system design is shown in figure 2.

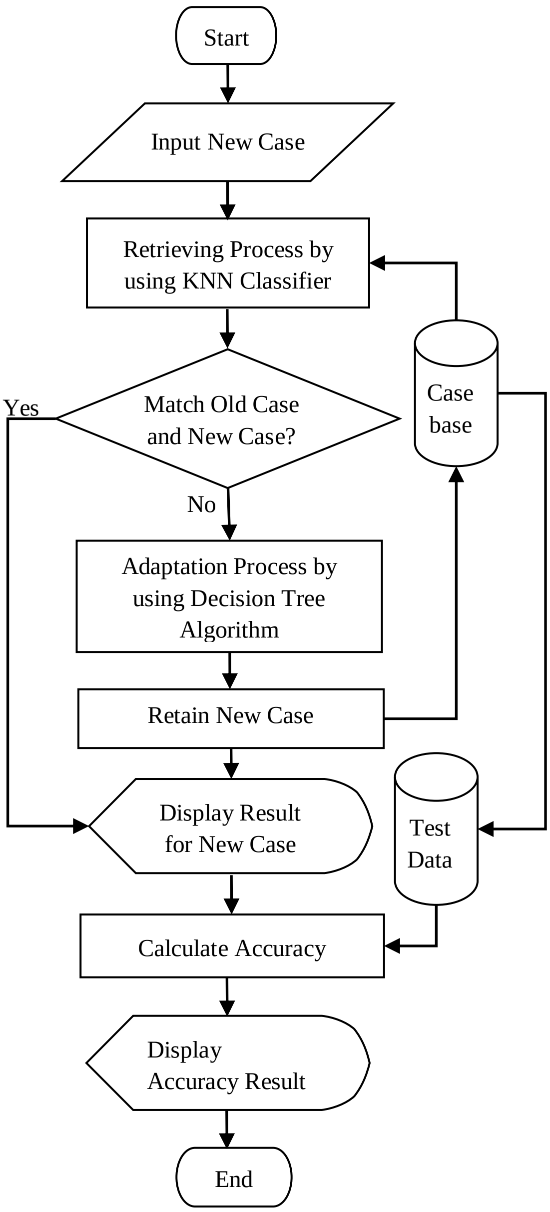

Fig. 2: Proposed System Design

6.1 Explanation of the System

User can get the required information by identifying relevant ones that have already asked by other users and are stored in the system's case base. Table 1 shows Kidney and Urinary Tract Diseases in the training dataset.

Table 1: Kidney and Urinary Tract Disease

ID	Age	Sex	Diabetes	Hypertension	Blood Pressure	Hemoglobin	Calcium	Phosphate	Potassium	Bone Pain	serum creatinine	RecentTX	Class
1	30	1	1	1	140	0	60	80	40	1	300	3	Acute
2	35	1	1	1	180	0	40	70	65	1	280	3	Acute
3	40	1	1	0	90	1	40	50	30	0	280	2	Chronic
4	42	0	1	1	160	0	48	70	65	1	490	3	Glomerular
5	30	0	0	0	150	1	60	55	60	0	300	2	Chronic
6	44	0	1	0	140	0	30	70	40	0	280	3	Chronic
7	50	1	0	0	100	1	40	60	35	0	270	0	RenalVascular
8	48	0	1	1	150	0	55	70	65	1	500	3	TubuloInterstitial
9	44	0	0	1	150	1	60	55	60	0	350	2	Acute
10	35	1	1	1	180	0	30	80	65	0	350	3	Chronic

At the first of the system, the user must input all attributes about kidney and urinary tract disease to produce the symptoms of this disease. Table 2 shows the unknown case of kidney and urinary tract disease.

Table 2: Unknown Case

Age	Sex	Diabetes	Hypertension	Blood Pressure	Hemoglobin	Calcium	Phosphate	Potassium	Bone Pain	SerumCreatinine	RecentTX	Class
30	1	1	1	200	0	60	45	63	1	350	3	?

Table 3: 6-Nearest Neighbor of New Case

ID	Age	Sex	Diabetes	Hypertension	Blood Pressure	Hemoglobin	Calcium	Phosphate	Potasium	Bone Pain	SerumCreatinine	RecentTX	Class
1	30	1	1	1	140	0	60	80	40	1	300	3	Acute
2	35	1	1	1	180	0	40	70	65	1	280	3	Acute
5	30	0	0	0	150	1	60	55	60	0	300	2	Chronic
6	44	0	1	0	140	0	30	70	40	0	280	3	Chronic
9	44	0	0	1	150	1	60	55	60	0	350	2	Acute
10	35	1	1	1	180	0	30	80	65	0	350	3	Chronic

After inputting new data to the system, the user needs to enter the value of K. After that the system will process the KNN classifier. In this phase, the system can retrieve either the exact match with the new case or the nearest match which has less distance with the new case. If the exact match is found, the system will give the exact case solution to the user directly. Otherwise, the system will retrieve the general solutions according to value K. This system calculates the Euclidean distances between the new case and sample cases. If the user enters the value of k=6, the system will show the 6 nearest neighbors of the new case. Table 3 shows the 6-Nearest Neighbor of the New Case. This system continues to classify the unknown case with decision tree algorithm while there is no identical previous case by the KNN algorithm. A decision tree is shown in figure 3.

