An integrated feature selection and cluster analysis techniques for case-based reasoning

Guo-Niu Zhu, Jie Hu*, Jin Qi, Jin Ma, Ying-Hong Peng

State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, PR China

ARTICLE INFO

Article history:
Received 21 April 2014
Received in revised form
9 November 2014
Accepted 13 November 2014

Keywords:
Case-based reasoning
Feature selection
Cluster analysis
Case organization

ABSTRACT

Feature selection and case organization are crucial steps in case-based reasoning (CBR), since the retrieval efficiency and accuracy even the success of the CBR system are heavily dependent on their quality. However, inappropriate feature selection and case selection together with ill-structured case organization may not only present a dilemma in case retrieval, but also greatly increase the case base. To obtain an efficient CBR system, selection of proper features and suitable cases with appropriate case organization are very important. This paper proposes a hybrid CBR system by introducing reduction technique in feature selection and cluster analysis in case organization. In this study, a minimal set of features is selected from the problem domain while redundant ones are reduced through neighborhood rough set algorithm. Once feature selection is finished, the growing hierarchical self-organizing map (GHSOM) is taken as a cluster tool to organize those cases so that the initial case base can be divided into some small subsets with hierarchical structure. New case is led into corresponding subset for case retrieval. Experiments on UCI datasets and a practical case in electromotor product design show the effectiveness of the proposed approach. The results indicate that the research techniques can effectively enhance the performance of the CBR system.

1. Introduction

Case-based reasoning (CBR) is an approach to problem solving by utilizing previous cases and experiences which are similar to the current one (Kolodner, 1993). It supposes that similar problems usually have similar solutions, and those problems may often take place. In CBR system, knowledge and previous experiences are stored as cases in a case memory (also known as case base or library). Typically, each case in the case memory is composed of two parts: one is the problem description part which illustrates the situation when the case happens, and the other is the solution description one which denotes the outcome of the corresponding problem. By similarity measurement between the current situation and the previous cases, a proposed solution is obtained for present problem according to the solution part of the similar case. Due to its unique way of knowledge representation, easy to implement and provision of an appropriate explanation for the output, CBR has been utilized in many areas, especially those complex and unstructured problems such as manufacturing, finance and medicine (Guo et al., 2012; Xie et al., 2013; Chuang, 2013; Ahn and Kim, 2009a,b; Ahmed et al., 2011).

However, traditional CBR is too much reliant on experts' experiences, especially at the early stages, such as feature selection and case organization (Guo et al., 2011). With the rapid development of CBR, large scale case base is becoming more common, with the number of instances ranging from thousands to millions (Cao et al., 2001). Although an enormous case base may improve the coverage of the problem domain, it also causes problems of retrieval efficiency and brings many noisy cases. How to select appropriate features and organize the case base to optimize the CBR system become critical issues. In past decades, a lot of studies have conducted to promote the performance of similarity measurement, but few focus on other aspects that can be further combined to strengthen the retrieval performance (Kang et al., 2014). They mainly pay attention to the feature weighting and similarity measurement techniques partially or simultaneously, ignoring the problems of feature selection and case organization, although their simultaneous optimization may work better.

In addition, traditional CBR mainly tends to provide the user the results by similarity measurement (Kang et al., 2014). It falls into troubles in situations where there are many uncertainties in retrieval requirements, such as customer driven product design. It has been gained widely recognized the importance to

* Corresponding author. Tel.: +86 21 34206552.
E-mail address: hujie@sjtu.edu.cn (J. Hu).

understand customer voice to develop a successful design. However, it is a difficult task especially at the early stages of the product design due to the deficiency of relative knowledge. Though there are a large number of past cases and experiences, it is incredible for customers to know the detailed properties of the product. They only concern the main features of the product to fulfill their tasks. Brief description with principal features after appropriate feature selection should be provided to help customers to know the key characteristics of the product. On the other hand, there are many uncertainties in customers' requirements. It is not always possible for customers to express their requirements accurately. Sometimes, they even do not know how to express their opinions. In such circumstance, a CBR with visualization of the hierarchical structure of the case library to reveal the inherent relations of the past cases can help to capture customers' true perceptions. Therefore, to develop a CBR with appropriate feature selection and case organization as well as case base visualization is an urgent need to assist product design.

