Case-Based Reasoning for the Diagnosis of Acute Respiratory Infections Using Minkowski Distance

M Salmin¹, F Tempola², A Fuad³, M Papuangan⁴
¹Department Of Informatics, Faculty Of Engineering, Universitas Khairun, Indonesia
*Corresponding author: munazatc28@gmail.com

Abstract. One of the computerized systems is highly developed in this century is the system of Case-Based Reasoning (CBR) where the system is able to complete a new case based on experience or previous cases. CBR provides solutions based on the degree of similarity in the old case with a new case. The system built in this research is the CBR system for diagnosing respiratory diseases by using the Minkowski distance as the similarity calculation. Calculation of accuracy in this study using K-fold cross-validation by using k = 5 and k = 7. Testing with k = 5, the best results are ≥ 60% threshold with 99.29% accuracy while using the lowest threshold value of ≥ 80% with accuracy is 88.57%. Testing with k = 7, the best results are ≥ 60% threshold with accuracy 99.29%, while the lowest was using a threshold value of ≥ 80% with accuracy is 89.29%.

Keywords – Acute respiratory infection, CBR, K-fold, Minkowski distance

1. Introduction
Acute respiratory tract infection is a health problem that exists in developing and developed countries. Acute respiratory infections are a disease that most (90%) are caused by viral infections that do not require antibiotic treatment. Based on data from Daerah Istimewa Yogyakarta (DIY) Health Office (2013), the STP Hospital outpatient report was processed with results that were not much different from the reports Pusat Kesehatan Masyarakat (PUSKESMAS) level for the Yogyakarta Special Region, namely that of the 10 major diseases dominated by respiratory tract infections, the most influenza diseases were 39,675 people and Pneumonia 8,865 people [1]. Cases could be obtained from experience someone or experience an expert in the field could be implemented to make a computerized system. One of the ability humans try imitated by experts is the ability to complete problem-based on experiences previous. Experts computer make system called with case-based reasoning (CBR) [2]. solving cases using CBR has been done by [3]. One method on CBR used in measuring the chemistry between the cases is the Minkowski distance. The Minkowski distance method has been used by comparing three methods for similarity calculations, namely the nearest neighbor similarity, Minkowski distance, Euclidean distance methods. The test results for the initial diagnosis of heart disease show that the system is able to recognize heart disease by calculating a 100% accuracy rate using the Minkowski distance method [4]. Weighted method Minkowski distance has also been used to implement CBR with a target case above 80, the system can recognize cardiovascular diseases (I21) correctly (sensitivity) of 100%, recognize I21 disease (specificity) of 83.33%, with an accuracy of 95.83% and error rate of 4.17% [5]. Based on the description of several studies above, the researcher conducted a study on the diagnosis of ARI disease which was implemented in the CBR with the Minkowski distance. Similar cases retrieved by manual are very inefficient in terms of time, so a computer-based CBR system is
needed. With a computerized system, this can help medical personnel to be able to diagnose acute respiratory infections quickly and have accurate accuracy.

2. Methods

2.1 Case-Based Reasoning

The term case-based reasoning (CBR) began to be widely known in various fields both in the field of information technology and computer science. CBR is a problem-solving approach based on comparing new problems that occur with past experience or previous cases [6].

2.2 Measurement of similarity

When there are new problems that arise, the first step is to take cases that are stored on a case basis by measuring similarity. Measurement of similarity will produce values that determine whether there are or not similarities between the new cases and the cases that exist on a case-by-case basis by comparing features in new cases with similar types on the basis of cases. The measurement of similarity used in this study includes:

a. Local similarity

Local similarity shows the similarity between new problem attributes to attributes stored on a case base. Equation (1) to calculate local similarity for numerical data types [7]

\[
f(S_k, T_k) = 1 - \frac{|S_k - T_k|}{R}
\]

Where \( S_k, T_k \) is the value of the feature you want to compare and \( R \) is the range of values for that feature.

