

A Case-Based Reasoning Method with Rank Aggregation

Jinhua Sun*, Jiao Du, Jian Hu

School of Management, Chongqing University of Technology, Chongqing, China

*Corresponding author e-mail: sjh1009@163.com

Abstract. In order to improve the accuracy of case-based reasoning (CBR), this paper addresses a new CBR framework with the basic principle of rank aggregation. First, the ranking methods are put forward in each attribute subspace of case. The ordering relation between cases on each attribute is got between cases. Then, a sorting matrix is got. Second, the similar case retrieval process from ranking matrix is transformed into a rank aggregation optimal problem, which uses the Kemeny optimal. On the basis, a rank aggregation case-based reasoning algorithm, named RA-CBR, is designed. The experiment result on UCI data sets shows that case retrieval accuracy of RA-CBR algorithm is higher than euclidean distance CBR and mahalanobis distance CBR testing. So we can get the conclusion that RA-CBR method can increase the performance and efficiency of CBR.

1. Introduction

Case-based reasoning (CBR), an artificial intelligence technique based on empirical knowledge that solves new problems by modifying or reusing the similar cases in the past[1]. The core of CBR method is case retrieval, which main function is to match the target case with the historical cases and then to solve the target cases by finding the historical case with high matching degree.

It is a NP-hard problem of case retrieval due to including multiple attributes in case, such as clear symbol attribute, interval number attribute, clear data attribute and blurry language attribute. The common approach is to use weights to integrate multiple case attributes. But whether it is the subjective determination method or the objective weight method, there are obvious shortcomings. Therefore, the attributes directly affect the case retrieval result. It has become one of important research directions that how to solve the problem of multiple attribute aggregation in the CBR method.

For solving above problems, a new ideal of ranking is put forward to get the similar case instead of similarity, which can make up these shortcomings of weights allocation of case attributes. The main thought is firstly getting ordering relation in each attribute subspace of historical cases through comparing with historical case. Then, a sorting matrix is got. The similar case retrieval process from ranking matrix is transformed into a rank aggregation optimal problem through Kemeny optimal principle.

2. Related Work

In classical CBR system, the similarity between pairwise cases is calculated by weights of case attributes. Often, the weights need to be artificially determined, such as decision-maker subjective allocation weight [2, 3], the Delphi method [4], the statistical method [5] and hybrid weighted mean method [6], which will be influenced by subjectively. According to the shortcoming of subjective weight, the objective weights allocation method is put forward by some scholars, such as entropy



method [7], neural networks algorithm [8], genetic algorithm [9] and membrane computing weight optimization [10]. However, the weight allocation is uneven through objective method, which is easily influenced by fuzzy randomness.

According to the shortcomings of weight determination method, some scholars propose the integrated model with CBR and ELECTRE [11] and TOPSIS [12]. P. Chanvarasuth et al addressed how to use outranking relationship in the ELECTRE III in combination with CBR for a case-based travel advisory system, which offered significantly better ranking result based on Euclidean distance [13]. Hui Li et al used principles of the Electre decision-aiding method in combination with CBR to construct corresponding CBR models, which include Electre-CBR-I and Electre-CBR-II [14]. Hui Li et al. proposed a new multi-criteria CBR approach that is used in a binary business failure prediction (BFP) similar to the positive and negative ideal case (SPNIC) as an implementation of this assumption, subject to order performance technology inspired by TOPSIS [15]. H. Malekpoora proposed a new TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) case-based reasoning (CBR) approach, which captures the oncologist's past experience and expertise to help oncologists to make a better decision between similarity measures, success rate and side effects of treatment [16]. C.K.Kwong and S.M Tam developed CBS-TX, which method can select and adapt some of the closest matching cases by the case retrieval, and similarity analysis, and Case evaluation of order preference based on the similarity of ideal solution (TOPSIS) algorithm [17].

