Comparison of Network and Data Correlation in Modeling Revise Stage of Case Based Learning

I Gede Santi Astawa[1,*], Desak Putu Sri Wulandari[2], Luh Putu Ida Harini[3]

[1,2,3] Udayana University, Badung, Bali – Indonesia
*santi.astawa@unud.ac.id

Abstract. Case-based reasoning (CBR) is a method that models and adapts experience to find the right solution for new problems. CBR stages consist of retrieve, reuse, revise, and retain. Revise stage is one of important key in CBR method to adapt a new case. This is because if a problem cannot be found the right solution in the knowledge base, then this stage will adapt experience to become an appropriate solution. Based on the research reviewed, the revise stage still uses active intervention by experts on the system. The involvement of experts directly in the system requires costs and is limited by space and time, so it is necessary to create a mechanism that can work automatically. The mechanism created is expected to be able to approach the expertise of an expert. However, in optimizing the process, it is necessary to carry out correlation analysis on knowledge data owned by CBR, so that data features that are highly correlated with the class can be selected. Otherwise neural network (NN) use to find the role revise model on the CBR data by learning its distance. In this study, we will create a CBR system which at the revise stage utilizes data correlation, CBR system which at the revise stage utilizes NN role modelling, and also creates a CBR system that still uses an expert, namely a teacher. The test results of the 10 test-data for both models, obtained an accuracy value of 70% for the CBR system at the revise stage using expert assistance, 90% for the CBR system at the revise stage using a data correlation model, and 87% for the CBR system at the revise stage using NN model. Based on these tests, the results of this study can be said that the CBR system with the data correlation model at the revise stage is able to approach or exceed the expertise of an expert.
.

Keywords: Case-Based Reasoning; Revise; Data Correlation; Student Performance.

1. Introduction

Case-based reasoning (CBR) is the method used to solve practical problems using experience adaptation and relying on a knowledge base [8]. This method is an artificial intelligence method that models and adapts experience to meet the rules as an

K. Dharmawan and N. A. Sanjaya ER (eds.), *Proceedings of the First International Conference on Applied Mathematics, Statistics, and Computing (ICAMSAC 2023)*, Advances in Computer Science Research 110,
https://doi.org/10.2991/978-94-6463-413-6_9

appropriate solution to answer new problems. Therefore, experience is the most important aspect in implementing this method. CBR has four stages including retrieve, reuse, revise, and retain. Retrieve means to get back, reuse means to reuse, revise means to review (revise), and retrain is to use a new solution as part of a new case. The CBR method has been implemented in various studies that have been conducted. Such as research related to the implementation of Case-based reasoning to detect hard drive damage [6]. In addition, there is research related to the application of case-based reasoning in decision support systems for handling mall tenant complaints [3]. In these studies, CBR is said to be able to find appropriate solutions to cases.

In CBR, the revise stage is the key to the CBR method. This is because if a problem is not found the right solution in the knowledge base, then this stage will adapt experience to become an appropriate solution. In several studies that have been studied, the revise stage still uses active intervention by experts on the system. Such as research related to identifying laptop damage using case-based reasoning with the nearest neighbor algorithm, obtaining an accuracy value of 100% [1]. In another study conducted regarding the prediction of student study time, the highest evaluation results with an accuracy value of 100%[2]. The accuracy value is obtained because at the revise stage it is done manually by a qualified expert interested. Thus, the system that has been created does not work automatically in providing solutions to new cases. The involvement of experts directly in the system requires costs, and is limited by space and time, so it is necessary to create a mechanism that can work automatically. The mechanism created is expected to be able to approach the expertise of an expert. One method that can be used in the revise stage is by looking at the relationship between data or data correlation. However, in optimizing the process, it is necessary to carry out correlation analysis on knowledge data owned by CBR, so that data features that are highly correlated with the class can be selected.

The CBR method can be used to measure performance. In this study, performance measurements were carried out on a dataset of student performance in mathematics. This research will create a CBR system which at the revise stage automatic by system and also creates a CBR system that still uses an expert, namely a teacher. Then performed a performance comparison between the models.

2. Research Methodology

This research consists of several stages such as data collecting and processing, design of Case-based Reasoning, and evaluation method.