For Boolean data types, local similarity functions apply [8]. Use equation (2).

\[
f(S_k, T_k) = \begin{cases} 1 & S_k = T_k \text{for } S_k, T_k \{\text{true, false}\} \\ 0 & \text{else} \end{cases}
\]

b. Measuring Confidence level

The level of confidence is reflecting the level of certainty desired by the gauge after deciding not to take very many measurements [2]. Equation (3) to calculate the level of confidence that a new problem (\( T \)) is part of a class on a case base (\( S \)) [2]

\[
P_k(S_k, T_k) = P_k(S_k) \ast \frac{j(S_k, T_k)}{J(T_k)}
\]

Where, \( P_k(S_k, T_k) \) is the normalization of similarity with the level of confidence, \( P_k(S_k) \) is the percentage of confidence level in a case \( k \) in the source case, \( J(S_k, T_k) \) is the total feature in the target case that appears in the source case, and \( J(T_k) \) is the total feature in the target case.

c. Global similarity

In this study, the method used to calculate the similarity of new cases with the old case is the Minkowski distance method. Minkowski distance is a method of calculating similarity based on distance. This method is also a generalization from other distance methods. If \( r = 1 \) is known as Manhattan / City Block Distance, if \( r = 2 \) is known as Euclidean distance and if \( r = \infty \) is known as Chebyshev distance [9]. Minkowski distance weighted is shown using equations (4) [10].

\[
\text{Sim} (S, T) = \left[ \sum_{k=1}^{n} \left| (w_k \cdot f_k(s))^T \cdot f_k(S_k, T_k) \right| / \sum_{k=1}^{n} \left| (w_k \cdot f_k(s))^T \right| \right]^{1/r}
\]
Where, \( Sim (S, T) \) is the global similarity between the source case \((S)\) and the target case \((T)\), \(n\) is many features, \(r\) is the Minkowski factor (value \(r = 3\)), \(w_k, p_k\) is the weight values to feature in the source case \(k\), \(f_k (S_k, T_k)\) is a function of local similarity to \(k\) in diseases of the source case and the target case, \(S_k\) is attributed to \( -k \) in the source case and \(T_k\) is an attribute to \(-k\) in the target case.

The value of \(r\) is a positive number \(\geq 1\), (between 1 and infinity). In previous studies using \(r\) = 3 with maximum accuracy results.

Equation (4) is modified by adding the confidence level factor from equation (3) to the case similarity calculation as shown in equation (5).

\[
Sim (S,T) = \sum_{k=1}^{n} (w_k f_k(s))^{r+1} f_k (S_k, T_k)^r \left( \sum_{k=1}^{n} (w_k p_k(s))^{r} \right)^{1/r} \times p_k (S_k) \times \frac{f_k (S_k, T_k)}{f(T_k)}
\]

Where \( Sim (S, T) \) is the global similarity between the source case \((S)\) and the target case \((T)\), \(n\) is many features, \(r\) is the Minkowski factor (value \(r = 3\)), \(w_k, p_k\) is the value of the weight of the feature to \(k\) in the source case, \(f_k (S_k, T_k)\) are common features of the disease to \(-k\) in source case and the target case or a local similarity function, \(S_k\) is Features to \(k\) in source case, \(T_i\) is to \(-k\) feature on the target case, \(P_k (S_k)\) is the percentage level of confidence in the case to \(-k\) in source case, \(J (S_k, T_k)\) is the total feature in case the target that appears on the source case, and \(J (T_k)\) is the total feature in the target case.

2. 3 System Testing

K-fold Cross Validation.

One of the system validation processes is using K-fold cross validation. Cross validation is a statistical method of evaluating and comparing learning algorithms by data divided into two segments one used to learn or train the model and the other used to validate the model. The K-fold cross validation model is used to avoid overlapping data testing [12].

