Yanrui LIANG,1,2) Hongbing WANG,1,2)* Anjun XU3) and Naiyuan TIAN3)

1) School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, 100083 China.
2) Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, 100083 China.
3) School of Metallurgy and Ecological Engineering, University of Science and Technology Beijing, Beijing, 100083 China.

(Received on December 1, 2014; accepted on February 17, 2015)

Case-Based Reasoning (CBR) is a kind of solving paradigm based on the past successful cases to get the solution for the current problem. When CBR is applied in complex industrial processes, solving efficiency is often not high due to too many influence factors involved. So it is necessary to reduce the number of attributes involved in CBR system for the fast modern industrial production, such as steel-making and continues casting process. A two-step CBR method is proposed for predicting the endpoint phosphorus content in BOF efficiently. First, the genetic algorithm is applied to find the optimal attributes subset based on the evaluation method of Correlation-based Feature Selection (CFS). Then, CBR system is applied for solving this problem with the reduced attributes. There are two kinds of similarity calculation method based on the euclidean distance and the gray distance, and two kinds of the weight decision method based on the even weight and the entropy weight for this CBR system. Four groups of experiment results show that the two-step CBR method has much more efficiency than the single CBR method, while maintaining almost the same prediction precision. The two-step CBR method can be used in the fast industrial process more efficiently.

KEY WORDS: attributes reduction; correlation-based selection; information gain; case-based reasoning; genetic algorithm.

1. Introduction

During the metallurgical process in BOF (Basic Oxygen Furnace), hot metal at about 1 350°C is converted into molten steel at about 1 650°C by a series of oxidation reaction of elements dissolved in it. This process consists of several steps:

1) Scrap and hot metal are charged into the furnace. Generally, hot metal should be preprocessed for desulfurization and desilication.

2) Oxygen is blown into the furnace through lance. Meanwhile, lime, dolomite and other auxiliary materials are added. Oxygen is reacted with the elements such as carbon, silicon and manganese to remove the impurities and rise the temperature continuously.

3) At a predetermined point in the blowing (about 85% of the whole blowing period), TSC (Temperature, Sampling and Carbon) detector is inserted into the molten steel to measure the temperature and the carbon content and take a sample. Comparing the measured carbon content with expected endpoint, the amounts of oxygen and coolant for the second blow period are calculated.

4) The blowing continues until the second blow period ends. Then, the endpoint temperature and carbon content of molten steel are measured by TSO (Temperature, Sampling and Oxygen) detector. If the endpoint contents are acceptable, the molten steel will be tapped. Otherwise, extra operations should be taken until steel components and temperature meet the technical standards.

With the development of iron and steel industries, the quality of steel is attracting increasing attention. The phosphorus in molten steel is the main influencing factor for cold brittleness, which means that steel products are easy to break down at the normal or low temperature. Therefore, dephosphorization has become one of the main functions in BOF. But, the phosphorus content can not be directly measured by TSC or TSO detector in the metallurgical process. The data about the phosphorus content can only be derived from the steel sample. The transfer and measurement may take several minutes. This delay will break the continuous production in BOF. Therefore, it is very important to develop the methods to predict the endpoint phosphorus content in BOF.

Case-Based Reasoning (CBR) is the process to use the existing experiences for solving the current problem, which has some similarity with the past case or cases.1) “A Case-based reasoned solves new problems by adapting solutions that were used to solve old problems”, the classical definition about CBR is put forward by Schank2) in 1989. The cases in the CBR system are often represented by many attributes. In the research field about the big data technology, the num-
number of attributes has increased a lot and it takes much time in the computation for case retrieval. The process of case retrieval is mainly focused in this paper. In fact, not all attributes are related to the problem target. A large number of invalid or unrelated attributes could be considered to reduce the solution efficiency greatly. Therefore, attributes reduction is essential to improve the solution efficiency of CBR, especially in the fast modern industrial production.

Related researches are as follows. Hammond3) designed CHEF in 1986 and CHEF included failures as part of the case and acquired reversible measures when failures occurred. Kyung-shik Shin4) utilizes a hybrid approach using genetic algorithms (GA) to case-based retrieval process in an attempt to increase the overall classification accuracy; Hui Li5) introduced outranking relations (OR), including strict difference, weak difference, and indifference, between cases to build up a new feature-based similarity measure mechanism in the principle of k-nearest neighbors; Aijun Yan proposed an Improved CBR-based fault Prediction method (ICBRP) to predict faults more accurately. This ICBRP is composed of a water filling theory-based weight allocation model and a Group Decision-Making based Revision (GDMR) model;6) Zhou Huifeng7) put forward a kind of fault diagnosis method which combined CBR method with fault tree analysis and this method effectively implemented the fault detection of the outside trajectory measurement and self-destruction system of cosmodrome; ShenJing8) used a case matching algorithm based on the similarity threshold combined with expert system to develop the disease diagnosis expert system; Chen Rui9) developed the disease diagnosis expert system; Dong Chuanheng11) gave the application of navigation equipment fault using CBR, in which the similarity between cases is computed based on hamming code; Ai Lixiang12) put forward the CBR technology for the diagnosis of Argon Oxygen Decarburization (AOD) furnace energy consumption. In this CBR system, the weights of influencing factors is obtained by Analytic Hierarchy Process (AHP), the similarity between cases is calculated based on the gray distance.