Leveraging the advantages of different technologies in CBR, some scholars begin to use multi-CBR system to solve case attributes aggregation problems. H. Li proposed a financial risk prediction of a CBR based on the majority of voting (Multi-CBR-MV) [18]. Based on the SVM (Multi-CBR-SVM), H. Li developed a new combining-classifiers system, which is called multiple CBR systems. Four uncommitted CBR systems with k-nearest neighbor algorithm are used as classifiers Combination, SVM as a combination of classifier algorithm[19].

3. Case-Based Reasoning with Rank Aggregation

3.1. Methodology Framework

Framework of CBR with rank aggregation is put forward in Fig.1. At first, case library is constructed through case representation which includes target cases and historical cases. Second, case ranking results in each attribute subspace of target case and historical case are obtained by calculating their distance. A ranking matrix $R_{M \times N}$ is constructed. Afterwards, case retrieval problem for multi attributes hybrid data is transformed into a ranking aggregation optimal problem. A new algorithm, RA-CBR, is designed.

3.2. Case Ranking in Each Attribute Subspace

In case-based reasoning system, C_0 is target case. $C_i (i=1,2,\dots,M)$ is the i th historical case. $v_j (j=1,2,\dots,N)$ expresses the j th attribute of case. The historical cases are ranked through calculating its distance from the historical case. This paper gives the sorting method of crisp symbol attribute, interval number attribute and crisp data attribute.

(1) Distance calculating of crisp symbol attribute

v_{0j} expresses a crisp symbol attribute in target case and v_{ij} expresses a crisp symbol attribute historical case. Then, their distance is calculated by formula(1).

$$Dis_1(v_{0j}, v_{ij}) = \begin{cases} 0 & \text{if } v_{0j} = v_{ij} \\ 1 & \text{if } v_{0j} \neq v_{ij} \end{cases} \quad (1)$$

(2) Distance calculating of crisp number attribute

v_{0j} and v_{ij} are numeric attribute value in target case and historical case. Then, their distance is calculated by formula(2).

$$Dis_2(v_{0j}, v_{ij}) = \frac{|v_{ij} - v_{0j}|}{\max v_{ij} - \min v_{ij}} \quad i \in M \quad (2)$$

(3) Distance calculating of interval attribute

$v_{0j} = [v_{0j}^L, v_{0j}^U]$ is the j th interval number in target case. $v_{ij} = [v_{ij}^L, v_{ij}^U]$ is j th interval number attribute value. Then, their distance is calculated by formula(3).

$$Dis_3(v_{0j}, v_{ij}) = \frac{\sqrt{(v_{ij}^L - v_{0j}^L)^2 + (v_{ij}^U - v_{0j}^U)^2}}{\max \left\{ \sqrt{(v_{ij}^L - v_{0j}^L)^2 + (v_{ij}^U - v_{0j}^U)^2} \mid i \in M \right\}} \quad (3)$$

(4) Distance calculating of Fuzzy linguistic variable

v_{0j} is fuzzy linguistic variable, which is the j th feature of target case. v_{ij} is the j th fuzzy linguistic feature in i th historical case. Then, fuzzy linguistic set is expressed as $F = \{f_1, f_2, \dots, f_T\}$. We use triangular fuzzy number to represent the fuzzy linguistic feature. f_i is expressed as $\tilde{f}_i = \{d_i^a, d_i^b, d_i^c\}$, which is calculated as follows[20].