Data Collection and Processing

The data that used in this research is data that taken from the UCI Machine Learning repository, namely student performance in math data with 347 data and 31 features, where the 31st feature is the class label of the student performance. The features in the data can be seen in table 1.

Table 1. Student Performance in Math

No.	Feature Name	No	Feature Name
1	Sex	17	Activities
2	Age	18	Nursery
3	Address	19	Higher
4	Famsize	20	Internet
5	Pstatus	21	Romantic
6	Medu	22	Famrel
7	Fedu	23	Freetime
8	Mjob	24	Goout
9	Fjob	25	Dalc
10	Guardian	26	Walc
11	Traveltime	27	Health
12	Studytime	28	Absences
13	Failures	29	G1
14	Schoolsup	30	G2
15	Famsup	31	G3
16	Paid		

In this research, data processing (pre-processing data) was carried out, such as data transformation, feature selection and data normalization. The purpose of data transformation is to change the measurement scale of the data into another form in order to fulfill the homogeneity of the variance and normal distribution of the data. The dataset of student performance in mathematics subjects has different data measurement scales, namely numerical and categorical. This data transformation uses will change categorical variables into numerical variables by performing a One-Hot-Encode process on categorical variables. In addition to changing the scale of data measurement, this process also converts the value of class features, namely G3 into 2 classes.

Feature selection is carried out to select the features to be used in research. The resulting features from the data transformation will be calculated for the correlation of the G3 features. After obtaining the correlation value, then the features with a correlation value above 0.05 are selected which are features for the revision stage in the CBR system. The data normalization is done with the aim of equating the range of values between data features. The data normalization method used in this research is Min Max Normalization.

Design of CBR System

In retrieve stage, new case will be searched for similarities with the knowledge base. This search stage starts with examining a new case with all the data in the knowledge base which aims to check the similarities between the features of the data that will determine the solution to the case. Feature similarity is calculated using cosine similarity with a threshold value of 0.95. This value is obtained because the dataset used has a high similarity so that the specified limit is high.

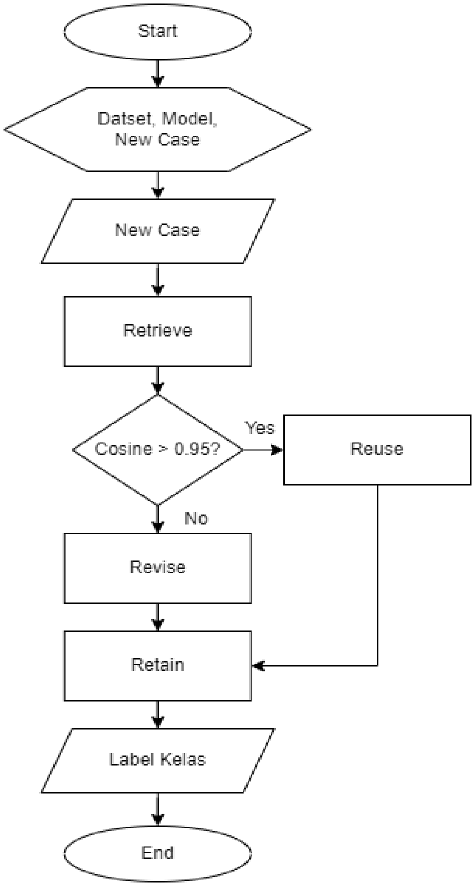

Fig. 1. Flow Diagram of CBR Method

In reuse stage, the case that has the highest similarity value above 0.95, the retieved case will be used directly as a solution for a new case. If the highest similarity value is below the threshold, then a new case solution will be sought using the revise stage. The revise stage, the process of adapting solutions to new cases is done by calculating the similarity distance between the new case data features and the cases in the knowledge base. This calculation uses features with a correlation coefficient value above 0.05[7]. This model will be combined with the K-Nearest Neighbor (KNN) algorithm as shown in figure 2.

When the system finds a suitable solution for a new case, the case will be entered into the knowledge base to be reused in finding solutions to similar problems in the future. In order to avoid duplication of data, at the retain stage an examination is carried out on the availability of data.

