khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Case-Base Reasoning (CBR) and Density Based Spatial Clustering Application with Noise (DBSCAN)-based Indexing in Medical Expert Systems

Herdiesel Santoso[1], Aina Musdholifah[2]
[1]Information Systems Study Program
Sekolah Tinggi Manajemen Informatika dan Komputer El Rahma
Yogyakarta
herdiesel.santoso@stmikelrahma.ac.id
[2]Department of Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta

Abstract-Case-based Reasoning (CBR) has been widely applied in the medical expert systems. CBR has computational time constraints if there are too many old cases on the case base. Cluster analysis can be used as an indexing method to speed up searching in the case retrieval process. This paper propose retrieval method using Density Based Spatial Clustering Application with Noise (DBSCAN) for indexing and cosine similarity for the relevant cluster searching process. Three medical test data, that are malnutrition disease data, heart disease data and thyroid disease data, are used to measure the performance of the proposed method. Comparative tests conducted between DBSCAN and Self-organizing maps (SOM) for the indexing method, as well as between Manhattan distance similarity, Euclidean distance similarity and Minkowski distance similarity for calculating the similarity of cases. The result of testing on malnutrition and heart disease data shows that CBR with cluster-indexing has better accuracy and shorter processing time than non-indexing CBR. In the case of thyroid disease, CBR with cluster-indexing has a better average retrieval time, but the accuracy of non-indexing CBR is better than cluster indexing CBR. Compared to SOM algorithm, DBSCAN algorithm produces better accuracy and faster process to perform clustering and retrieval. Meanwhile, of the three methods of similarity, the Minkowski distance method produces the highest accuracy at the threshold ≥ 90.

Keywords: case-base reasoning; clustering; dbscan; indexing; som.

1. Introduction

Expert system is a part of artificial intelligence that has been developed widely to help diagnose of diseases. The method commonly used in expert systems is rule-based reasoning, or case-based reasoning [1]. Case-based reasoning (CBR) methods have been widely applied in the medical field [2] - [6], due to the ability of CBR to work like an expert by retrieval of previous cases to solve new cases according to the given diagnosis [7]. The more old cases stored in the case base, the CBR system will be smarter in finding solutions for a given case. Problems with computation time and memory space requirements become a challenge especially when too many old cases exist on the case base. That is because the system must calculate the value of the similarity of new cases with all the old cases on the case base. A solution that can be used to shorter computational time is by finding solution that does not need to involve all data on the case base, but sufficient with some of the closest cases, so that the indexing process is needed [8].

Research focusing on the indexing process in CBR has been carried out with various methods, such as Fuzzy algorithm [9], back propagation classification algorithm [10], K-means clustering algorithm [11], and Local Triangular Kernel-Based Clustering (LTKC) algorithm [12]. K-means algorithm needs data of number of clusters that will be formed, because the assumption of the number of clusters determined at the beginning does not necessarily produce an optimal cluster. This method also has a low tolerance for data that contains noise and outliers. The back propagation and LTKC training process require quite long time because they have to try the training parameters one by one to get the best cluster. Clustering can group data sets that are not labeled into several data clusters based on similarity and dissimilarity [13]. Basically these algorithms work by grouping cases based on the specified features. When the retrieval process is carried out on the CBR, searching for similarity values can be conducted to cases that have the same index as new cases. Clustering algorithm can describe the patterns and tendencies contained in data groups. Each group represented by the value of the center of the cluster (cluster centroid). Cluster center enables measurement of similarity between new data and all cluster centers so it can determine the most similar data groups.

The proposed clustering method uses Self-Organizing Maps (SOM) compared to Density Based Spatial Clustering Application with Noise (DBSCAN). SOM is an artificial neural network-based learning algorithm that is good in exploration and visualization of high-dimensional data [14]. The training process on the SOM algorithm does not require supervision, the SOM network will learn without having a target in advance [15]. This is different from some artificial neural network methods such as back propagation which requires a target during the learning process. Density-based clustering methods such as DBSCAN have the characteristics of clusters with high density surrounded by clusters that have with low density. DBSCAN has advantages such as: being able to handle large amounts of data in short time, having tolerance to data containing noise and outliers, being able to recognize irregular shapes, being able to handle high dimensional data, and unnecessary to know the number of clusters to be formed [16] [17].

Each clustering algorithm requires testing to determine the quality of the clustering results. The validation of the results of clustering in this study was performed by evaluating the results of the clustering algorithm based on the structure that has been determined in the data set using Davies-Bouldin index and Silhouette index [18]. The process of looking of similarity between new cases and old cases in this study uses the nearest neighbor retrieval technique, by calculating the value of similarity or closeness between new cases and old cases. Three methods were used and compared, that are manhattan distance similarity, euclidean distance similarity and minkowski distance similarity.

