Journal of
Information Systems Engineering
and Business Intelligence

Vol. 4, No. 1, April 2018
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Hepatitis Diagnosis Using Case-Based Reasoning with Gradient Descent as Feature Weighting Method

Yufika Sari Bagi[1)]*, Suprapto[2)]

[1)]*STMIK Multicom Bolaang Mongondow*
Sutoyo Street No.12,Kotamobagu,Indonesia
[1)]yufika@stmikmulticom.ac.id

[2)]*Department of Computer Sciences and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada*
Yogyakarta,Indonesia
[2)] sprapto@ugm.ac.id

Article history:

Received 2 February 2018
Revised 3 April 2018
Accepted 6 April 2018
Available online 28 April 2018

Keywords:

Case-Based Reasoning
Gradient Descent
Feature Weighting
Hepatitis Diagnosis

Abstract

Retrieval is one of the stages in case-based reasoning system which find a solution to new problem or case by measuring the similarity between the new case and old cases in the case base. Some of the similarity measurement techniques are involving feature weights that show the importance of the feature in a case. Feature weights can be obtained from a domain expert or by using a feature weighting method either locally or globally. Gradient descent is the feature weighting method which computes global weights for each feature. This research implemented gradient descent to obtain feature weights in case-based reasoning for hepatitis diagnosis and the similarity measurement using weighted Euclidean distance. There were four variations of case base size and test data used in the research; i.e., 50% of case base and 50% of test data, 60% of case base and 40% of test data, 70% of case base and 30% of test data, and 80% of case base and 20% of test data were variation 1, 2, 3, and 4 respectively. In addition, each variation used four scenarios based on the way how to mark the test data. In scenario 1, 2, 3 and 4, the test data were respectively marked at the end, the beginning, the half of beginning and half of end, and the middle. The result showed that the accuracy of the system reaches 100% at scenario 1 in variation 4. Overall of all four variations and four kinds of scenario, the average accuracy of the system was 77.55%, average recall of system was 69.74%, and the average of precision was 78.39%. In addition, the level of accuracy was also influenced by the number of case base and the scenario of case selection for the case base. This is because more cases in the case base, the bigger chance for the system to find similar cases.

I. Introduction

Case-Based Reasoning (CBR) is a problem-solving method using old experiences with the specific way. Case base is old experiences of problems that have solutions. Every case in case base consists of problem and solution [1][2] [3].

The retrieval process or process of finding old cases that have similarity to a new case is one of the most important processes in CBR system [4][5]. Some of the similarity measurement techniques are involving feature weights. The feature weight can provide information about the importance level of the feature in a case. Thus, the weight of case features is very important in the similarity calculation that involves feature weight. The feature

[*] Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.4.1.25-31

weight can be assigned by an expert or by computation - using a weighting method. The feature weights given by an expert will depend on the experience of the expert [3].

Feature weights can be assigned automatic using a feature weighting method and this way will very useful in a domain that has many features, where it is almost impossible to assign manually by an expert. This way also will very helpful when no one exactly knowledge about feature weight [4]. There is much research about hepatitis diagnosis, one of them is hepatitis diagnosis using hybrid CBR and Particle Swarm Optimization (PSO) used UCI dataset [5][6]. In this research, every case has 19 number of feature. The case that has many features will be quite difficult to assign feature weight [4]. Some feature weighting methods assume that weights are different among local areas of instance space while most methods learn weight settings globally [6].

One of feature weighting method that assigns weight globally is Gradient Descent (GD) method, where this weighting method is significantly superior to other models [2]. Gradient Descent is an unsupervised method that has an advantage that performs the task of feature weighting without clustering the feature space explicitly and does not need to know the number of clusters present in the feature space [2]. Gradient descent method was used in a research about medical data classification using a combination of CBR and Fuzzy Decision Tree (FDT) [7]. The weight of the feature was used in the second stage that is classification cases into several groups using a method of measuring the distance called weighted distance metric. The results of the CB-FDT performance test showed that average accuracy was 99.5% in breast cancer and 85% in liver disease. There is a research about CBR for diagnosis of hepatitis disease using medical record data of hepatitis patients [8]. In this research, feature weight assigned by expert and accuracy of the system was 94.29%.