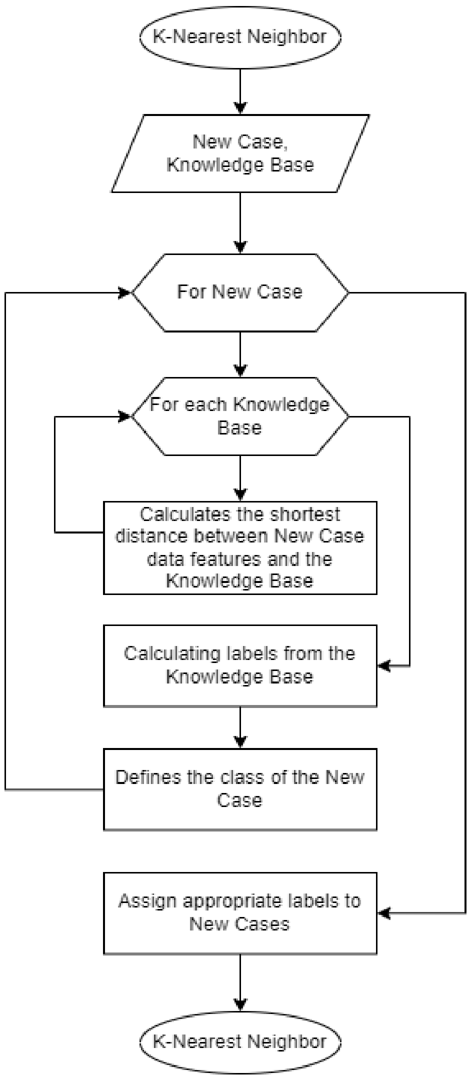

Fig 2. Flow Diagram of Revise Stage

Method Evaluation

In this research, the testing of the implementation of the method were carried out by comparing the accuracy of the CBR system, which was revised using a automatic model with which was revised using expert assistance. The data used in this test is 25 student performance data. The CBR system with the help of experts will be given the test data which then the expert will label the student's performance. These results will be compared with the results of the CBR system in which the revision stage applies automatic models by calculating accuracy. The accuracy calculations use the confusion matrix with the following formula.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \; x \; 100\% \quad (1)$$

Explanation:
TP: When predictions and reality are true.
FP: When predictions doesn't happen but reality does.
FN: When predictions happen but reality doesn't happen.
TN: When predictions and reality don't happen

3. Results and Discussion

3.1 Data Correlation based model

Based on the doing research, the obtained features that have a correlation coefficient value above 0.05 which is 16 features as shown in the table 3. These features will used in the revise stage

Table 2.

Feature	Correlation Value
G2	0.965046
G1	0.888697
schoolsup_no	0.219075
Medu	0.19988
Fjob_teacher	0.18332
Fedu	0.164606
Mjob_health	0.149392
address_U	0.145944
Mjob_services	0.114722
studytime	0.113614
higher_yes	0.110312
sex_M	0.107215
internet_yes	0.106972
activities_yes	0.067379
famsup_no	0.058864
romantic_no	0.058054

In the revise stage, the use of data correlation will be combined using the KNN algorithm. The use of this algorithm is because at the revise stage it will look for the distance between data features in new cases and data in the knowledge base. Based on the distance found, a comparison of the distance class of the data is carried out by equation 2.

$$\nabla_{class} = \frac{\sum_{i=1}^{n}(correlation_i \times \nabla_i)}{\sum correlation} \dots (2)$$

The results of the data similarity comparison will determine the class of the new case. If the ∇_{class} more then the cut off point, the class of new data will be different with the class of nearest data on CBR. But if ∇_{class} less then the cut off point, the class will be same. If the similarity of the data refers to a certain class then that class will be the solution to the case. For example, if the results of the similarity of new cases refer to data that has class "0", then the case has a solution in the form of class "0".

3.2 NN based model

Neural network (NN) generally used in classification problems, automatically NN can learn the CBR data to determine it class. In this research used backpropagation NN with single hidden layer to learn different of new data and nearest CBR data. Input layer consist of 31 neuron that represent the distance of feature of two randomize data. And the output NN is a single neuron that represent class changed or not. In this model NN learn how the distance of two data can effect the class of data.