2. Method

a. Knowledge Acquisition

This study used case data of medical record of patients with severe malnutrition at RSUP Dr. Sardjito Yogyakarta [3]. The malnutrition disease data consists of 90 data sets divided into 70 data as training data and 20 data as test data. The second case data is the medical record of patients with heart disease in the Medical Record Installation of RSUP Dr. Sardjito Yogyakarta [6]. The heart disease case data consists of 135 data sets divided into 115 data as training data and 20 data as test data. The third data is the diagnosis data on suspected thyroid disease from the Garvan Institute. The thyroid disease case data consists of 1428 data sets divided into 1000 data as training data and 428 data as test data.

b. Case Representation

The case representation used the frame model. Cases are represented as collections of features that characterize cases and solutions for handling these cases. Weighting of features is important to determine the level of significance of the feature to the disease. The weighting of each feature for each case is performed by an expert. If there are new cases, the weighting of disease features is divided into two categories, that are No and Yes. The value for each category is 0 for no symptoms and 1 for symptoms. After the old cases in the case base are clustered, the old case data is represented again by adding new knowledge derived from cluster center. Table 1 is a representation of cases of malnutrition in children under five who added new knowledge derived from the value of the cluster center.

Table 1. Representation of cases of malnutrition after clustering.

No	Case	Information
A	**Indication**	
1	G003	Rounded and swollen face
2	G009	Xylophone ribs
3	G019	Edema
4	G021	Very thin
B	**Patient data**	
1	Age	35 month
C	**Disease**	
1	P003	Marasmus-Kwashiorkor
D	**Indexing**	
1	Cluster	1

c. Indexing

The indexing method in this system used clustering method, i.e Density Based Spatial Clustering Application with Noise (DBSCAN) compared to Self-Organizing Maps (SOM). DBSCAN or SOM is used to group old case data into groups based on similarity and dissimilarity, so in each group contains similar data.

1) Data Normalization

The data normalization used the Min Max Normalization method. Normalization features include age, TSH, T3, TT4, and T4U since they have significant vulnerability. Min Max Normalization requires Minimum and Maximum age features. For example the age feature of malnutrition cases is a minimum value of 0 months and a maximum value of 60 months, and the age feature of a heart case minimum value is 0 years and the maximum value is 100 years. Equation 1 is the Min Max Normalization formula.

$$v' = \frac{v - \min}{\max - \min} \tag{1}$$

2) Selft Organizing Map (SOM)

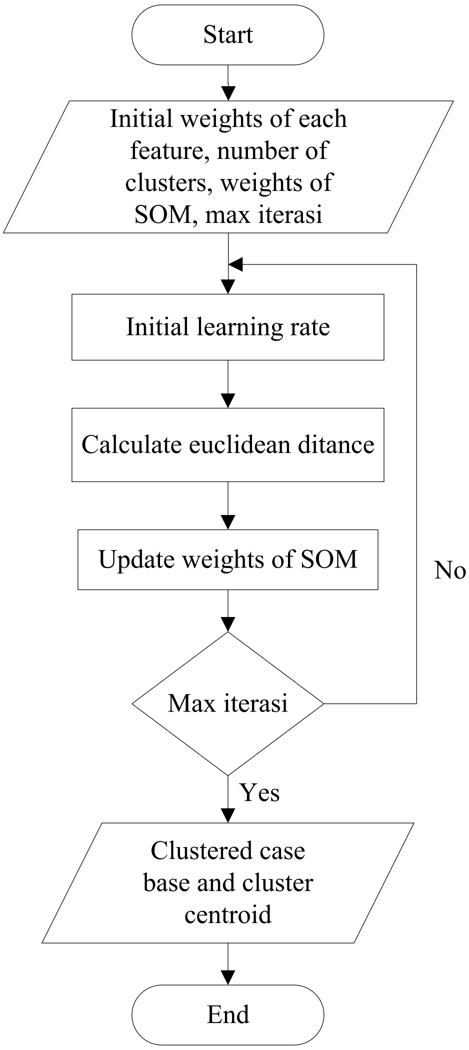
Start
Initial weights of each feature, number of clusters, weights of SOM, max iterasi
Initial learning rate
Calculate euclidean ditance
Update weights of SOM
Max iterasi
No
Yes
Clustered case base and cluster centroid
End

Figure 1. Clustering design using SOM algorithm.

Self-Organizing Map (SOM) algorithm or often referred as Kohonen Artificial Neural Network is one of the topology of Unsupervised Artificial Neural Network (Unsupervised ANN) in which the training process does not require supervision (target output). The clustering design using the SOM method is shown in the flowchart of Figure 1 [15]. Explanation of the flowchart diagram of Figure 1 is as follows:

a) Initializing the weights of each feature in the case base (ix) as input from SOM, number of clusters (k), initial weight (wi), and maximum iteration as SOM parameters.

b) Determine the learning rate (η) and decrease learning rate (α).

c) For each case base (xi) calculate the euclidean distance (Dj) to all initial weights of SOM (wij) using equation (2). After knowing the euclidean distance to each weight, look for the index that has the smallest value.

$$\mathrm{D}_{\mathrm{j}} = \sum_{i}^{n} \left(w_{ij} - x_i\right) \qquad (2)$$

d) Each wij weight within the radius of Dj neighborhood, the weight is updated by equation (3).

$$w_{ij}(new) = w_{ij}(old) + \alpha(x_i - w_{ij}(old)) \qquad (3)$$

e) Update the learning rate every 1 iteration with equation (4).

$$\eta(new) = \eta(old) \times \alpha \qquad (4)$$

f) As long as the maximum number of iterations has not been reached, repeat steps c through e.

g) Output clustering using the SOM method is a clustered case database and new weights are used as cluster center values.