In this research, gradient descent was used to obtain feature weight globally. Gradient descent technique has been used to optimize a model to solve the problem of feature selection, called grafting [9]. Similarity measurement in this research was using weighted Euclidean distance method. Weighted Euclidean distance has been used in research about turbine diagnose and accuracy was 90% [10]. This method also has been implemented in similarity measure of CBR system and the accuracy was 94.83% [11].

II. Methods

A. Design of Feature Weighting Using Gradient Descent

The process to calculate feature weight using gradient descent method is using feature weight learning approach to minimize feature evaluation index function using equation (1) [12]).

$$E(w) = \frac{2}{N*(N-1)}\sum_{p}\sum_{q(q\neq p)}\frac{1}{2}\left[Sim_{pq}^{(w)}(1-Sim_{pq}^{(1)})+Sim_{pq}^{(1)}(1-Sim_{pq}^{(w)})\right] \quad (1)$$

Where N is the number of the case in case base and $Sim_{pq}^{(w)}$ is similarity value of a pair cases $e_p$ande_q, with involving trained weights (w) that compute using equation (2).

$$Sim_{pq}^{(w)} = \frac{1}{1+\alpha d_{pq}^{(w)}} \quad (2)$$

Where α is positive constant and $d_{pq}^{(w)}$ is weighted Euclidean distance between case p and q that calculated using equation (3).

$$d_{pq}^{(w)} = \sqrt{\sum_{j=1}^{n} w_j^2 X_j^2(x_{pj}, x_{qj})} \quad (3)$$

Where w is feature weights, *wj*is weight of*j*th feature and *Xj* is distance between*j*th feature of case p and q.

To extract feature weight, the feature evaluation index should be made as minimum using gradient descent by updating a value of w_j (denoted byΔw_j) using equation (4).

$$\Delta w_j = -\lambda\frac{\partial E}{\partial w_j} \quad (4)$$

For j=1,2, …, n or number of feature, and λ is learning rate. The training algorithm to compute feature weight using gradient descent method is described as follows [4]:

1. Input parameter α and *learning rate* (λ). This research use α=0.06 and λ=0.05.
2. Initialize weight with random values in [0,1]
3. For each j, compute Δwj using equation (4).
4. For each j, update weight wj with wj+ Δwj if wj+ Δwj in [0,1]

Compute feature evaluation index (*E*) using equation (1) and check the stop conditions. The stop conditions in this research are if no one features weight was updated and if the value of *E* is increased.