The system generates decision rule from a decision tree. Decision rules for the unknown case are as follows:

(a) **IF** Calcium = 60 **AND** Hypertension = 1 **THEN** Result = Acute

(b) **ELSE IF** Calcium = 60 **AND** Hypertension = 0 **THEN** Result = Chronic

(c) **ELSE IF** Calcium = 40 **THEN** Result = Acute
(d) **ELSE IF** Calcium = 30 **THEN** Result = Chronic

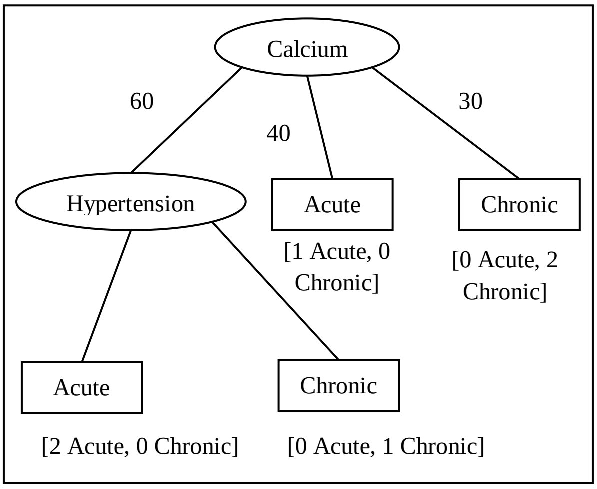

Fig. 3: Decision tree

After generating decision rules, the system can retrieve the solution for the unknown case and retain the modified solution as a case for future problem-solving. Table 4 shows the result for the unknown case.

Table 4: Result for unknown case

Age	Sex	Diabetes	Hypertension	Blood Pressure	Hemoglobin	Calcium	Phosphate	Potasium	Bone Pain	SerumCreatinine	RecentTX	Class
30	1	1	1	200	0	60	45	63	1	350	3	Acute

6.2 Experimental Result of the System

This system is tested by using about 450 total kidney and urinary tract disease cases. They are divided into 300 training cases and 150 testing cases. In this system, the user can test the accuracy of the system by using different k-nearest neighbor size (k-value) of the unknown case. This system used k= 35 in test 1, k= 42 in test 2, k= 50 in test 3 and k= 31 in test 4 respectively. Experimental results are shown in figure 4.

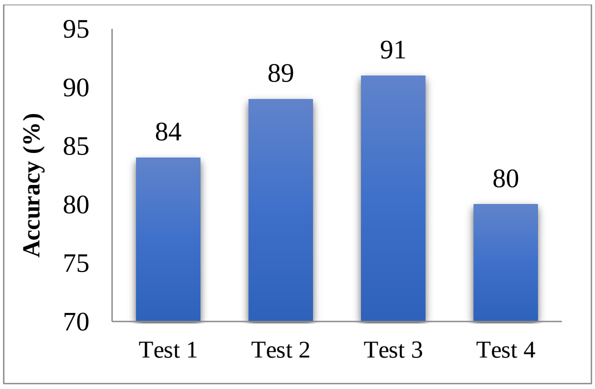

Fig. 4: Result of the System

7. CONCLUSION

The proposed system supports a medical diagnosis by using the symptoms of Kidney and Urinary Tract that the patients suffer. The system helps medical staff to get the correct diagnosis in time in emergency cases and prevent delay in the commencement of medical treatment. CBR is a suitable methodology for most of the medical domains and tasks for the following reasons; cognitive adequateness, explicit experience, the duality of subject knowledge and system integration. The development of Kidney and Urinary Tract disease case-based system will directly benefit to the users and physicians. By using this system, the physicians can check easily the patient's detailed Kidney and Urinary Tract disease information and can easily know which stage of Kidney and Urinary Tract disease.

8. REFERENCES

[1] B. Shahina and U. A. Mobyen, "Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments", IEEE, vol. 41, no. 4, pp. 421-434, 2011.
[2] H. Tabatabaee and H. Fadaeiyan, "Using Case-Based Reasoning for Diagnosis in Medical Field", Bulletin of Environment Pharmacology and Life Sciences, vol. 4, no. 11, pp. 102-114, 2015.
[3] A. Nega and A. Kumlachew, "Data Mining based Hybrid Intelligent System for Medical Application", I. J. Information Engineering and Electronic Business, vol. 4, pp. 38-46, 2017.
[4] S. V. Shokouhi and P. Skalle, "An Overview of Case-Based Reasoning Applications in Drilling Engineering", Springer, 2012.
[5] Y. F. Safri and R. Arifudin, "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor", Scientific Journal of Informatics, vol. 5, no. 1, pp. 9-18, 2018.
[6] M. A. Jabbar, "Prediction of Heart Disease using K-Nearest Neighbor and Particle Swarm Optimization", Biomedical Research, vol. 28, no. 9, pp. 4154-4158, 2017.
[7] G. Bhumika, R. Aditya, and J.Akshay, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining", International Journal of Computer Applications, vol. 163, no. 8, pp. 15-19, 2017.