The main purpose of this paper is to further enhance the performance of CBR. This paper presents an integrated reduction technique and cluster analysis approach to manipulate the problems of feature selection and case organization in large CBR system. Neighborhood rough set method is introduced as a reduction tool for feature weighting and selection in case representation. Then the growing hierarchical self-organizing map (GHSOM) network is conducted in the initial case clustering and organization. Although feature selection and cluster analysis are widely applied in other data processing and classification areas, past studies only focus on partially one of them, the integration of both techniques are seldom used in CBR.

The rest of this paper is organized as follows. Section 2 gives a brief review about the related work. Section 3 introduces the research approach. Section 4 conducts experiments on UCI datasets and a practical case about electromotor product design to evaluate the performance of the proposed method. Section 5 presents the conclusion and makes suggestion about the future work.

2. Related work

Case retrieval efficiency and accuracy are attractive topics among the development of CBR system. Many researchers conducted a lot of work to improve the retrieval performance by enhancing the process of similarity measurement (Li and Ho, 2009; Kar et al., 2012; Qi et al., 2009, 2011). However, traditional CBR relied heavily on experts' experiences, the fundamental roles of feature selection and case organization are ignored.

Feature selection and weighting generally determine the representative features and their weights, remove redundant ones while case selection deals with the selection of typical cases to construct case memory. Reduction techniques are crucial in this phase by eliminating redundant features and erroneous cases to strengthen the retrieval efficiency and accuracy. Among various feature selection techniques, rough set method is widely used in feature selection and case base reduction in CBR. Salamó and Golobardes (2001), Salamó and López-Sánchez (2011) introduced two reduction methods to manipulate case memory reduction on the basis of rough set algorithm. They conducted case memory reduction through studying the representativeness of the whole cases and employing a different strategy to select the best cases, and further performed feature selection for dimensionality reduction. Jiang et al. (2006) presented a novel model to calculate feature weights and conduct feature reduction based on fuzzy similarity and rough set algorithm. Li et al. (2006) performed feature reduction by rough set and combined it with case selection to handle large documents. Fernandez-Riverola et al. (2007) adopted a rough set based reduction approach to minimize the case base. The method calculated the relevance of each attribute and was used to reduce the memory size of a fuzzy CBR system according to the contribution of each case feature.

However, classical rough set algorithm can only work with categorical features. It cannot deal with numerical variables that are continuous valued. Genetic algorithm (GA) and many other methods were proposed to enhance the processes of feature selection and case reduction. Liu et al. (2008) introduced association rules mining in case base reduction. They attempted to reduce the case base and discover a compact collection of association rules to substitute the homogeneous cases in the initial case memory to improve retrieval efficiency. Ahn and Kim (2009a,b) tried to promote the prediction accuracy of CBR by using GA to optimize the processes of feature weighting, case selection as well as the size of neighbors simultaneous. Li et al. (2009) developed a novel feature selection model based on mutual information in software cost estimation. Xiong and Funk (2006) proposed an approach to assess the quality of feature selection based on the performance of CBR. They further suggested a hierarchical memetic algorithm to conduct feature selection and similarity modeling simultaneously (Xiong and Funk, 2010). In addition, similarity clustering and artificial immune system algorithm were also reported in the feature selection in CBR (Massie et al., 2007; Lin and Chen, 2011).

Although many methods are proposed, most of them are too dependent on the parameters that are selected. They differ from different values due to poor stability. Thus neighborhood rough set method is proposed as a reduction technique to address this issue. It can manipulate both numerical and categorical attributes without discretization and it is powerful in identification of effective subsets of features.