In this process, a subset of existing case data is formed. For this research, 5-fold validation is used, meaning 4 subsets are used as training sets (case base) and 1 subset is used as a testing set, with 5 iterations. The measurement results are the average values of 5 tests [11].

![Cross validation process 5-fold](image)

Accuracy is the level of proximity of the measurement of the quantity to the actual value. Accuracy measurement is done by comparing the number of correct diagnoses by the system with the number of test data. For the measurement of accuracy in this study used by k- fold cross-validation. Comparison can be written in the form of equation (6) [11].

\[
P = \frac{c}{n} \times 100\% 
\]
3. Result and Discussion

The stage in making a diagnosis is to count new cases with old cases, starting with calculating the local similarity then proceed with calculating global similarity. The following is an example calculation using local and global similarity.

Table 1. Example Of New Case Data

| No | Patient Data | Information | New Case Data |
|----|--------------|-------------|---------------|
| 1  | Age          | 18 year     | Age           |
| 2  | Gender       | Woman       | Gender        |
| 3  | Symptoms     |             | Symptoms      |
| 4  | G01          | Fever       | G01           |
| 5  | G02          | Emotional Pain | G02         |
| 6  | G03          | Throat Pain | G03           |
| 7  | G08          | Cough       | G08           |
| 8  | G10          | Malaise     | G10           |
|    | G16          | Sputum      | G16           |
|    | G18          | Hoarseness  | G18           |
|    | G21          | Gag         | G21           |
|    | G22          | Dizzy       | G22           |
| 1  | R03          | Infection   | R03           |

Table 2. Example case based in frame

| Case data 001 | Information | Weight | Case data 002 | Information | Weight |
|---------------|-------------|--------|---------------|-------------|--------|
| Data patient  |             |        |               |             |        |
| Age           | 16          | 4      | Age           | 23          | 4      |
| Gender        | woman       | 3      | Gender        | man         | 3      |
| Symptoms      |             |        | Symptoms      |             |        |
| G01           | Fever       | 3      | G01           | Fever       | 3      |
| G02           | Emotional pain | 7     | G02           | Emotional pain | 7   |
| G03           | Throat pain | 5      | G03           | Throat pain | 5      |
| G05           | Cold        | 2      | G04           | headache    | 3      |
| G08           | Cough       | 2      | G08           | Cough       | 2      |
| G10           | Malaise     | 3      | G10           | Malaise     | 3      |
| G16           | Sputum      | 4      | G11           | Anorexia    | 2      |
| G18           | Hoarseness  | 2      | G16           | Sputum      | 4      |
|               |             |        | G22           | Dizzy       | 2      |
| Factor risk   |             |        | Factor risk   |             |        |
| R03           | Infection   | 3      | R03           | Infection   | 3      |
| R05           | History of disease | 3 | Type of disease | J029        |
|               |             |        |               |             |        |
| Type of disease | J029    |     |               | J029        |
| Expert confidence level | 100% |        | Expert confidence level | 100% |

Process Diagnosis is done by calculating the local similarity which is divided into 2 (two) parts, namely age, sex, symptoms and risk factors. Age calculation uses equation (1), while gender closeness, symptoms and risk factors use equation (2).

Discussion of Test Results
This test uses the fold cross-validation method, where the total data on cases of is a diseases are 140 cases divided into 5 subsets (k = 5) and 7 subsets (k = 7) with details of each subset of 28 case data and 20 case data. The division of cases into subsets is done randomly where all types of diseases are filled at each subset evenly. For measuring the accuracy of system performance built in this study using equation (6).