TABLE 1
ATTRIBUTES DESCRIPTION

No.	Attributes	Factor	Value
1	Age	General	Age of patient
2	Sex	General	1 = Male, 2 = Female
3	Cough	Symptom	1 = Yes, 0 = No
4	Productive cough	Symptom	1 = Yes, 0 = No
5	Itchy in body	Symptom	1 = Yes, 0 = No
6	Myalgia	Symptom	1 = Yes, 0 = No
7	Body pain	Symptom	1 = Yes, 0 = No
8	Fatigue	Symptom	1 = Yes, 0 = No
9	Dysentry	Symptom	1 = Yes, 0 = No
10	Dysuria	Symptom	1 = Yes, 0 = No
11	Fatty liver	Symptom	1 = Yes, 0 = No
12	Fever	Symptom	1 = Yes, 0 = No
13	Diarrhea	Symptom	1 = Yes, 0 = No
14	Itchy in foot and hand	Symptom	1 = Yes, 0 = No
15	Icterus	Symptom	1 = Yes, 0 = No
16	Swollen foot	Symptom	1 = Yes, 0 = No
17	Tarsalgia	Symptom	1 = Yes, 0 = No
18	Dizzy	Symptom	1 = Yes, 0 = No
19	Diaphoresis	Symptom	1 = Yes, 0 = No
20	Spasmodic torticollis	Symptom	1 = Yes, 0 = No
21	Nausea	Symptom	1 = Yes, 0 = No
22	Vomit	Symptom	1 = Yes, 0 = No
23	Dysgeusia	Symptom	1 = Yes, 0 = No
24	Decreased appetite	Symptom	1 = Yes, 0 = No
25	Lost appetite	Symptom	1 = Yes, 0 = No
26	Tarsalgia in right foot	Symptom	1 = Yes, 0 = No
27	Tarsalgia in left foot	Symptom	1 = Yes, 0 = No
28	Abdomen pain	Symptom	1 = Yes, 0 = No
29	Gastralgia	Symptom	1 = Yes, 0 = No
30	Low back pain	Symptom	1 = Yes, 0 = No
31	Low back pain (right)	Symptom	1 = Yes, 0 = No
32	Low back pain (left)	Symptom	1 = Yes, 0 = No
33	Backache	Symptom	1 = Yes, 0 = No
34	Arthralgia	Symptom	1 = Yes, 0 = No
35	Flatulence	Symptom	1 = Yes, 0 = No
36	Pharyngitis	Symptom	1 = Yes, 0 = No
37	Headache	Symptom	1 = Yes, 0 = No
38	Sore throat	Symptom	1 = Yes, 0 = No
39	Sclera icterus	Symptom	1 = Yes, 0 = No
40	Dyspnea	Symptom	1 = Yes, 0 = No
41	Swollen hand	Symptom	1 = Yes, 0 = No
42	Urine is yellow	Symptom	1 = Yes, 0 = No
43	Urine is brown	Symptom	1 = Yes, 0 = No
44	Allergy	Risk	1 = Yes, 0 = No
45	Smoke	Risk	1 = Yes, 0 = No
46	Alcohol	Risk	1 = Yes, 0 = No
47	Drugs	Risk	1 = Yes, 0 = No
48	Past medical history	Risk	1 = Yes, 0 = No
49	Family medical history	Risk	1 = Yes, 0 = No

B. Data and Testing

This research using medical record of hepatitis patient from previous research [8]. Numbers of cases are 117, where each case has 52 attributes. The first attribute is the id of a case, the second attribute is the age of the patient, and the third attribute is sex of the patient. The fourth to fifth attribute is symptom factors and the forty-sixth to fifty-first attribute are risk factors of a patient. The fifty-second attribute is hepatitis disease type. In this research the problem feature for each case are age, sex, symptom factors and risk factors, so that number of feature problem for each case are 50 features, while hepatitis disease type that uses in this research are hepatitis A, B, and C. There are two data types those are case base and test data, where case base used on weighting process. Overall of attributes are shown in Table 1.

The attribute of symptom and risk factor will have value 1 if the patient has the symptom or risk factor, otherwise, the attribute will have value 0. Examples of data are shown in Table 2.

TABLE 2
DATA EXAMPLE

Case ID.	Age	Sex	Symptom Factors	Risk Factors	Hepatitis Type
K001	19	Female	Fatigue, fever, dizzy, Nausea, Vomit, Lost appetite, abdomen pain, sclera icterus, urine is yellow	allergy, drugs, past medical history	Hepatitis A
K002	20	Male	Fever, icterus, dizzy, nausea, vomit, decreased appetite, abdomen pain, gastralgia, urine is brown		Hepatitis A
K003	50	Female	Fatigue, fever, dizzy, nausea, vomit, decreased appetite, gastralgia, sclera icterus, urine is yellow	Smoke	Hepatitis B
K005	23	Female	Itchy in body, fatigue, fever, tarsalgia, dizzy, nusea, vomit, gastralgia, arthralgia, urine is yellow	Past medical history	Hepatitis B
K006	73	Female	fatigue, diarrhea, itchy in foot and hand, tarsalgia, dizzy, nusea, vomit, lost appetite, abdomen pain, flatulence, sore throat, urine is yellow	allergy, past medical history, family medical history	Hepatitis C