Most of the researches above combined CBR with other techniques to improve the performance from the viewpoint of case retrieval. Currently, big data technology has attracted many people’s attention and the characteristic of high dimension which will affect the efficiency of case storage, case retrieval and the whole system should be considered. So it is necessary to reduce the number of attributes involved in CBR system for the fast modern industrial production, such as steel-making and continues casting process.

Attributes reduction is the process of searching the optimal subset of attributes. There have been a lot of algorithms. Kai Zheng13) proposed an Enhancement for Heuristic Attribute Reduction (EHAR) in rough set; Yanyan Yang14) developed two algorithms to find all β-distributed reducts based on the algorithm of finding all minimal elements; Jin Qian15) design a novel structure of key-value pair to speed up the computation of equivalence classes and attribute significance and parallelize the traditional attribute reduction process based on MapReduce mechanism. Chang Chunguang16) put forward a kind of reduction technique for case attributes based on rough set theory to improve the efficiency of case retrieving in CBR; Li Fenggang17) proposed the attribute selecting strategy on the basis of the attribute-oriented reduction techniques; Wu Zhengjiang18) gave an attribute reduction algorithm based on GAs and discernable matrixes which can calculate the weak reduction easily and the weak reduction effectiveness is evaluated through k-nearest neighbor accuracy; Yaoxu19) and Mao Yong20) introduced the advantages and disadvantages of several methods attributes selection algorithm.

In this paper, the genetic algorithm is applied to find the optimal attributes subset based on the evaluation method of Correlation-based Feature Selection (CFS). Here, GA has the global search ability so as to avoid the local optimization. CFS is based correlation, which is naturally mapped into the influence on the target in the problem to be solved.

In this paper, the solution efficiency in CBR is focused from the point view of attributes reduction. The two-step CBR method based on attributes reduction is proposed and verified by predicting the endpoint phosphorus content in BOF.

2. Two-step Case-based Reasoning System

The workflow of the two-step CBR system is shown in Fig. 1. First, GA is applied to find the optimal attributes subset based on the evaluation method of CFS. Then, CBR system is applied for solving this problem with the reduced attributes.

2.1. Case Representation

There are many kinds of case representation methods in CBR system and the attribute-value method was used to represent the cases in this paper. The attribute-value method is a kind of representation method which is simple to storage, easy to understand, efficient to retrieval and relatively easy to implement. There are two main case representation methods: the pure-case representation and the case-based representation. In this paper, the pure-case representation method is adopted. The case-based representation method is a kind of representation method which is simple to storage, easy to understand, efficient to retrieval and relatively easy to implement. There are two main case representation methods: the pure-case representation and the case-based representation. In this paper, the pure-case representation method is adopted.

Fig. 1. Workflow of the two-step CBR system.
to code. The attribute-value method is usually used in computer system.

### 2.2. Attribute Reduction

CBR system without the attributes reduction can also achieve its prediction, but too much case attributes may reduce the speed of matching computation and take a long time for the system to respond, which seriously influenced the fast pace of industrial production. There are two mainly processes about attribute reduction, attribute subset produce process and evaluate attribute subset. The CFS method and GA method are combined in this paper to reduce the number of attributes. GA is used to attribute subset produce process, and the CFS method is used to evaluate the subset.

Attribute subset produce process is the process of searching feature subset, and the process is responsible for providing feature subset for the evaluation function. Many search algorithms have been used for feature selection. GA is an optimization method developed by Holland.\(^{21}\) GA is a method based on the principle of the evolution of Darwin’s. It uses the basic genetic operation repeatedly to the group containing the feasible solution, in order to generate a new group constantly and makes the population evolve. At the same time, GA searches for best individual to meet the requirements of obtaining the optimal solution. GA can realize parallel global search with large search space, and it can also adjust the search space to the direction containing the optimal solution in the process of exploring. So GA not only can be conducive to avoid the local minimum points and get the optimal solution or near optimal solution, but also can search more efficiently. Therefore, GA has been proven to be an effective computational method and used in many fields. Sun used GA to deal with high-dimensional data clustering.\(^{22}\) Zhao proposed improved genetic algorithm for text feature selection.\(^{23}\) Tsai uses genetic algorithms in feature and instance selection.\(^{24}\) GA is a global optimization algorithm and it is often used for feature selection.