| Testing | 1 (%) | 2 (%) | 3 (%) | 4 (%) | 5 (%) | Average of accuracy |
|---------|-------|-------|-------|-------|-------|---------------------|
| thr \( h_0 \) \( \geq 60\% \) | 100   | 100   | 96.43% | 100   | 100   | 99.29%             |
| thr \( h_0 \) \( \geq 65\% \) | 100   | 100   | 96.43% | 100   | 96.43% | 98.57%             |
| thr \( h_0 \) \( \geq 70\% \) | 100   | 96.43% | 89.29% | 100   | 96.43% | 96.43%             |
| thr \( h_0 \) \( \geq 75\% \) | 96.43% | 89.29% | 89.29% | 100   | 96.43% | 94.29%             |
| thr \( h_0 \) \( \geq 80\% \) | 92.86% | 85.71% | 82.14% | 89.29% | 92.86% | 88.57%             |

Result Of Average Accuracy For K = 5
Testing with k = 5 uses the threshold \( \geq 60\% \) accuracy is 99.29%, testing using a threshold \( \geq 65\% \) accuracy is 98.57%, testing using a threshold \( \geq 70\% \) accuracy is 96.43%, testing using a threshold \( \geq 75\% \) accuracy is 94.29%, testing using a threshold \( \geq 80\% \) of the accuracy is 88.57%. To be more clear about the details of testing, see Table 3.

| Testing | 1(%) | 2(%) | 3(%) | 4(%) | 5(%) | 6(%) | 7(%) | Average of accuracy |
|---------|------|------|------|------|------|------|------|---------------------|
| thr \( h_0 \) \( \geq 60\% \) | 100  | 100  | 100  | 95   | 100  | 100  | 100  | 99.29%             |
| thr \( h_0 \) \( \geq 65\% \) | 100  | 100  | 100  | 95   | 100  | 100  | 95   | 98.57%             |
| thr \( h_0 \) \( \geq 70\% \) | 100  | 100  | 90   | 90   | 100  | 100  | 95   | 96.43%             |
| thr \( h_0 \) \( \geq 75\% \) | 100  | 100  | 85   | 85   | 100  | 95   | 95   | 94.29%             |
| thr \( h_0 \) \( \geq 80\% \) | 90   | 90   | 85   | 80   | 95   | 90   | 95   | 89.29%             |

The average yield accuracy for k = 7
Testing with k = 7 uses the threshold \( \geq 60\% \) accuracy is 99.29%, testing using a threshold \( \geq 65\% \) accuracy is 98.57%, testing using a threshold \( \geq 70\% \) accuracy is 96.43%, testing using a threshold \( \geq 75\% \) accuracy is 94.29%, testing using a threshold \( \geq 80\% \) accuracy is 89.29 %. To be more clear about the detailed information, see Table 4.

4. Conclusions and Recomendation

4.1 Conclusion

Based on the results of previous studies and the analysis process in this study, the following conclusions are obtained:

1. [13], [14] implements CBR using the weighted minkowski method with a target case above 80%, with an accuracy rate of 95.83%.
2. Diagnosed schizophrenia by comparing \( r = 1, r = 2 \) and \( r = 3 \) where the value of \( r = 3 \) has very good accuracy compared to the others[15], [16].

3. This research provides the results of the CBR system for the diagnosis of ARI disease by taking into account the closeness between new cases and old cases based on features of age, gender, symptom features and risk factor features that have been given weight and expert confidence level.

4. The test results with \( k = 5 \) (5-fold) from the data obtained in the patient's medical record with each subset filled with 28 data, the best result is using the threshold \( \geq 60\% \) with an accuracy of 99.29\% while the lowest uses a threshold of \( \geq 80\% \) with an accuracy value of 88.57\%. Testing with \( k = 7 \) (7-fold) from the data obtained in the patient's medical record with each subset filled with 20 data, the best result is by using the threshold \( \geq 60\% \) with an accuracy of 99.29\% while the lowest using the threshold \( \geq 80\% \) with an accuracy value of 89.29\%.

5. Testing using a threshold of \( <60\% \) is not done because on the basis of cases there are certain diseases that have few features so that the input of the age and sex of the patient is already can diagnose an illness.