While feature selection deals with preprocessing of the CBR system, case organization, management and maintenance are the procedures of refining the case base. They focus on ways for revising the structure or contents of the case memory to facilitate further retrieving and reasoning. Many clustering algorithms and data mining techniques were introduced in case organization and maintenance. Cao et al. (2001, 2003) proposed a fuzzy-rough method for case library maintenance and adaptation knowledge mining. They used fuzzy-rough method to learn feature weights, divide the case base into some clusters, mine adaptation rules and select representative cases. Yang and Wu (2001) developed an interactive CBR which used clustering algorithm to merge similar cases together and create decision forests to help divide a large case library into several small ones. In order to resolve the problems in case adaptation, Jung et al. (2009) built a RBF network on the basis of typical cases generated by k-means clustering, and obtained the most similar typical one through the network after appropriate adjustment. Zhuang et al. (2009) used self-organizing maps (SOM) as a data mining approach to partition enormous pathology ordering data into several homogenous clusters and extract useful knowledge from the clusters, and further combined with CBR for decision support. Fan et al. (2011) suggested a case-based weighted clustering methodology to divide the original dataset into several clusters, and combined with a fuzzy decision tree and GA to develop a decision-making system for medical diagnosis. Kim and Han (2001) introduced learning vector quantization and SOM to calculate the centroid values of clusters and presented a cluster-indexing method of CBR for bond rating prediction. Freyne and Smyth (2010) exploited an online case representation scheme in CBR to offer users the ability to visualize complex datasets. Chakraborti et al. (2008) studied the problems of visualizing and evaluating complexity in case base maintenance. They suggested a method to visualize case library by features and cases clustering and presented a complexity measurement algorithm.

Among those data mining methods in CBR, cluster analysis is a popular technique which was paid more attention (Tseng et al., 2005). By clustering, the case base is transformed into several small ones in which the cases are more similar in each cluster. The most similar cluster would be retrieved and compared with the new case to find the suitable one. Although clustering is a promising technique in case organization, most of traditional clustering algorithms suffer from many serious problems. First of all, most of the clustering algorithms need to define auxiliary information in advance (Bajo et al., 2010). However, it is not always possible to predefine the proper clustering strategy, the topology of the clustering and the number of clusters. Second, traditional clustering algorithms can only show the similarities by distance measurement, not the inherent hierarchical structure of the dataset. In other words, they lack ability for visualizing the inherent relationships of the clustering results. To address these issues, this paper introduces GHSOM as a clustering algorithm in case organization. It adopts a flexible architecture which only depends on the training process. Most importantly, it can offer the user the proper clusters and the visualization of the inherent hierarchical relations of the case library by stacking similar cases close to each other. The visualization is extremely useful in situations where there are many uncertainties in retrieval requirements such as customer driven product design to help to identify the user's true perception.

While neighborhood rough set based reduction approach plays an important role in the phase of feature selection, GHSOM concerns with how to effectively organize the case memory. Thus these two methods can be combined together to optimize the processes of feature selection and case organization as well as provide the visualization of the case library to further promote the performance of CBR.

3. Proposed method

3.1. Framework of the proposed method

For the purpose of further improving the retrieval performance of CBR system, this paper proposes an integrated method by introducing neighborhood rough set method as a reduction technique in feature selection while conducting cluster analysis in case organization using GHSOM. With the help of feature selection, redundant features are removed. By using GHSOM, the most similar cases and the corresponding cluster can be found for further retrieving. The framework of the research model is shown in Fig. 1.

3.2. Feature selection

Generally, previous knowledge and experiences are stored as cases in the case library while suitable features are selected to represent them. Each case in the case memory contains three parts: one is the case number, which is allocated by the system in order to provide a unique identification for each case; the other is problem description, which is used to describe the situation when the case happened; another is solution description, which is used to represent the results of the corresponding cases. The model is depicted as

{*case number*, *problem description*, *solution description*}

However, each case may have too many features. It is incredible to use all of them to represent a case for retrieving due to the reason that many redundant features may cost a large number of computational times and deteriorate system performance. Usually, the case is represented by a few principal features and redundant features are removed. Thus, more attention should be paid to the issue of how to select principal features from a large amount of initial ones, namely feature selection. This can be done with the help of the concept of usefulness or significance of the feature which is decided by its redundancy and relevancy. A feature is redundant or not lies on whether it is correlated with other features, if it does then it is redundant, otherwise it is not redundant. Accordingly, a feature is relevant or not depends on whether it can affect the prediction of the solution, if it does then it is relevant, otherwise it is irrelevant. Hence the final feature subset should be one that all the problem features are mutually irrelevant, but closely related to the solution ones. When the features are determined, it will be used to describe the cases, and used for further retrieving in CBR.