3.3 Research data

In this research, an assessment was carried out by experts and systems that used the model at the revision stage. Assessments made by experts have results as shown in table 4.13. Based on this table, the accuracy obtained by experts in predicting the class of a data reaches 83.75%. Meanwhile, the research carried out by the system with data correlation model as shown in table 4.15 is capable of producing an accuracy value of 90%. And by NN model the accuracy value is 94%

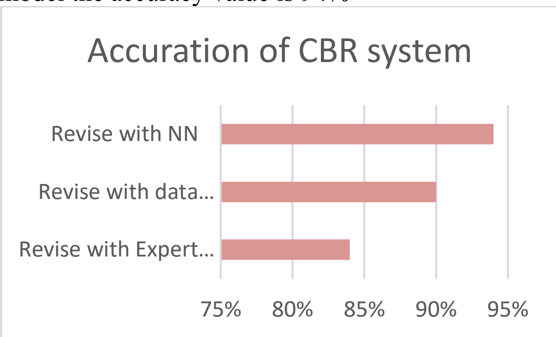

Fig 3. Comparison of CBR System

Based on the results of testing process conducted by experts and the system, it was found that the system's accuracy value was better than the expert's accuracy value. Accuracy comparison graph can be seen in Figure 3. From these results, it can be determined that the CBR system is capable of producing precise accuracy when at the revision stage using a data correlation model compared to using expert assistance.

4. Conclusion

Conclusion

Based on the presentation of the research, it can be concluded that the results of the comparison of student performance predictions using CBR with data correlation models at the revision stage are much better than CBR with expert involvement where the accuracy of the CBR system with the model is 90% compared to the accuracy of the CBR system with expert assistance of 83.75%. The accuracy of the system is much better because it uses a correlation model from the knowledge base used in the CBR system. Meanwhile, the accuracy produced by experts is much lower. This is because these results depend on the perspective and knowledge of each expert.

Suggestion

In the next research, it is suggested to try other method in revise stage of CBR which better from method in this research to produce better accuracy values.

References

[1] Afrano Amran, Agustinus. (2018). *Implementasi Metode Penalaran Case-based reasoning (CBR) Dengan Algoritma Nearest Neighbor Dalam Identifikasi Kerusakan Laptop*. Skripsi. *Jurusan Teknik Informatika. Fakultas Sains dan Teknologi.* Universitas Sanata Dharma. Yogyakarta..

[2] Endah, Wahanani, Hamindia P.S, Made, Ali Akbar, Fawwar. (2020). *Case Based Reasoning Prediksi Waktu Studi Mahasiswa Menggunakan Metode Euclidean Distance dan Normalisasi Min-Max*. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 7(6), 1279-1288.

[3] Kartikasari, Meivi, Santoso, Purnomo B. and Yudaningtyas, Erni. (2015). *Penerapan Case Based Reasoning pada Sistem Pendukung Keputusan Penanganan Komplain Penyewa Mall*. Jurnal EECCIS, 9(2), 138-143.

[4] Richter, Michael M dan Weber, Rosina O. (2013). Case-Based Reasoning: A Textbook. Springer. New York.

[5] Rahma, Abdul. A.N., Rezza, Darmawan, Agung, Rianto, Yan, dan Syafrullah, M. (2019). Prediction Of Students Academic Success Using Case Based Reasoning. Proceeding of International Conference on Electrical Engineering, Bandung: 18 – 20 September 2019, 171-176.

[6] Ritha, Nola and Sutoyo, M. Nurtanzis. 2016. Case Based Reasoning Untuk Mendeteksi Kerusakan Harddisk. Jurnal Sustainable, 5(1), 5-21.

[7] Sammut, Claude. and Webb, Geoffrey I. (2017). Encyclopedia of Machine Learning. Springer. New York.

[8] Supriana, I Wayan dan Dwi Prebiana, Kiki. (2020). *Metode Penalaran Berbasi Kasus (Case Base Reasoning) Dalam Penentuan Kelayakan Sekolah Perawatan*. Jurnal Rekayasa Sistem Komputer, 3(1), 57-65.