3) Density Based Spatial Clustering Application with Noise (DBSCAN)

Density Based Spatial Clustering Application with Noise (DBSCAN) is one of the density-based clustering algorithms. The DBSCAN algorithm works by expanding high density regions into clusters and placing irregular clusters in the spatial database as noise. The clustering design using the DBSCAN method is shown in the flow chart of Figure 2 [16]. DBSCAN has 2 parameters, that are Eps or ε psilon (maximum radius of the neighborhood) and MinPts (minimum number of points in the Eps-neighborhood of a point).

Explanation of the flow diagram of Figure 2 is as follows:

a) Initializing the weights of each feature in the case base as DBSCAN input, the maximum radius of the neighborhood (Eps) and the minimum number of points in the Eps-neighborhood of a point (MinPts) as a DBSCAN parameter.

b) Specify one data as a random starting point (p).

c) For each case data in the case base, calculate the value of ε psilon or all distances that are density reachable to p using equation (5).

$$\mathrm{D}_{\mathrm{ij}} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2} \qquad (5)$$

d) If the amount of case data that meets ε psilon is more than MinPts, then p is a core point and one cluster is formed.

e) If there is no case data that is density reachable to p or the amount of case data that meets Eps is less than MinPts, then p is Noise.

f) Repeat steps c through e until all cases of case data base are processed.

g) Calculate the cluster center value (cluster centroid) using the average value for each cluster group.

h) The output of the case database is clustered and the average value is used as the cluster center value.

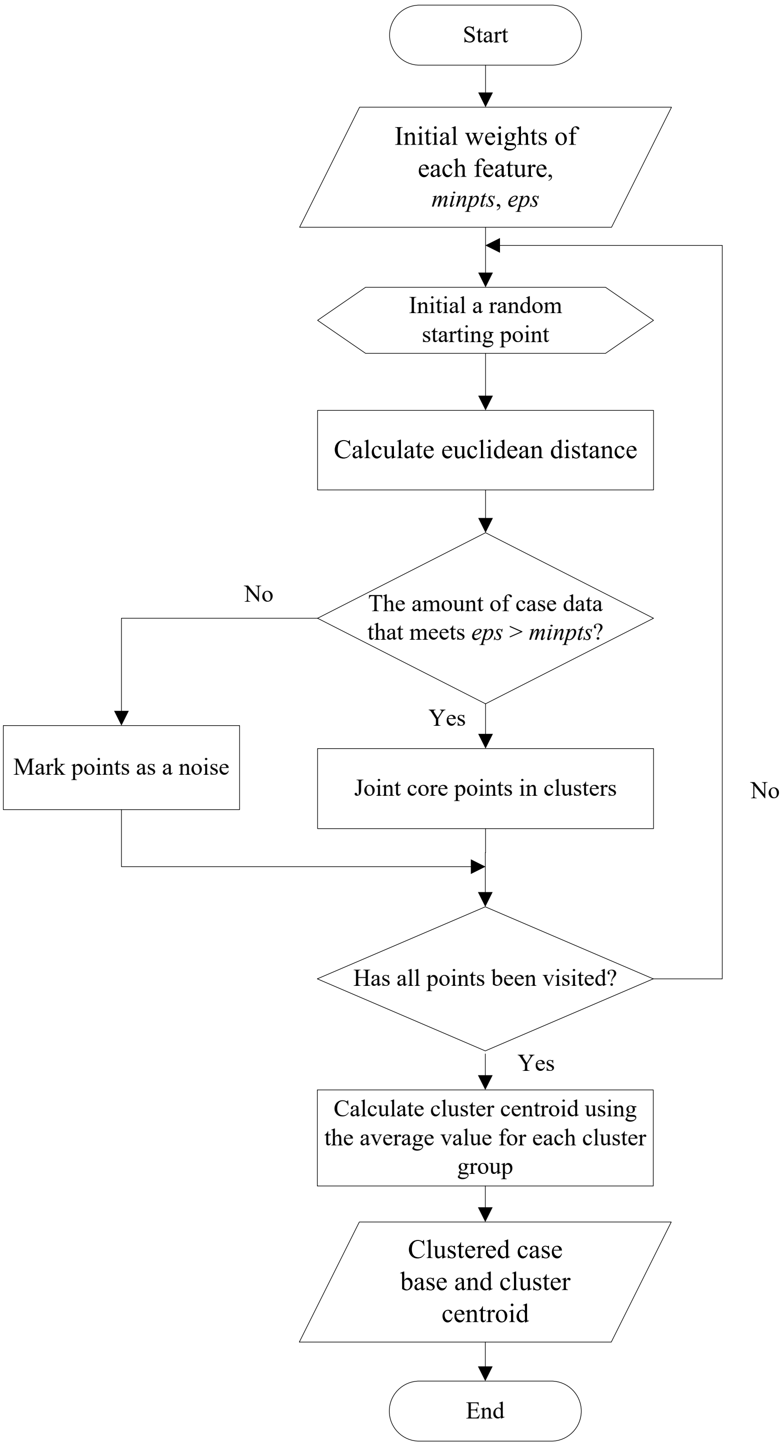
Start
Initial weights of each feature, *minpts*, *eps*
Initial a random starting point
Calculate euclidean distance
The amount of case data that meets *eps* > *minpts*?
No
Yes
Mark points as a noise
Joint core points in clusters
No
Has all points been visited?
Yes
Calculate cluster centroid using the average value for each cluster group
Clustered case base and cluster centroid
End

Figure 2. The design of clustering with DBSCAN algorithm.

c. Cluster Evaluation

The evaluation methods used in this system are the silhouette index and the Davies-Bouldin index methods. These methods are used to test the quality of the results of clustering. These methods are cluster validation methods that combines cohesion and separation methods. To calculate the value of silhouette index and Davies-Bouldin index, the distance between data is acquired by using the euclidean distance formula.