This paper introduces neighborhood rough set algorithm as a feature selection approach. Generally, rough set method is a popular approach widely used in attribute reduction. Neighborhood rough set algorithm is a variant of classical rough set which introduces the concept of neighborhood relations (Hu et al., 2008). Formally, an information system is represented by $S = \langle U, A \rangle$, where U denotes the universe, A is a set of attributes, $A = \{a_1, a_2, \ldots, a_n\}$ and $A = C \cup D$, C is a set of condition attributes and D is the decision one (Walczak and Massart, 1999).

Given $B \subseteq C$, $\forall x_i \in U$, the neighborhood $\delta_B(x_i)$ of x_i in B is defined as

$$\delta_B(x_i) = \{x_j | x_j \in U, \quad \Delta_B(x_i, x_j) \leq \delta\} \tag{1}$$

where Δ is a distance function.

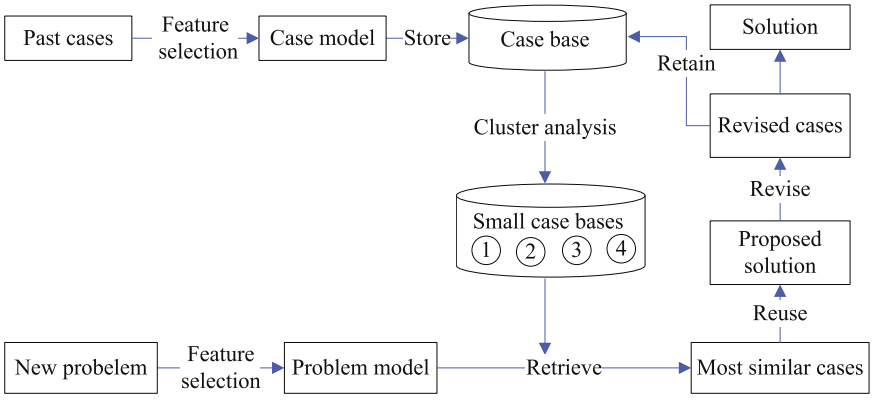

Fig. 1. Framework of the proposed method.

$\forall x_1, x_2 \in U$, their distance is defined as

$$\Delta_P(x_1, x_2) = \left(\sum_{i=1}^{n} |f(x_1, a_i) - f(x_2, a_i)|^P \right)^{1/P} \tag{2}$$

where $f(x, a_i)$ is the value of x in a_i.

Thus the set of neighborhood $\{\delta(x_i) | x_i \in U\}$ composes a granule space that covers U rather than partitions it. The neighborhood relation N on U can be denoted by a relation matrix $R(N) = (r_{ij})_{n \times n}$, where

$$r_{ij} = \begin{cases} 1, & \Delta(x_i, x_j) \le \delta; \\ 0 & \text{otherwise.} \end{cases}$$

Then the lower approximation ($\underline{NX}$), upper approximation ($\overline{NX}$) and boundary (BNX) of X in $\langle U, N \rangle$ are defined as

$$\underline{NX} = \{x_i | \delta(x_i) \subseteq X, x_i \in U\} \tag{3}$$

$$\overline{NX} = \{x_i | \delta(x_i) \bigcap X \neq \varnothing, x_i \in U\} \tag{4}$$

$$BNX = \overline{NX} - \underline{NX} \tag{5}$$

Therefore a neighborhood decision system is represented as $NDS = \langle U, A, N \rangle$, the lower approximation ($\underline{N_B D}$), upper approximation ($\overline{N_B D}$) and boundary ($BN(D)$) of D on B are defined as

$$\underline{N_B} D = \bigcup_{i=1}^{n} \underline{N_B} X_i \tag{6}$$

$$\overline{N_B} D = \bigcup_{i=1}^{n} \overline{N_B} X_i \tag{7}$$

$$BN(D) = \overline{N_B} D - \underline{N_B} D \tag{8}$$

where $X_1, X_2, \ldots, X_n$ are the object subsets corresponding with decision 1 to n, $\underline{N_B} X = \{x_i | \delta_B(x_i) \subseteq X, x_i \in U\}$, $\overline{N_B} X = \{x_i | \delta_B(x_i) \bigcap X \neq \varnothing, x_i \in U\}$. The lower approximation ($\underline{N_B D}$) is also known as the positive region of decision D on condition B, denoted by $POS_B(D)$.