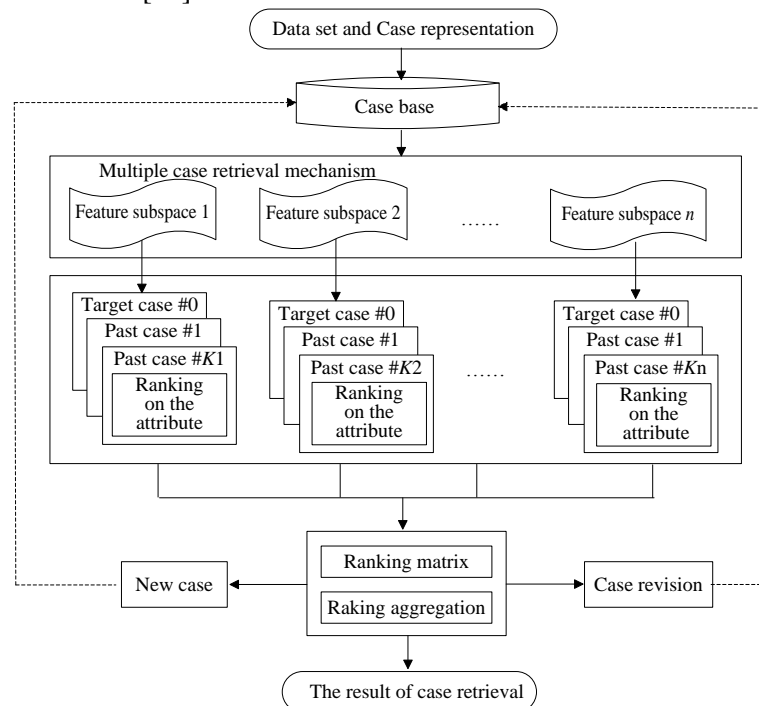


Fig.1 The framework of case-based reasoning

$$\tilde{f}_i = \{d_i^a, d_i^b, d_i^c\} = \{\max[(i-1)/T, 0], i/T, \min[(i+1)/T, 1]\} \quad (4)$$

So v_{0j} can be expressed as $\tilde{v}_{0j} = \{v_{0j}^a, v_{0j}^b, v_{0j}^c\}$, v_{ij} can be represented as $\tilde{v}_{ij} = \{v_{ij}^a, v_{ij}^b, v_{ij}^c\}$.

Their distance can be defined by formula(5).

$$Dis_4(v_{0j}, v_{ij}) = \frac{\sqrt{(v_{ij}^a - v_{0j}^a)^2 + (v_{ij}^b - v_{0j}^b)^2 + (v_{ij}^c - v_{0j}^c)^2}}{\max \left\{ \sqrt{(v_{ij}^a - v_{0j}^a)^2 + (v_{ij}^b - v_{0j}^b)^2 + (v_{ij}^c - v_{0j}^c)^2} \mid i \in M \right\}} \quad (5)$$

According to above equations, if the distance from target case is the closer, the historical case ranking is the better. v_{mj} and v_{nj} express the j th attribute values of m th and n th historical case. The ranking method of historical cases can be calculated by formula(6).

$$Rank(v_{mj}, v_{nj}) = \begin{cases} rank_m \geq rank_n & \text{if } Dis(v_{0j}, v_{mj}) \leq Dis(v_{0j}, v_{nj}) \\ rank_m < rank_n & \text{else} \end{cases} \quad (6)$$

3.3. Ranking Aggregation Theory

Definition 1 K-distance. R_k and $R_l \in R_{M \times N}$ ($k, l = 1, 2, \dots, N$) respectively express the ranking results between cases in k th and l th attribute subspace. If there are two elements in R_k and R_l satisfying $R_{mk} <$

R_{nk} and $R_{ml} > R_{nl}$, which elements are called pairwise disagreement. Then, the number pairwise disagreement in R_k and R_l is K -distance, which is expressed as $K(R_k, R_l)$.

Definition 2 Borda count. Borda count is a voting mechanism method. R_{mk} is a element in R_k . $R_k(m)$ is the voting value, which is the order in R_k . $|R_k|$ is the total number of elements in R_k . Then, the borda count of element R_{mk} is calculated as follows.

$$W_{R_k}(m) = 1 - \frac{R_k(m) - 1}{|R_k|} \quad (7)$$

Definition 3 Condorcet criterion. Each element is compared with other elements in a permutation. If the borda count of one element is higher than another element, it beats that element, which is called condorcet winner. This method is condorcet criterion.