1) Silhoutte index

Silhoutte index was used to measure the quality and strength of a cluster, how well an object is placed in a cluster. The step of calculating the silhoutte index value starts with calculating the average distance from object i to all objects in a cluster. The calculation will produce an average value called ***ai***. Next, calculate the average distance from object ***i*** to objects in other clusters. Of all the average distances, take the smallest value, the value is called bi. Next, calculate the silhoutte index using equation (6) [18].

$$S_i = \frac{\left(b_i - a_i\right)}{\max\left(a_i, b_i\right)} \qquad (6)$$

Where ***s (i)*** is a Silhouette index value, ***a (i)*** is the average distance between point ***i*** and all points in ***A*** (the cluster where point ***i*** is located), ***b (i)*** is the average distance between point ***i*** to all points in clusters other than *A*. The silhoutte index value can vary between -1 to 1. The clustering result is good if the silhoutte index value is positive (*ai* <*bi*) and ai approaches 0, so that the maximum silhoutte index value is 1.

2) Davies-Bouldin index (DB index)

Davies-Bouldin Index has characteristics in validating clusters based on the calculation of quantity and derived features of the datas et. DB index value is calculated using equation (7) [18].

$$DB = \frac{1}{c}\sum_{c=1}^{c} Max_{\,i \neq j}\left\{\frac{d(x_i) + d(x_j)}{d\left(c_i, c_j\right)}\right\} \qquad (7)$$

Where ***DB*** is Davies-bouldin value, ***c*** is the number of clusters, ***d (xi)*** and ***d (xj)*** case data in clusters ***i*** and clusters ***j, d (ci, cj)*** is the distance between clusters ***ci*** and ***cj***. The smaller value of Davies Bouldin Index shows that the cluster configuration scheme is optimal and the cluster quality is getting better.

d. Retrieve and Reuse

CBR systems built with cluster-indexing can provide additional knowledge derived from previous cases. This knowledge is acquired from cluster center values generated from cluster analysis and added as a representation on a case base. After the case is represented by adding knowledge to the cluster center value, the case is then stored in a database. Figure 3 shows the architecture of the CBR system architecture with cluster-indexing.

If there are new cases, the system initializes the symptoms experienced by the patient and represents them as new cases. The system will search for the most relevant clusters by calculating the similarity of symptoms of new cases to the cluster center values. Similarity calculation is performed by comparing the euclidean distance between new cases with the cluster center value using the Cosine Coefficient method. After obtaining an index or cluster that is relevant to the new case, then a calculation is performed to find the similarity value between the new case and the cases in the case base that are in the same cluster.

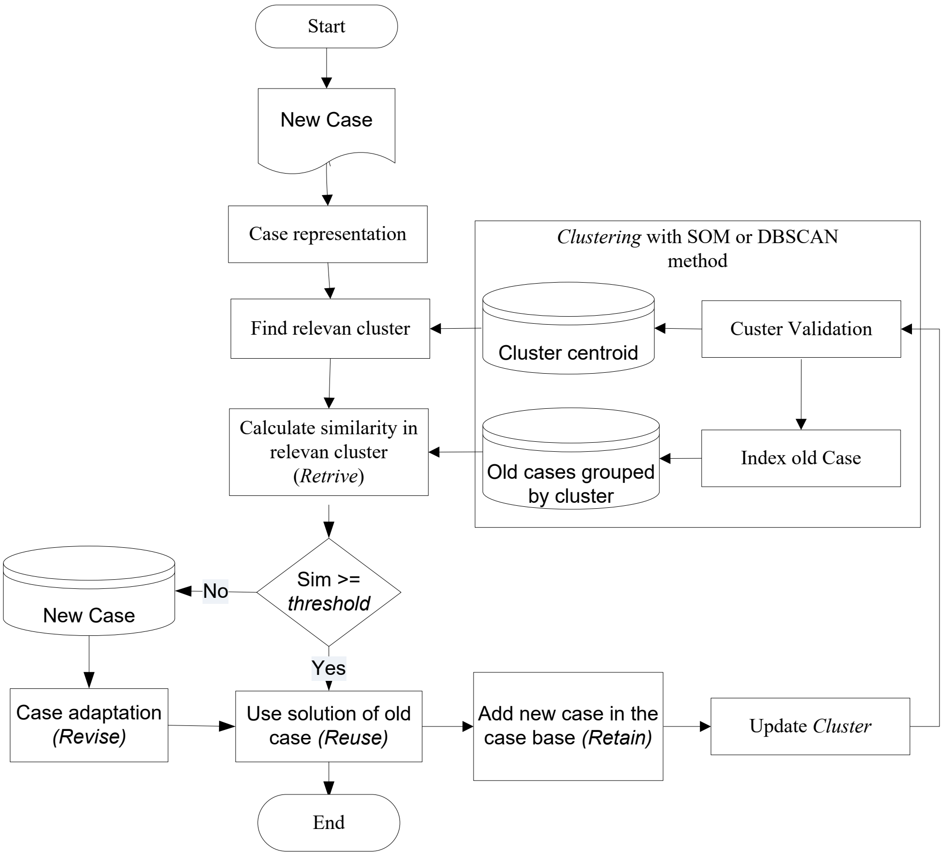
Start
New Case
Case representation
Clustering with SOM or DBSCAN method
Find relevan cluster
Cluster centroid
Custer Validation
Calculate similarity in relevan cluster (*Retrive*)
Old cases grouped by cluster
Index old Case
New Case
No
Sim >= *threshold*
Yes
Case adaptation (*Revise*)
Use solution of old case (*Reuse*)
Add new case in the case base (*Retain*)
Update *Cluster*
End