The dependency of D on B is defined as

$$\gamma_B(D) = \frac{|POS_B(D)|}{|U|} \tag{9}$$

For a given $NDS = \langle U, A, N \rangle$, $B \subseteq C$, B is a relative *reduct* if: $\gamma_B(D) = \gamma_A(D)$; $\forall a \in B, \gamma_B(D) > \gamma_{B-a}(D)$.

The significance of attribute a is defined as

$$sig(a, B, D) = \begin{cases} \gamma_B(D) - \gamma_{B-a}(D), & \text{if } a \in B \\ \gamma_{B \cup a}(D) - \gamma_B(D), & \text{if } a \notin B \end{cases} \tag{10}$$

Hence attribute a is redundant if $sig(a, B, D) = 0$, otherwise it is indispensable. Therefore the feature selection algorithm according to neighborhood rough set is conducted as follows:

Step 1: input: information system $\langle U, A \rangle$, size of the neighborhood δ, convergence criterion ε; output: *reduct*;

Step 2: $\varnothing \rightarrow reduct$;

Step 3: for each attribute $a_i \in C - reduct$, calculate $\gamma_{reduct \cup a_i}(D)$ and $sig(a_i, reduct, D)$;

Step 4: find a_l: $sig(a_l, reduct, D) = \max_i(sig(a_i, reduct, D))$;

Step 5: if $sig(a_l, reduct, D) > \varepsilon$, then $reduct \cup a_l \rightarrow reduct$, go to Step 2;

else

return *reduct*;

end

By using neighborhood rough set algorithm, feature weights are determined and redundant features are removed according to the weights of each feature.

3.3. Case organization

When the feature selection is accomplished, emphasis should be placed on how to organize those cases. In this paper, GHSOM is introduced to perform cluster analysis to the case library so the large case base can be partitioned into small groups to reduce the size of stack memory and optimize the retrieval performance.

Due to its outstanding ability in the visualization of topological relationship for high-dimensional dataset, SOM is widely employed as a powerful tool for exploratory data analysis in the identification and visualization of clustering in the dataset (Kohonen, 1982). However, traditional SOM has to predefine the topology of the network and it cannot provide the visualization of the clustering results with a hierarchical structure. To overcome these deficiencies of traditional SOM, this paper introduces GHSOM for case organization and clustering in case base construction. The GHSOM adopts a flexible and hierarchical architecture with multiple layers while each of them contains many independent SOMs where the numbers of layers, maps as well as neurons are all determined by training. It can offer the user the proper clusters as well as the inherent hierarchical relations of the case memory (Rauber et al., 2002). Initially, only one SOM is utilized in layer 1. For each node on the structure, a new SOM might be increased to the subsequent layer if the deviation reaches the given criterion. The mechanism is repeated until it reaches the stopping criterion. The architecture of the GHSOM is shown in Fig. 2.

The construction of the GHSOM is described as follows:

Step 1: Parameter initialization: such as the initial map size, learning rate, neighborhood range, growing-stopping criterion, hierarchical stopping criterion, label threshold and the number of labels.

Step 2: Set up a virtual layer 0 which contains only one node (unit). Its weight is assigned as the mean of the input data. The mean quantization error (mqe_0) is obtained by measuring the Euclidean distance between the input data and the weight.

Step 3: Determine the map size of layer 1 on the basis of the standard SOM training procedure. As shown in Fig. 2, the map in layer 1 is made up of 3×2 nodes and offers a rough organization to the input data.