Definition 4 Kemeny optimal. Permutation lists $(R_1, R_2, \dots, R_N) \in R_{M \times N}$, permutation R^* is a new ranking according to the N Rankings. S_R is calculated as follows.

$$S_R(R_1, R_2, \dots, R_N) = \arg \min \sum_{l=1}^N K(R^*, R_l) \quad (8)$$

If $S_R(R_1, R_2, \dots, R_N)$ achieve the minimum, permutation R^* is the Kemeny optimal of lists (R_1, R_2, \dots, R_N) . Kemeny optimal satisfies condorcet criterion. However, the number of permutation lists is more than four, and Kemeny optimal is NP-hard problem. So a locally optimal method is used to solve above Kemeny optimal problem.

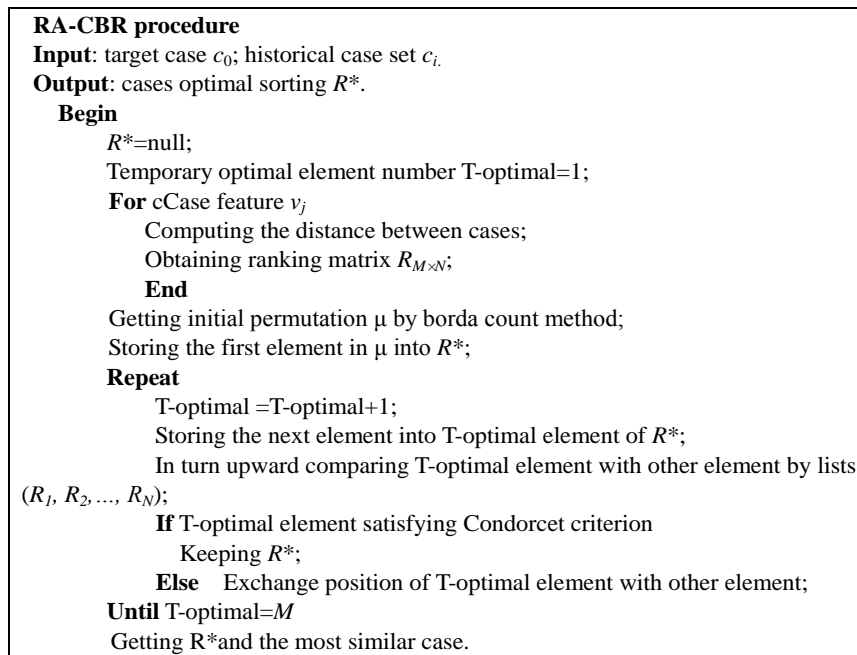
Definition 5 Locally kemeny optimal of permutation lists. The permutation σ does not exist which satisfies $SR(\sigma, R_1, R_2, \dots, R_N) < SR(R^*, R_1, R_2, \dots, R_N)$, when the position of adjacent element pairs is arbitrarily converted. Then, permutation R^* is the locally kemeny optimal.

Definition 6 Permutation consistent. Permutation μ and σ generate from permutation list (R_1, R_2, \dots, R_N) . If $K(\mu, \sigma) = 0$, then μ and σ are consistent.

Definition 7 Locally kemeny optimal of permutation μ . Initial permutation μ is got from (R_1, R_2, \dots, R_N) through borda count method. We say that permutation R^* is locally kemeny optimal of permutation μ , if (1) R^* is consistent with μ and (2) if we restrict attention to the set S consisting of the previous K ($1 \leq K \leq M$) elements in μ , then the projection of R^* onto S is a locally kemeny optimal aggregation of the projections of R_1, R_2, \dots, R_N onto S by condorcet criterion.

4. RA-CBR Algorithm Procedure

The CBR method is changed into a ranking aggregation optimal problem according to above ranking aggregation theory. The greedy algorithm is used to solve the local optimal solution, which is called RA-CBR. The procedure RA-CBR algorithm is shown in Fig.2.

**Fig.2** RA-CBR algorithm

5. Experiment Study and Analysis

In order to validate the effectiveness of RA-CBR method, a test is done on the UCI database in Table 1. Then, five-fold cross validation is adopted. Therein, four folds data are randomly extracted as historical cases and other data are target cases. The experimental computer is Intel Core iu-7500, 16G memory. RA-CBR method is achieved by Matlab 8.0.