Figure 3. CBR system architecture design with cluster-indexing.

The threshold value of similarity are 0.7, 0.8, and 0.9 which means that if the highest similarity is greater than the threshold and close to 1, so this indicates that the new case has the exact same resemblance to the old case then the solution from the source case will be given to the user (reuse). If the similarity value decreases or is below the threshold, then the case will be stored in the database as a revision case which later the case under the threshold will be adjusted from the solution of the previous cases by the expert (revise). The new case is then saved to the case base by considering the cluster center value to become new knowledge (retain).

1) Determination of the Closest Cluster

During the process of finding a solution for a case, the CBR system will search for clusters that are most relevant to the new case by calculating the similarity of the symptoms of the old case with the cluster center value. Similarity calculation performed by comparing distances using the cosine coefficient method [19]. If given 2 vectors ***X*** and ***Y***, then the similarity value can be found by equation (8):

$$\text{Cos}\ (X, Y) = \frac{\langle X, Y \rangle}{\|X\| \bullet \|Y\|} \tag{8}$$

where **“‹ ›”** denotes the multiplication of vectors ***X*** and ***Y***, and **“| *X* || *Y* |”** denotes the norm for each vector. For vectors with non-negative elements, the cosine similarity value always lies between 0 and 1, where 1 indicates the two vectors are really the same, and 0 indicates the opposite.

The retrieval process used the nearest neighbor method. Nearest neighbor works by calculating the value of similarity, that is, the closeness between new cases and old cases based on matching weights of a number of existing features. There are two types of similarity measurements that are local similarity and global similarity [6]. Local similarity is a measurement of proximity at the feature level, whereas global similarity is a measurement of proximity at the object level (case).

Local similarity used in this study can be divided into two types, which are numerical and symbolic. The features included in the symbolic type are the symptom features and risk factors, while the numerical features are the sex and age features. Numerical data is calculated using equation (9)

$$f(S_i, T_i) = 1 - \frac{|S_i, T_i|}{|f_{\max} - f_{\min}|} \tag{9}$$

Note: ***f (Si, Ti)*** is the similarity of the ***i-feature*** of the old case or source case (***S***) with the new case or target case (***T***), Si is the value of the ***i-feature*** of the old case (source case), ***Ti*** is the ***i-feature*** value of the new case (target case), ***fmax*** is the maximum value of the ***i-feature*** on the case base and fmin is the minimum value of the ***i-feature*** on the case base. Meanwhile, symbolic data will be calculated using equation (10).

$$f(S_i, T_i) = \begin{cases} 0, if S_i \neq T_i \\ 1, if S_i = T_i \end{cases} \tag{10}$$

Note: ***f (Si, Ti)*** is the ***i-th*** feature similarity of the ***S*** (source) and ***T*** (target) cases, ***Si*** is the ***i-th*** value feature of the old (source) case and ***Ti*** is the ***i-th*** value feature of the new case (target).

Global similarity was used to calculate the similarity between new cases and cases on the case base. The methods to calculate global similarity in this study are Manhattan distance similarity in equation (11), euclidean distance similarity in equation (12), and minkowski distance similarity in equation (12) [20].

$$Sim(X, Y) = \frac{\sum_{i=1}^{n} f(S_i, T_i) * w_i}{\sum_{i=i}^{n} w_i} \tag{11}$$

$$Sim(X, Y) = \left(\frac{\sum_{i=1}^{n} w_i^r \times |f(S_i, T_i)|^r}{\sum_{i=i}^{n} w_i^r} \right)^{1/r} \tag{12}$$

Note: ***Sim (Si, Ti)*** is the value of similarity between the old case (***S***) and the new case (***T***), ***fi (Si, Ti)*** is the similarity of the ***i-th*** feature of the old case and the new case, the similarity of the ***i-th*** feature of the source case and target case, ***n*** is the number of features in each case, ***i*** is the individual feature, between ***1 s / dn***, ***wi*** is the weight given to the ***i-th*** feature, and ***r*** is the minkowski factor (positive integer). The value of ***r*** is equal to 2 for euclidean distance and equal to 3 for minkowski distance similarity.

e. CBR System Testing

Testing is performed by applying new cases, which are 20 data as test data for cases of malnutrition and heart disease and 428 data as test data for thyroid disease cases. The results of the system are then compared with the data contained in the medical record data. System accuracy is calculated by comparing the number of correct diagnosis with the amount of test data. The accuracy in this study is acquired by comparing the number of correct decision results and the amount of test data in accordance with equation (13).

$$\text{Accuracy} = \frac{\sum_{i=i}^{n} k_i}{n} \times 100\% \tag{13}$$

Note: ***ki*** is the ***i-th*** decision (***ki*** is 1 if the decision is right and 0 is if the decision is wrong), ***n*** is the amount of test data.