Feature subset evaluation methods can be divided into the filter and wrapper. The methods based on correlation, distance, information gain and consistency belong to the filtering methods and the classifier error rate belongs to wrapper. In this paper, evaluation methods based on CFS is used. CFS is a method based on information gain. CFS is a classic filter feature selection algorithm which considering the relationship between the attributes. The greater correlation between each attribute and the class attribute and the smaller attribute redundancy between the attributes will get the higher value. The correlations are calculated by information gain (IG).

Entropy is a physical quantity expressing chaotic degree of the system, the more chaos the system is the higher its entropy is; According to information theory, entropy mainly adopts numerical form to express the uncertainty of a random variable values in order to score the result of information content.\(^{25}\) In recent years, a lot of research on the information entropy, and widely applied to feature selection.\(^{26-31}\)

The CFS method is as follows:

$$M_n = \frac{m\bar{\tau}_a}{\sqrt{m + m(m-1)\bar{\tau}_a}}$$

Here, \(M_n\) refers to the value of the subset with \(m\) attributes and \(n\) refers to the number of the subset, \(\bar{\tau}_a\) refers to the average correlation between attributes and categories, \(\tau_c\) refers to average correlation between different attributes. The correlation between \(X\) and \(Y\) is computed by the information gain, which is been normalized.

The basic principle of information gain is assuming that there is a discrete variable \(Y\), values of \(Y\) are included in \(\{y_1, y_2, ..., y_m\}\), probability for \(y_i\) is \(P_i\). Information entropy of \(Y\) is defined as

$$H(Y) = -\sum_{i=1}^{m} P_i \log_2 P_i$$

In the condition that attaching to another variable \(X\), and \(X = x_o\), conditional information entropy of \(Y\) is expressed as

$$H(Y \mid X) = \sum_{i=1}^{m} P_{X=x_i} H(Y \mid X = x_i)$$

Knowing that before and after \(X\), information gain of \(Y\) is defined as

$$IG(Y \mid X) = H(Y) - H(Y \mid X)$$

To ensure that the various attributes can compare with each other and make different properties to have the same effect, information gain should be normalized as follows.

$$U_{XY} = 2.0 \times \frac{IG(Y \mid X)}{H(Y) + H(X)}$$

This principle above is applied to evaluate attribute subset, evaluation values of different attributes subsets will be compared. The attribute subsets which can score high would be taken as the next generation. Attributes subset selection process is as shown in Fig. 2.

![Fig. 2. Workflow of selecting attribute subset.](image_url)
2.3. Normalization

In CBR system, normalized processing methods are often used to make influence factors fully demonstrate their proportion. Normalization not only can prevent one or several dimensions of data to have strong influence when the data in high dimension, but also can simplify the calculation and speed up the case retrieval. In this CBR system, normalization of \([0, 1]\) is used to work on cases and problems.

2.4. Similarity Calculation Method

In this paper, there are two methods for the calculation of similarity: based on euclidean distance and based on gray distance. In CBR system, the problem should be matched with the stored cases in order to find the most similar case or several similar cases. Here, the result of the most similar case can be as a result of problem. There are two kinds of methods to calculate the similarity.

1) The similarity calculation method based on euclidean distance. The data should be normalized before calculating the similarity. Assuming that the attributes of case is a n dimensional vector, \(X_j\) refers to the first \(j\) attribute of the cases stores in the case base, \(Y_j\) refers to the first \(j\) attribute of the cases stores in the problem base, \(W_j\) refers to the weight of the first \(j\) attribute, and \(W_1 + W_2 + \cdots + W_j + \cdots + W_n = 1\). So formula of the distance between the case in case base and problem is as follows:

\[
d(X,Y) = \sum_{j=1}^{n} W_j \left( Y_j - X_j \right)^2
\]

The similarity between the stored case in case base and problem is as follows:

\[
G_{sim}(X,Y) = \frac{1}{1 + d(X,Y)}
\]

2) The similarity calculation method based on gray distance. Assuming that \(S_0\) is a problem case, the stored case base is \([S_i = 1, 2, 3, \ldots, m]\), \(m\) is the number of the stored cases. The formula of grey distance between \(S_0\) and \(S_i\) is as follows:

\[
G_{sim}(S_0, S_i) = \frac{1}{1 + \sqrt{\sum_{j=1}^{m} G^2_{sim}(S_0(j), S_i(j))}}
\]

Here, \(G_{sim}(S_0(j), S_i(j))\) refers to the grey distance between case \(S_0\) and \(S_i\) in the first \(j\) attribute.

2.5. Weight Decision Methods

Different attributes may have different influences on the target, so the weights of attributes need to be considered. In this CBR system, two methods of calculating the attribute weights are adopted: the even weight method and the entropy weight method.