Table 1 Testing data sets

Data sets	Attribute number	Record number	Class number
adult	14	16281	2
car	6	1728	4
breast-cancer	9	286	2
post-operative	8	90	3

5.1. The Algorithm Accuracy Rate

The experimental results are shown in Table 2. The average accuracies on the four data sets are 90.11%, 91.76%, 80.40% and 84.44%. The accuracies of RA-CBR algorithm for adult and car data sets are higher than post-operative and breast-cancer data sets. The testing results show that RA-CBR algorithm has a high accuracy rate.

Table 2 The experimental results accuracy rate (%)

<div style="text-align: center;"> <div style="transform: rotate(-45deg); display: inline-block;">Experiment</div> <div style="display: inline-block;">Data sets</div> </div>	Test 1	Test 2	Test 3	Test 4	Test 5	Average
adult	90.60	90.29	89.07	91.83	88.75	90.11
car	91.59	90.43	92.46	91.88	92.46	91.76
breast-cancer	78.94	82.46	80.70	77.19	82.70	80.40
post-operative	88.89	83.33	77.78	88.89	83.33	84.44

5.2. Performance Comparison

In order to test the performance of RA-CBR algorithm, Euclidean distance CBR and Manhattan distance CBR are compared with RA-CBR. The subjective weight allocation method is used for

ECBR and MCBR. The experimental results show that RA-CBR has higher accuracy than ECBR and MCBR, which are shown as Fig 3.

6. Conclusions

The rank aggregation theory is firstly used in CBR system in this paper, and a new ensemble framework for CBR is put forward. In this framework, the case ranking method in each attribute subspace is proposed, and the connotation and related definition of ranking aggregation are given. On the basis, the similar case retrieval process is transformed into a rank aggregation optimal solving problem. At the same time, RA-CBR algorithm is designed, which effectively settles the problem of determining attribute weights in the case retrieval process. The experiment results show that RA-CBR method is feasible and efficient. To extend application scope of RA-CBR algorithm, it is an important research field for how to apply this method on real dataset.

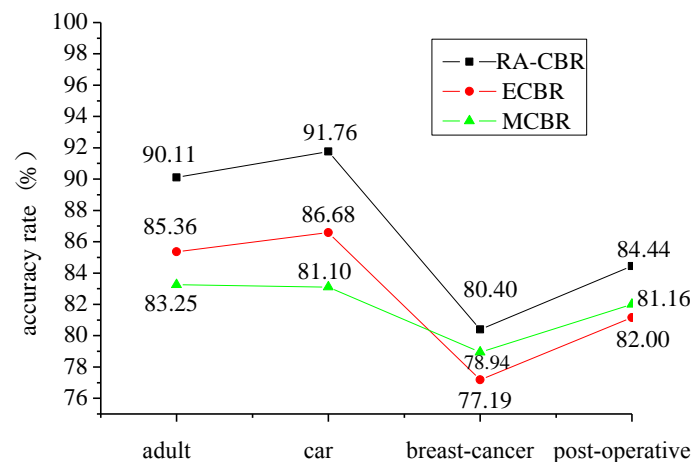


Fig.3 Comparison of different algorithms

Acknowledgment. This work was partially supported by the National Natural Science Foundation of China (Grant No. 71401021, Grant No. 71301181), by the Humanities and Social Science Project of Chongqing Municipal Education Commission(16SKGH143), by the Science and Technology Project of Chongqing Municipal Education Commission (KJ1600939), by Chongqing Social Science Federation planning project (2013YBGL125).