3. Results and Discussion

a. Case Base Clustering Process

The process of clustering of old cases on a case base used the SOM and DBSCAN clustering algorithms. The SOM method requires three parameters, which are number of clusters, maximum iteration, and learning rate. While the DBSCAN method requires two parameters, that are minimum points and epsilon. The parameter value is optimal if it produces the minimum Davies-Bouldin index value and the highest silhoutte coefficient and accuracy. The optimal parameter determination process carried out by clustering each case base data set using several combinations of parameters. Then each combination of these parameters is used to calculate the accuracy of the CBR retrieval process. Table 2 shows the SOM parameters and Table 3 shows the DBSCAN parameters.

The results of clustering with the SOM method depend on the initial weight given and the number of neurons in the output layer. Meanwhile, in DBSCAN the greater the ***minPts*** value, the more noise will be, this also affects the quality of the cluster. Therefore, determining the ε***psilon*** and ***minPts*** values at the beginning of the clustering process is very important. The quality of clustering results for the SOM and DBSCAN methods can be seen from the Davies-Bouldin index value, Silhoutte index and accuracy. The smaller the Davies-Bouldin index value, shows that the cluster parameters are optimal and the better the cluster quality. Meanwhile, for the Silhoutte index value getting closer to 1 shows that each case data is in the right cluster and there is no overlapping classes. Accuracy is determined by comparing the system diagnosis results and the actual diagnosis without applying a threshold value. The accuracy values of each trial are compared and the highest value for each data set is searched on the case base. The highest accuracy is used as the optimal clustering parameter.

Table 2. Optimal SOM parameters.

SOM attribute	Malnutrition Case Data	Heart Case Data	Thyroid Case Data
Amount of Clusters	3	5	5
Iteration	50	50	500
Learning Rate	0.1 – 0.2	0.4 – 0.5	0.7 – 0.8
Silhoutte index	0.378	0.279	0.303
DB index	0.812	0.324	0.587
Time (s)	0.439	1.167	11.48
Accuracy	100%	100%	87.15%

Table 3. Optimal DBSCAN parameters.

DBSCAN attribute	Malnutrition Case Data	Heart Case Data	Thyroid Case Data
Epsilon	1	13	0.5
MinPoints	3	3	10
Amount of Clusters	4	4	11
Amount of Noise	2	6	163
Silhoutte index	0.365	0.268	0.688
DB Index	0.888	0.462	0.420
Time (s)	0.124	0.282	8.37
Accuracy	100%	100%	90.89%

b. System Capability Analysis

The process of analyzing the ability of the system is divided into three scenarios. The first scenario is the diagnosis of the system using CBR non-indexing, the second scenario is the diagnosis of the CBR system with indexing using the SOM algorithm and the third scenario is the diagnosis of the CBR system with indexing using the DBSCAN algorithm. The searching process of relevant clusters with CBR cluster-indexing used the cosine similarity method and the similarity calculation process for all three scenarios used the Manhattan distance similarity method, euclidean distance similarity and minkowski distance similarity. Testing is performed by applying new cases, which are 20 data as test data for cases of malnutrition and heart disease and 428 data as test data for thyroid disease cases. Then the amount of the correct data is calculated, and the accuracy is determined according to the threshold and the average retrieval time for each similarity method. Based on the 3 testing scenarios, there are differences in the results of each scenario, as seen in Table 4 for cases of malnutrition, Table 5 for cases of heart disease and table 6 for cases of thyroid disease.

Table 4. Comparison of system capability in CBR non-indexing and CBR cluster-indexing for cases of malnutrition.

Scenario	Method	Threshold	Manhattan Distance	Euclidean Distance	Minkowski Distance
Scenario 1	**CBR non-indexing**	≥70	18 (90%)	18 (90%)	17 (85%)
		≥80	17 (85%)	18 (90%)	17 (85%)
		≥90	9 (45%)	18 (90%)	17 (85%)
Average retrieve time (seconds)			0.02598	0.02792	0.02925
Scenario 2	**CBR SOM indexing**	≥70	20 (100%)	20 (100%)	20 (100%)
		≥80	18 (90%)	20 (100%)	20 (100%)
		≥90	9 (45%)	20 (100%)	20 (100%)
Average retrieve time (seconds)			0.02269	0.02323	0.02425
Scenario 3	**CBR indexing DBSCAN**	≥70	20 (100%)	20 (100%)	20 (100%)
		≥80	18 (90%)	20 (100%)	20 (100%)
		≥90	9 (45%)	20 (100%)	20 (100%)
Average retrieve time (seconds)			0.02245	0.02288	0.02305

Table 5. Comparison of system capabilities in CBR non-indexing and CBR cluster-indexing for cardiac case data.