1) Even weight method. In this weight method, assuming that there are \(n\) attributes. The weight of attribute \(j\) is \(W_j = \frac{1}{n}\).

2) The entropy weight method. Entropy weight method is a kind of objective value assignment method, and it calculates the attribute weight according to degree of variation. The detailed process of entropy weight is shown in the paper written by Wang Guowei.25)
ingle-step CBR and the two-step CBR. Here, the single-step CBR is the CBR without attribute reduction. The two-step CBR includes the first step of attribute reduction and then the second step of case retrieval.

3.2. Results Comparison about the Predicting Hit Percentage

The differences between predicting and actual values are used to illustrate the hit percentage of predicting the endpoint phosphorus content. The distribution of the hit percentage is shown in Figs. from 4 to 11 and Tables from 2 to 3. Figures from 4 to 11 respectively show the distribution about the hit percentage of predicting endpoint phosphorous content using different types of CBR methods. The abscissa represents the error interval between the predicting value and the actual value. The ordinate shows the hit percentage of predicting endpoint phosphorous content. For example, Fig. 4 shows the hit percentage of predicting endpoint phosphorous content using single-step CBR with even weight attribute and similarity based on grey distance. In the error interval [0 ppm, 10 ppm], the hit percentage of predicting endpoint phosphorous content is 13.25%. Table 2 and Table 3 show the comparison of predicting hit percentage between the single-step CBR and the two-step CBR. For the error interval [–10 ppm, 10 ppm] in Table 2, the hit percentage is 25% using the single-step CBR with even weight attribute and similarity based on grey distance and the hit percentage is 22.25% using the two-step CBR with even weight attri-
but and similarity based on grey distance. The difference of hit percentage using these two types of CBR methods is 2.75% in the error interval [–10 ppm, 10 ppm]. All the figures show a standard normal distribution. Both single-step and two-step CBR can effectively predict the endpoint phosphorous content in BOF. In the error interval [–40 ppm, 40 ppm], both single-step and two-step CBR can get more than 62.00% hit percentage.
Sometimes, the single-step CBR gets a little higher hit percentage of predicting than the two-step CBR. Sometimes, the two-step CBR has a little higher hit percentage of predicting than the single CBR. For example, for the similarity calculation method based on grey distance and the weight decision method based on entropy weight, the two-step CBR has got 2.25% higher hit percentage of predicting in the error interval [–30 ppm, 30 ppm] than the single-step.

Fig. 9. Percentage of predicting endpoint phosphorous content hit using two-step CBR with entropy weight attribute and similarity based on grey distance.

Fig. 10. Percentage of predicting endpoint phosphorous content hit using single-step CBR with entropy weight attribute and similarity based on euclidean distance.

Fig. 11. Percentage of predicting endpoint phosphorous content hit using two-step CBR with entropy weight attribute and similarity based on euclidean distance.
CBR. But, in the error interval [-20 ppm, 20 ppm], the single-step CBR has got 2.50% higher hit percentage than the two-step CBR. The difference of predicting hit percentage between the two-step CBR and the single-step CBR is less than 3.75%. So, the two-step CBR almost gets the same predicting precision as the single-step CBR.

### 3.3 Results Comparison about the Running Time

The advantage of two-step CBR is the reduction of the running time. This has much significance for the fast modern industrial production. The comparison about the running time of predicting the endpoint phosphorus content is shown in Table 4. Here, the predicting for the 400 problems is done for 10 times to get the average time for the 400 problems and then get the average time for the single problem after a division calculation. The average running time of each problem using the two-step CBR can be reduced about 30% comparing with the single-step CBR.

The two-step CBR method can be used in the fast industrial process more efficiently, such as the steel-making process and the continuous casting process. The reduction in running time of the two-step CBR has much real significance for fast modern industrial process.
4. Conclusion

In this paper, the two-step CBR method is presented and its feasibility is verified by predicting the endpoint phosphorus content in BOF. First, the optimal attribute subset is obtained by GA with the evaluation method of CFS. The original 21 attributes for predicting the endpoint phosphorus content can be reduced into 11 attributes. Four groups of experiment results show that the two-step CBR method has much more efficiency than the single CBR method, while maintaining the same prediction precision. The two-step CBR method has much significance for the fast industrial process.

Further research is required to develop a new CBR model to improve the hit percentage of predicting the endpoint phosphorus content in BOF to realize the real industrial application.

Acknowledgements

This is financially supported by Beijing Higher Education Young Elite Teacher Project (YETP0382), 2012 Ladder Plan Project of Beijing Key Laboratory of Knowledge Engineering for Materials Science (Z121101002812005) and Fundamental Research Funds for the Central Universities (FRF-TP-14-047A2).

REFERENCES