The testing process is done with 4 variations number of case base and test data those are: first variation using 50% of data as case base and 50% as test data second variation using 60% of data as case base and 40% as test data, third variation using 70% of data as case base and 30% as test data and fourth variation using 80% of data as case base and 20% as test data. For each variation, using 4 kinds of scenario to mark the test data those are in first scenario the test data mark at the end of data, in second scenario the test data mark at the begin of data, in third scenario the test data mark half at the begin and half at the end of data and in the fourth scenario the test data mark in the middle of data. Ten alpha values α = {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1} were used in similarity measurement process for each scenario. Similarity measurement using weighted Euclidean distance with equation (2). After the test, the accuracy, recall, and precision will be measured using equation (5), (6) and (7) respectively [13].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \text{x } 100 \tag{5}$$

$$Recall = \frac{TP}{TP + FN}\ x\ 100 \tag{6}$$

$$Precision = \frac{TP}{TP + FP}\ x\ 100 \tag{7}$$

Where :

TP	: Number of correct diagnosis result for the positive data test
TN	: Number of correct diagnosis result for the negative data test
FP	: Number of incorrect diagnosis result for the positive data test
FN	: Number of incorrect diagnosis result for the negative data test

III. RESULTS

The result of accuracy, recall and precision measurement of CBR diagnosis using gradient descent (GD) are shown in Table 3, Table 4, and Table 5 respectively.

TABLE 3
THE RESULT OF ACCURACY MEASUREMENT

Variations	**Accuracy (in %)**				
	Scenario				**Average**
	1	**2**	**3**	**4**	
1	72.41	74.48	64.48	66.90	69.57
2	78.72	81.28	68.30	65.96	73.56
3	94.29	80	71.43	68	78.43
4	100	86.36	95.45	72.73	88.64

In Table 3, the accuracy of the system using GD was reached 100% where use variation 4 and scenario 1. The average accuracy using GD weights overall can calculate as follows:

$$\text{Average accuracy} = \frac{69.57\% + 73.56\% + 78.43\% + 88.64\%}{4} = 77.55\%$$

Next, the recall measurement is shown in Table 4.

TABLE 4
THE RESULT OF RECALL MEASUREMENT

Variations	**Average recall (in %)**			
	Hepatitis types			**Overall**
	A	**B**	**C**	
1	90.30	42.69	41.67	58.22
2	89.52	57.28	50	65.60
3	89.75	65.63	60.71	72.03
4	98.08	70	81.25	83.11

The average recall CBR system using GD weights overall can calculate as follows:

$$\text{Average accuracy} = \frac{58.22\% + 65.60\% + 72.03\% + 83.11\%}{4} = 69.74\%$$

Next, the precision measurement is shown in Table 5.

TABLE 5
THE RESULT OF PRECISION MEASUREMENT

Variations	**Average precision (in %)**			
	Hepatitis types			**Overall**
	A	**B**	**C**	
1	70.68	68.53	66.72	68.64
2	75.51	67.40	82	74.95
3	79.64	76.79	83.34	79.92
4	87.70	82.50	100	90.07

The average precision using GD weights overall was 78.39%.

IV. DISCUSSION

According to the result of accurate measurement in Table 3, then the accuracy graph of CBR using gradient descent in every variation shown in Figure 1.

By the graph, in Figure 1 we can see that the accuracy was raised while a number of case base increases and the highest average accuracy were in variation four. From Table 3, we can also see that the accuracy of CBR system using GD was reaching 100% where use variation 4 and scenario 1, wherein this variation using 80% of data as case base and 20% as test data. According to Table 4, the graph of recall was made and shown in Figure 2.