Scenario	Method	Threshold	Manhattan Distance	Euclidean Distance	Minkowski Distance
Scenario 1	**CBR non-indexing**	≥70	16 (80%)	20 (100%)	19 (95%)
		≥80	13 (65%)	20 (100%)	19 (95%)
		≥90	6 (30%)	12 (60%)	19 (95%)
Average retrieve time (seconds)			0.0535	0.0565	0.0469
Scenario 2	**CBR SOM indexing**	≥70	17 (85%)	20 (100%)	19 (95%)
		≥80	13 (65%)	20 (100%)	19 (95%)
		≥90	6 (30%)	12 (60%)	19 (95%)
Average retrieve time (seconds)			0.0417	0.0423	0.0424
Scenario 3	**CBR indexing DBSCAN**	≥70	17 (85%)	20 (100%)	20 (100%)
		≥80	13 (65%)	20 (100%)	20 (100%)
		≥90	6 (30%)	12 (60%)	19 (95%)
Average retrieve time (seconds)			0.0411	0.0418	0.0421

Table 6. Comparison of system capabilities in CBR non-indexing and CBR cluster-indexing for thyroid case data.

Scenario	Method	Threshold	Manhattan Distance	Euclidean Distance	Minkowski Distance
Scenario 1	**CBR non-indexing**	≥70	392 (91.56%)	393 (91.82%)	393 (91.82%)
		≥80	392 (91.56%)	393 (91.82%)	393 (91.82%)
		≥90	385 (89.95%)	392 (91.56%)	392 (91.56%)
Average retrieve time (seconds)			0.124	0.127	0.130
Scenario 2	**CBR SOM indexing**	≥70	353 (82.45%)	373 (87.15%)	373 (87.15%)
		≥80	352 (82.24%)	373 (87.15%)	373 (87.15%)
		≥90	324 (75.70%)	352 (82.24%)	352 (82.24%)
Average retrieve time (seconds)			0.112	0.114	0.119
Scenario 3	**CBR indexing DBSCAN**	≥70	389 (90.89%)	389 (90.89%)	389 (90.89%)
		≥80	389 (90.89%)	389 (90.89%)	389 (90.89%)
		≥90	371 (86.68%)	389 (90.89%)	389 (90.89%)
Average retrieve time (seconds)			0.105	0.106	0.107

The testing results of the three scenarios shows that the best accuracy and retrieval time at the threshold ≥ 90 for malnutrition disease data, acquired using the Minkowski distance method which is implemented on the CBR with indexing using the DBSCAN method. The accuracy is 100% with an average retrieval time of 0.02305 seconds. Research [3] with the same case data, reached the best accuracy of 85% with a threshold ≥ 0.75. So in the case of malnutrition, CBR with indexing using the DBSCAN method can improve accuracy. The best accuracy and retrieval time value at threshold ≥ 80 of heart disease data acquired using the Minkowski distance method implemented on CBR with DBSCAN indexing. The accuracy is 100% with an average retrieval time of 0.0421 seconds. This accuracy is as good as research [6] which produces 100% accuracy at threshold ≥ 80. The best retrieval time for thyroid disease data at threshold ≥ 90 is acquired using CBR with DBSCAN indexing of 0.107 seconds. This value is better than research [12] with an average retrieval time of 0.3045 seconds. But for accuracy calculation, CBR non-indexing is able to guess 392 correct data from 428 test data and produce an accuracy of 91.56%. Whereas CBR with cluster-indexing is able to guess 389 data from 428 test data and produces an accuracy of 90.89% which is implemented with the DBSCAN algorithm. This accuracy is smaller than the accuracy produced by research [12] using the Minkowski distance method with an accuracy of 92.52%.

In CBR with cluster-indexing the number of clusters greatly influences the retrieval time. Because the increasing number of clusters will make the cluster size of each cluster being relatively reduced. The retrieval time of the old case matching process will also be reduced, as the number of clusters decreases. On the other hand the time to search for relevant clusters will also increase in the process of finding the cluster center along with the increasing number of clusters. The number of clusters in the SOM algorithm is determined based on the number of output neurons while the initial weighting of the initial neurons is determined randomly. In DBSCAN, the larger value of ***εpsilon*** the wider scope of the cluster. While too small εpsilon will produce a large number of clusters and the distance of objects are very close each other. Likewise, too large ***minPts*** will produce a lot of noise. This will affect the accuracy of the CBR system with cluster-indexing.

Non-indexing CBR always provides the highest similarity value as a solution. The solution is required by comparing new cases with all cases on a case base. If the CBR non-indexing finds cases with the same similarity value, the cases are sorted by the earliest calculation process and the top case is taken to be a solution. The diagnosis with the highest similarity is not always the same as the diagnosis given by experts. This is because the similarity method does not consider the level of confidence in the new cases. For the next research, it is necessary to add the level of expert confidence in diagnosing the disease since the different features that exist in a particular case.

4. Conclusion

The results of clustering with SOM algorithm are depend on the initiation of the initial weight given to the cluster and the number of neurons in the output layer. Initial weight initiation in the SOM algorithm is generated randomly so it is possible to obtain different clustering results for the same parameters. Likewise with the DBSCAN algorithm, the results of clustering are depend on the value of εpsilon and minPts specified at the beginning. Therefore, a proper method is needed to determine the most appropriate parameters for the SOM and DBSCAN algorithms in order to produce the best cluster.