4.2. Recommendation

In this research there are still many shortcomings, so it is recommended for further research with similar topics or methods, namely the following:

1. Some criteria need to be added such as physical examination, laboratory and others in order to get better results.

2. This study is limited to 4 types of ARI so that in future studies it can be added more to the type of ARI so that more ARI can be diagnosed in the CBR system.

5. References

[1] Dinkes D.I.Y., “Profil Kesehatan D.I.Y,” Denise Kesehatan DIY, 2013. [Online]. Available: http://www.depkes.go.id/resources/download/profil/PROFIL_KES_PROVINSI_2012/14_Profil
_Kes.Prov.DIYogyakarta_2012.pdf.

[2] S K Pal and S C K Shiu 2004 Foundation of Soft Case-Based Reasoning, 1st ed. New Jersey: Wiley-Interaction

[3] Tempola F, Musdholifah A, and Hartati 2018 “Case-Based Reasoning for Determining the Feasibility of Scholarship Grantees Using Case Adaptation,” Proc. - 2018 5th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2018, pp. 370–374.

[4] Wahyudi E, Hartati S dan Musdholifah A 2017, “Case-Based Reasoning Untuk Diagnosis Penyakit Jantung,” IJCCS (Indonesian J. Comput. Cybern. Syst.), vol. 11, no. 1, pp. 1–10.

[5] Faizal H E 2018, “Weighted Minkowski similarity method with CBR for diagnosing cardiovascular disease,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 12, pp. 305–310.

[6] Katedee T, Sanrach S C and Thesawadwong 2010 Case-Based Reasoning System for Histopathology Diagnosis. Thailand, 2010.

[7] B M K Jha, D Pakhira and Chakraborty 2013, “Diabetes Detection and Care Applying CBR Techniques,” IJSCE, vol. 6, no. 2, pp. 132–137.

[8] Y Nurdiansyah and Hartati S 2014, Case-Based Reasoning Untuk Pendukung Diagnosis Gangguan pada Anak Autis (Yogyakarta: Universitas Gadjah Mada).

[9] J M Mergio and M Casanovas 2008 “The Induced Minkowski Ordered Weighted Averaging Distance Operator, ESTYLF08, Cuencas Mineras (Mieres-Langreo),” in Congreso Espanol sobre Tecnologias Logica Fuzzy, pp. 35–41.

[10] H Nunez, M S Marre, U Cortes, J Comas, M Martinez, I R Roda and M Poch 2004 “A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations,” Environ. Model. Softw., vol. 19, no. 9, pp. 809–819.

[11] M. Bramer 2007 Principles of Data Mining (London: Springer-Verlag).

[12] Madhukar, M., Aguian, S., and Chronopoulos, A.T., 2012, New Decision Support Tool for Acute Lymphoblastic Leukemia Classification, Dalam Journal of SPIE-IS&T/ Vol. 8295 829518-1.
[13] E. Faizal, “Case Based Reasoning Diagnosis Penyakit Cardiovascular Dengan Metode Simple Matching Coefficient Similarity,” J. Teknol. Inf. dan Ilmu Komput., vol. 1, no. 2, p. 83, 2014.

[14] Y. S. Bagi and S. Suprapto, “Hepatitis Diagnosis Using Case-Based Reasoning with Gradient Descent as Feature Weighting Method,” J. Inf. Syst. Eng. Bus. Intell., vol. 4, no. 1, p. 25, 2018.

[15] A. Labellapansa, A. Efendi, A. Yulianti, and A. K. Evizal, “Lambda value analysis on Weighted Minkowski distance model in CBR of Schizophrenia type diagnosis,” 2016 4th Int. Conf. Inf. Commun. Technol. ICoICT 2016, pp. 137–140, 2016.

[16] Rismawan, T., dan Hartati, S “Case-Based Reasoning untuk Diagnosa Penyakit THT (Telinga Hidung dan Tenggorokan),” IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 6, no. 2, pp. 67–78, 2013.