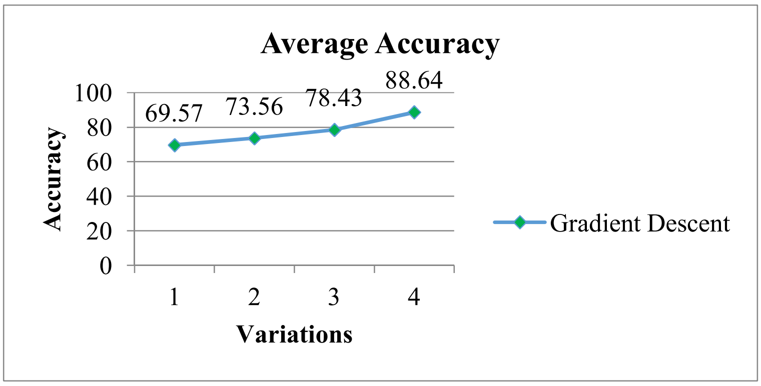

Figure.1 Average Accuracy Graph of CBR Using Gradient Descent

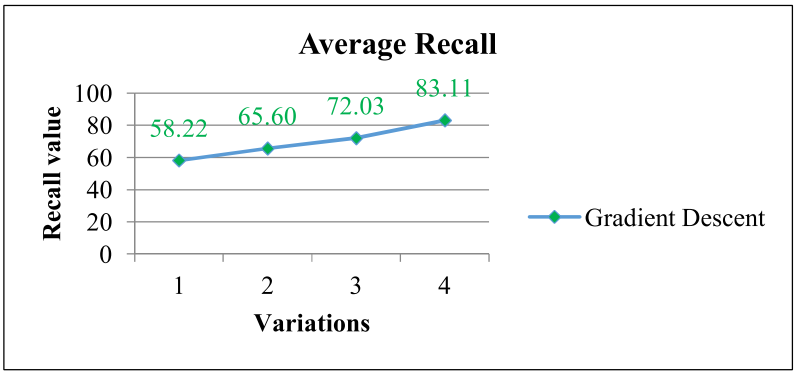

Figure.2 Average Recall Graph of CBR Using Gradient Descent

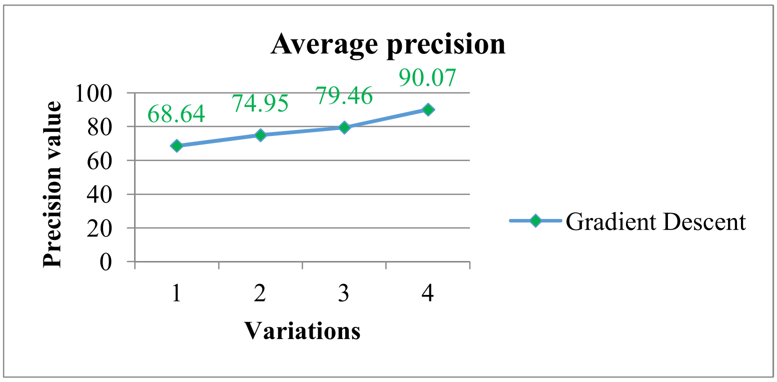

Figure.3 Average Precision of CBR System Using Gradient Descent

By the graph, in Figure 3 we can see that the precision was also raised like accuracy and recall while the number of case base increases.

V. Conclusions

Based on the experimentations, the number of cases in the case base (or case base size) influenced the accuracy of the system (the average accuracy at variation 4 was the highest). The scenario of case selection for case base also influences the accuracy. Based on the test result, the accuracy of the system reaches 100% at scenario 1 in variation 4. Overall of all four variations and four kinds of scenario, the average accuracy of the system was 77.55%, average recall of system was 69.74%, and the average of precision was 78.39%. In addition, the level of accuracy was also influenced by the number of case base and the scenario of case selection for the case base. This is because more cases in the case base, the chances of a system to finding similar cases will be more.

From the result of this research, the suggestion needs to do a research for the different domain using the weighting methods (i.e. gradient descent) to find out if domain will affect the accuracy of a system using both methods. Also, need to compare gradient descent with others methods to see the difference.