In the case of malnutrition and heart disease data testing, CBR with cluster-indexing has better accuracy and shorter processing time than non-indexing CBR. Whereas in the case of thyroid disease the accuracy of non-indexing

CBR is better than non-indexing CBR, even though CBR with cluter-indexing has a better average retrieval time. Cluster-indexing method with DBSCAN algorithm has a better accuracy, faster processing and retrieval time than SOM. Whereas, of the three similarity methods, the Minkowski distance method produced the highest accuracy at the threshold of ≥ 90. Further research needs to consider the level of confidence in the new case and the level of expert confidence of a case in calculating the value of similarity due to differences in features that exist in a particular case.

References

[1] P. Berka, "NEST : A Compositional Approach to Rule-Based and Case-Based Reasoning," *Adv. Artif. Intell.*, vol. 2011, 2011.

[2] N. Rumui, A. Harjoko, and A. Musdholifah, "Case-Based Reasoning for Stroke Disease Diagnosis," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 12, no. 1, pp. 33–42, 2018.

[3] Nurfalinda and N. Nikentari, "Case Based Reasoning untuk Diagnosis Penyakit Gizi Buruk pada Balita," *J. Sustain. J. Has. Penelit. dan Ind. Terap.*, vol. 06, no. 02, 2017.

[4] M. Benamina, B. Atmani, and S. Benbelkacem, "Diabetes Diagnosis by Case-Based Reasoning and Fuzzy Logic," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 3, pp. 72–80, 2018.

[5] L. G. Vedayoko, E. Sugiharti, and M. A. Muslim, "Expert System Diagnosis of Bowel Disease Using Case Based Reasoning with Nearest Neighbor Algorithm," *Sci. J. Informatics*, vol. 4, no. 2, pp. 7–10, 2017.

[6] E. Wahyudi and S. Hartati, "Case-Based Reasoning untuk Diagnosis Penyakit Jantung," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 11, no. 1, pp. 1–10, 2017.

[7] S. Mulyana and S. Hartati, "Tinjauan Singkat Perkembangan Case – Based Reasoning," *semnasIF UPNVN Yogyakarta*, pp. 17–24, 2009.

[8] A. Sarkheyli and D. Söffker, "Case Indexing in Case-Based Reasoning by Applying Situation Operator Model as Knowledge Representation Model," *IFAC-PapersOnLine*, vol. 28, no. 1, pp. 81–86, 2015.

[9] J. Lu, D. Bai, N. Zhang, T. Yu, and X. Zhang, "Fuzzy Case-Based Reasoning System," *Appl. Sci.*, vol. 6, no. 7, p. 189, 2016.

[10] T. Rismawan and S. Hartati, "Case-Based Reasoning untuk Diagnosa Penyakit THT (Telinga Hidung dan Tenggorokan)," *Indones. J. Comput. Cybern. Syst.*, vol. 6, no. 2, pp. 67–78, 2012.

[11] S. Guo, F. Yang, Q. Lu, and X. Liu, "Combination Case-Based Reasoning and Clustering Method for Similarity Analysis of Production Manufacturing Process," *Proc. - 2015 Int. Conf. Ind. Informatics - Comput. Technol. Intell. Technol. Ind. Inf. Integr. ICIICII 2015*, pp. 97–101, 2015.

[12] D. Riyadi and A. Musdholifah, "Local Triangular Kernel-Based Clustering (LTKC) for Case Indexing on Case-Based Reasoning," *Indones. J. Comput. Cybern. Syst.*, vol. 12, no. 2, pp. 139–148, 2018.

[13] D. L. Olson, *Descriptive Data Mining*, 1st ed. Singapore: Springer Singapore, 2017.

[14] R. Popovici and R. Andonie, "Music genre classification with Self-Organizing Maps and edit distance," *Proc. Int. Jt. Conf. Neural Networks*, 2015.

[15] R. Umar, A. Fadlil, and R. R. Az Zahra, "Self Organizing Maps (SOM) untuk Pengelompokan Jurusan di SMK," *KHAZANAH Inform.*, vol. 4, no. 2, pp. 131–137, 2018.

[16] H. Shah, K. Napanda, and D. Lynette, "Density Based Clustering Algorithms," *Int. J. Comput. Sci. Eng.*, vol. 3, no. 11, pp. 54–57, 2015.

[17] A. Musdholifah, S. Hashim, and S. Zaiton, "Cluster Analysis on High-Dimensional Data: A Comparison of Density-based Clustering Algorithms," *Aust. J. Basic …*, vol. 7, no. 2, pp. 380–389, 2013.

[18] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus External cluster validation indexes," *Int. J.*, vol. 5, no. 1, pp. 27–34, 2011.

[19] H. Seetha, M. N. Murty, and B. K. Tripathy, *Modern Technologies for Big Data Classification and Clustering*. Hershey PA: IGI Global, 2018.

[20] J. M. Merigó and M. Casanovas, "A new minkowski distance based on induced aggregation operators," *Int. J. Comput. Intell. Syst.*, vol. 4, no. 2, pp. 123–133, 2011.