References

- [1] Aamodt, E. Plaza Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, vol.7, no.1, pp.39-59,1994
- [2] B.S.ZHANG, Y. L.YU. Hybrid similarity measure for retrieval in case-based reasoning system. *Systems Engineering-theory & Practice*, no. 3, pp. 131-136, 2009
- [3] Z.P. Fan, Y.H. Li. Hybrid similarity measure for case retrieval in CBR and its application to emergency response towards gas explosion. *Expert Systems with Applications*, Vol. 41, no. 5, pp. 2526–2534, 2014
- [4] W.L. Chang, A CBR-based Delphi model for quality group decisions. *Cybernetics & Systems*, Vol.42 ,no.6, pp.402–414, 2011
- [5] Liang, D.Gu, I. Bichindaritz, X.Li, C.Zuo, W. Cheng. Integrating gray system theory and logistic regression into case-based reasoning for safety assessment of thermal power plants. *Expert Systems with Applications*, Vol. 39, no. 5, pp. 5154–5167, 2012
- [6] J.Qi, J.Hu, Y.H. Peng. Hybrid weighted mean for CBR adaptation in mechanical design by exploring effective, correlative and adaptative values. *Computers in Industry*, Vol 75, pp.58-66, 2016
- [7] K.Zhao, X. Yu, A case based reasoning approach on supplier selection in petroleum enterprises. *Expert Systems with Applications*, Vol.38, no.6, pp.6839–6847, 2011
- [8] S. Ha, A personalized counseling system using case-based reasoning with neural symbolic feature weighting (CANSY). *Applied Intelligence*, Vol.39 , no.3, pp. 279–288, 2008
- [9] H. Ahn, K. Kim, Global optimization of case-based reasoning for breast cytology diagnosis. *Expert*

- Systems with Applications, Vol.36, no.1, 724–734,2009
- [10] A.J.Yan, H.S.Shao, Z.Guo. Weight optimization for case-based reasoning using membrane computing. Information Sciences, Vol. 287, pp. 109–120, 2014
 - [11] M. Tavana, Z. Li, M. Mobin, M. Komaki, and E. Teymourian, Multi-objective control chart design optimization using NSGA-III and MOPSO enhanced with DEA and TOPSIS, Expert Systems with Applications, 2016, (50), pp. 1739.
 - [12] S. Corrente, S.Greco, R. Słowiński, Multiple criteria hierarchy process for ELECTRE Tri methods, European Journal of Operational Research, 2016, 252(1), pp. 191-203.
 - [13] P. Chanvarasuth, L. Boongasame, Remove from marked Records Hybridizing principles of the ELECTRE III method with case-based reasoning for a travel advisory system: case study of Thailand, Asia Pacific Journal of Tourism Research, 2015, 20(5), pp. 585-598.
 - [14] H. Li, J. Sun, Hybridizing principles of the Electre method with case-based reasoning for data mining: Electre-CBR-I and Electre-CBR-II, European Journal of Operational Research, 2009, 197(1), pp. 214-224.
 - [15] H. Li, H. Adelib, J. Sun, J.G. Han, Hybridizing principles of TOPSIS with case-based reasoning for business failure prediction, Computers & Operations Research, 2011, 38(2), pp. 409-419.
 - [16] H. Malekpoora, N. Mishrab*, S. Sumalyac, S. Kumarid, An efficient approach to radiotherapy dose planning problem: a TOPSIS case-based reasoning approach, International Journal of Systems Science: Operations & Logistics, 2016.
 - [17] C.K. Kwong, S.M. Tam, Case-based reasoning approach to concurrent design of low power transformers, Journal of Materials Processing Technology, 2002, 128(1-3), pp. 136-141.
 - [18] H. Li, J. Sun, Majority voting combination of multiple case-based reasoning for financial distress prediction. Expert Systems with Applications, 2009, 36(3), pp. 4363-4373
 - [19] H. Li, J. Sun, Predicting business failure using multiple case-based reasoning combined with support vector machine, Expert Systems with Applications, 2009, 36(6), pp. 10085-10096.
 - [20] A. Skeete, M.Mobin. Aviation Technical Publication Content Management System Selection Using Integrated Fuzzy-Grey MCDM Method. In Proceedings of the 2015 Industrial and Systems Engineering Research Conference, 2015: 1-10