REFERENCES

[1] M. M. Richter and R. O. Weber, *Case-Based Reasoning (A Textbook)*. New York: Springer, 2013.

[2] S.M.F.D.Syed Mustapha, "Case-based reasoning for identifying knowledge leader within online community," *Expert Systems with Applications*, vol. 97, pp. 244-252, May 2018.

[3] Hassan Y.A.Abutair and Abdel fettah Belghith, "Using Case-Based Reasoning for Phishing Detection," *Procedia Computer Science*, vol. 109, pp. 281-288, 2017.

[4] S. K. Pal and S. C. K. Shiu, *Foundations of Soft Case-Based Reasoning*. New Jersey: John Wiley & Sons, Inc, 2004.

[5] Yan Aijun, Yu Hang, and Wang Dianhui, "Case-based reasoning classifier based on learning pseudo metric retrieval," *Expert Systems with Applications*, vol. 89, pp. 91-98, December 2017.

[6] Amalia Utamima and Angelia Melani Andrian, "Penyelesaian Masalah Penempatan Fasilitas dengan Algoritma Estimasi Distribusi dan Particle Swarm Optimization," *Journal of Information Systems Engineering and Business Intelligence*, vol. 2, no. 1, pp. 11-16, 2016.

[7] Lu Ling and Li Bofeng, "Combining Different Feature Weighting Methods for Case Based Reasoning," Malardalen University, Vasteras, Swedia, One Year Master Program 2014.

[8] Debarun Kar, Sutanu Chakraborti, and Balaraman Ravindran, "Feature Weighting and Confidence Based Prediction for Case Based Reasoning Systems," in *Proceeding of 20th International Conference, ICCBR 2012*, Lyon, 2012, pp. 211-225.

[9] M. Neshat, M. Sargolzaei, A. N. Toosi, and A. Masoumi, "Hepatitis Disease Diagnosis Using Hybrid Case Based Reasoning and Particle Swarm Optimization," *International Scholarly Research Network (ISRN)*, vol. 2012, 2012.

[10] Dietrich Wettschereck, David W. Aha, and Takao Mohri, "A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms," *Kluwer Academic Publishers*, vol. 11, pp. 273-314, 1997.

[11] Chin-Yuan Fan, Pei-Chann Chang, Jyun-Jie Lin, and J. C. Hsieh, "A Hybrid Model Combining Case-Based Reasoning and Fuzzy Decision Tree for Medical Data Classification," *Applied Soft Computing*, vol. 11, pp. 632-644, Desember 2009.

[12] Miswar Papuangan, Retantyo Wardoyo, and Aina Musdholifah, "Penerapan Case-Based Reasoning untuk Diagnosis Penyakit Hepatitis," Universitas Gadjah Mada, Yogyakarta, Tesis 2016.

[13] Simon Perkins, Kevin Lacker, and James Theiler, "Grafting : Fast, Incremental Feature Selection by Gradient Descent in Function Space," *Journal of Machine Learning Research*, vol. 3, pp. 1333-1356, 2003.

[14] Aisha Yousuf and William Cheetham, "Case-Based Reasoning for Turbine Trip Diagnostics," in *Proceeding of 20th International Conference, ICCBR 2012*, Lyon, 2012, pp. 458-468.

[15] Eka Wahyudi and Sri Hartati, "Case-Based Reasoning untuk Diagnosis Penyakit Jantung," *IJCCS*, vol. 11, pp. 1-10, January 2017.

[16] S. K. Pal, R. K. De, and J. Basak, "Unsupervised Feature Evaluation: A Neuro-Fuzzy Approach," *IEEE Transactions on Neural Networks*, vol. 11, pp. 366-376, 2000.

[17] I. H. Witten and E. Frank, *Data Mining : Practical Machine Learning Tools and Techniques*. San Fransisco: Morgan Kaufman, 2005.