Neurocomputing 149 (2015) 1245–1252

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An improved case-based reasoning method and its application in endpoint prediction of basic oxygen furnace

Min Han*, ZhanJi Cao

Dalian University of Technology, Dalian 116023, China

ARTICLE INFO

Article history:
Received 3 March 2014
Received in revised form
27 April 2014
Accepted 5 September 2014
Communicated by T. Heskes
Available online 16 September 2014

Keywords:
Case-based reasoning
Fuzzy *c*-means clustering
Mutual information
Support vector machine
BOF

ABSTRACT

Case retrieval and case revise (reuse) are core parts of case-based reasoning (CBR). According to the problems that weights of condition attributes are difficult to evaluate in case retrieval, and there are few effective strategies for case revise, this paper introduces an improved case-based reasoning method based on fuzzy *c*-means clustering (FCM), mutual information and support vector machine (SVM). Fuzzy *c*-means clustering is used to divide case base to improve efficiency of the algorithm. In the case retrieval process, mutual information is introduced to calculate weights of each condition attribute and evaluate their contributions to reasoning results accurately. Considering the good ability of the support vector machine for dealing with limited samples, it is adopted to build an optical regression model for case revise. The proposed method is applied in endpoint prediction of Basic Oxygen Furnace (BOF), and simulation experiments based on a set of actual production data from a 180 t steelmaking furnace show that the model based on improved CBR achieves high prediction accuracy and good robustness.

1. Introduction

As a branch of artificial intelligence, CBR receives more and more scholars and experts' attentions and has a rapid development. CBR solves problems by reusing solutions to similar problems stored as cases in a case base [1]. It is applicable to the fields with no accurate mathematical models, but with rich experiences and historical cases, like steelmaking industry. It can realize incremental learning and has a good interpretability [2]. A typical CBR process is mainly composed of four parts, namely case representation, case retrieval, case reuse and case retain.

A problem can be represented as a case by extracting some attributes to describe it. The description of a case in this paper contains condition attributes and solution attributes. Condition attributes reflect characteristics of the problems, and solution attributes record the solving methods. The basic assumption of CBR is that if the condition attributes in problem space are close, the corresponding solution attributes will also be close to each other in solution space. Case retrieval is the core content of case-based reasoning method, and its purpose is to find the cases with high similarity to the problem in case base quickly [3]. Scale of the case base and weights of condition attributes often have a great influence on retrieval results.

First, too large a scale of case base will lower the efficiency because every case in the case base is needed to calculate similarity to purpose problem individually. One way to improve retrieval accuracy and efficiency is to divide the case base, and retrieval is conducted in subcase bases. Many algorithms like fuzzy clustering have been introduced in dividing the case base. Wang [4] had introduced a rough set to build some equivalent classes of "problems and solutions" to improve retrieval efficiency. Self-Organization Map (SOM) and Hierarchical Self-Organization map (GHSOM) [6] were used to map the case base into a two dimensional coordinate system, and position relationship between purpose and source cases can be seen directly in the maps. Mata [5] used Growing Cell Structure (GCS) to organize case base. Besides that, Directed Acyclic Graph Support Vector Machine (DAGSVM) and association rules were introduced in preprocessing and analysis of case base [7]. These methods had achieved good results for solving a certain problem, and they needed to be consistent with the corresponding retrieval methods. However, efficiency of algorithm is a major bottleneck and the improved case base must be convenient to maintain.

Second, weights of condition attributes directly affect results of similarity computation. Similarity is determined by the consistency of condition attributes, and the importance of each attribute is different. For example, some attributes are critical factors in distinguishing each case, and they are assigned with heavy weights in order to increase their effect for calculation results. In general, the weights are assigned according to experts' experience in corresponding field. But it is often objective and limited, and the

* Corresponding author. Tel.: +86 411 84708719.
E-mail address: minhan@dlut.edu.cn (M. Han).

http://dx.doi.org/10.1016/j.neucom.2014.09.003
0925-2312/

degree of automation is low. While the method based on statistics of attributes values, like mean square error, is restricted by the sampling data. Furthermore, some experts and scholars introduce a rough set in weights determination [8,9], but attributes need discrete processing and will lose some useful information. Besides that, we can only achieve discrete values for weights after encoding and decoding process of genetic algorithm [10]. The method in which weight matrices are calculated as parameters to be optimized by optimization methods, like Particle Swarm Optimization (PSO) method, achieves good results, but its generalization performance needs consideration [11]. Juell introduced a method to update weights based on feedback of calculation deviation [12]. Hyuk [13] adapted neural network to calculate weights and Jinxiang Pian proposed a self-learning method in weights determining and it was applied successfully in BOF. The methods mentioned above are widely used in similarity calculation, but there are still some shortcomings and limitations, so it is very necessary to design an efficient method for weights calculation to improve validity of case retrieval, and guarantee attributes' physical meaning at the same time. In addition, under the premise of reasonable and effective case representation, attribute reduction is another means to improve efficiency of retrieval, and it can be seen as a special situation of weights determination. The reduced attributes are assigned with weights equal to zero in view of similarity calculation.

Third, since retrieved cases cannot be completely consistent with target problems, the solution attributes need to be adjusted and reused [14]. The common method is to calculate weighted average of history solutions according to similarity. It is also a good way to extract rules for case revise by data mining technique [15], and the rules are added to each case as a special attribute, but this will increase the burden on database storage. Besides that, the algorithm like expert system and neural networks are introduced to establish revise models based on training data, but it needs enough history data with high similarity [16]. However, there are often a small number of similar cases, and processing capabilities of limited samples are very necessary. Knuth–Morris–Pratt (KMP) pattern matching algorithm was applied in the case revise and case retain, and it had better performance in documenting object-oriented application frameworks [17]. Under certain conditions, the above methods can achieve satisfied results, but efficiency and validity of the algorithm for case revise are critical problems that need to be solved.

In response to above problems, this paper introduces an improved case-based reasoning method based on fuzzy clustering, mutual information (MI) and support vector machine (SVM). The improved CBR method can be called Hybrid_CBR for short. Fuzzy c-means clustering (FCM) is a widely used method for cluster analysis with low computational complexity and high computing speed. It is applied to divide case base into sub-categories where case retrieval is conducted. Mutual information is capable for estimating a general dependence between two variables. It is a relatively reasonable method to fully mine the implication information between condition and solution attributes, estimate weights and realize attributes reduction, improving the validity of case retrieval. In addition, the model for case revise is established based on SVM to make full use of retrieved similar cases. SVM has processing capacity for small sample data sets and high robustness. It is suitable for application in process of case revise. In the end, the improved CBR model is applied in endpoint carbon content prediction of basic oxygen furnace. Experiments based on actual production data are conducted to test and verify effectiveness of the proposed method.

This paper is organized as follows. Section 1 covers relevant work on CBR, while Section 2 describes the improved CBR method and corresponding algorithm in each part of case-based reasoning. In this section, fuzzy c-means clustering, mutual information and support vector machine are introduced Section 3 introduces simulation experiments, including background, experimental conditions and obtained results. Finally, Section 4 presents the conclusion and further work.

2. Hybrid case-based reasoning algorithm

Case-based reasoning is an analogical reasoning method which breaks through the bottleneck that domain knowledge is difficult to describe and extract accurately in rule-based reasoning (RBR). CBR implies the knowledge and experience in cases, then retrieves and reuses the similar ones for reasoning [18]. Case retrieval and case reuse are the most important parts that affects reasoning results, so the improvements focus on these two parts. In case retrieval process, fuzzy c-means clustering is introduced to divide the case base into small ones. At first, we need to calculate similarity between purpose case and each clustering center. The similarity values are sorted in descending order and then retrieve similar cases in the cluster with largest similarity. In the stage of similarity calculation, attributes weights are estimated by mutual information between each condition attribute and solution attribute. The condition attributes with least mutual information values can be seen as redundant ones and be reduced. In case reuse process, SVM is used as the reuse model. The retrieved similar cases are chosen as training samples to build a regression model for case reuse after k-cross-validation, and solution for purpose case can be inferred through the proposed regression model. Three methods mentioned above are utilized together to improve effectiveness and robustness of CBR, and the flow diagram of the improved CBR is shown in Fig. 1.

As shown in Fig. 1, case base plays an important role in CBR. There are often many cases stored in database, so the computation loads are heavy when computing similarity with every case in the database. Fuzzy c-means clustering method can solve this problem by dividing case base, and choose only one certain sub-base. It can low the calculation amount in a great degree. In Fig. 1, B1, B2 and Bc representative different sub-bases where c is the number of clusters. The introduction of mutual information solves the

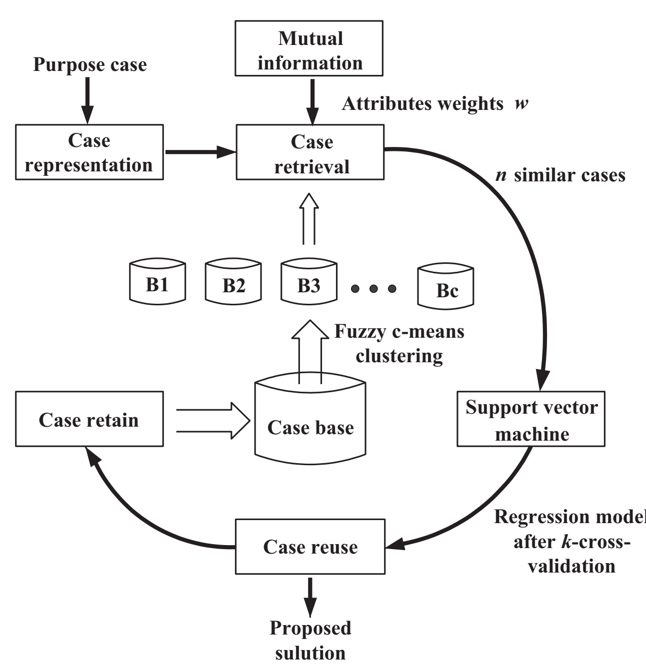

Fig. 1. Flow diagram of the improved CBR.