Step 4: Grow horizontally of the map. By computing the mqe of all nodes in the current layer, the mapping quality is evaluated and

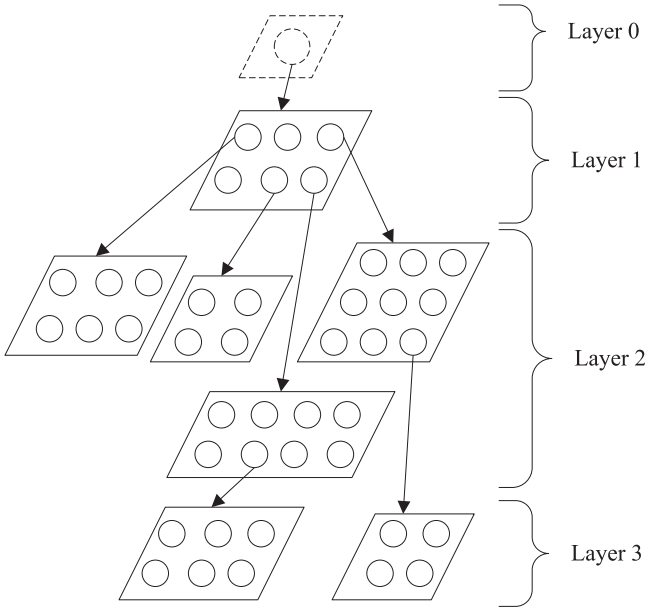

Fig. 2. Architecture of the GHSOM.

the error unit is identified. Then new units are inserted by rows or columns between the error one and its most dissimilar neighbor, and their weights are assigned as the average of their neighbors. Finally, compute *MQE* of the present map which is equivalent to the average *mqe* of all the nodes. The map growth continues until the *MQE* reaches a certain fraction τ_1 of the mqe_u of the corresponding node in the former layer:

$$MQE_m < \tau_1 \cdot mqe_u \tag{11}$$

Step 5: Expand the hierarchical structure. The hierarchy grows until all the mqe_i in the present layer reaches the given fraction τ_2 of the mqe_0 of layer 0.

$$mqe_i < \tau_2 \cdot mqe_0 \tag{12}$$

Step 6: Set up the further layers by repeating the procedure of Step 4 and 5. The down layer is the subset of the corresponding node in the upper layer which has been mapped onto. Then the maps in the bottom layer provide a more detailed data organization.

Once the GHSOM is created, all previous cases of the CBR system are divided into several small groups. By changing the values of τ_1 and τ_2, the size of each group can be controlled.

When a new problem occurs, it is first represented by the problem feature vectors. Then it will be put together with the past cases to conduct clustering. Finally, it is guided into corresponding subset by GHSOM, in which the new problem is mapped with the previous cases and the most similar one can be obtained by case retrieval using similarity measurement such as cosine similarity or fuzzy similarity.

4. Case study

4.1. Evaluation on benchmark datasets

Before used in CBR, experiments are conducted to evaluate the performance of the neighborhood rough set method, GHSOM algorithm and their integrated model. The experiments are implemented on a PC platform with Intel Core i5-3550 3.30 GHz CPU, 8GB RAM, Windows 7 and Matlab development environment. Ten datasets in UCI machine learning repository[1] are selected as test benchmarks. The properties of the datasets are outlined in Table 1.

In evaluation of the feature selection method, two other popular methods including ReliefF (Robnik-Šikonja and Kononenko, 2003) and PCA (Jolliffe, 2005) together with the original one without reduction are used to measure the performance of the proposed neighborhood rough set method (NRSM). Correspondingly, in the clustering algorithm evaluation, four other prevalent techniques including Kmeans, FCM, KNN and SOM, are utilized to compare with the GHSOM algorithm. Then their integrated method (NRSM–GHSOM) is developed and compared with the Kmeans and GHSOM clustering. Classification accuracy is used as an index of performance evaluation in various experiments. The comparison of the feature selection approach, clustering algorithm and their integrated method are shown in Tables 2, 3 and 4, respectively.

In Table 2, the classification accuracy in feature selection evaluation is calculated by Kmeans. From Table 2, it can be concluded that the neighborhood rough set feature selection method has higher classification accuracy than other methods in general. While in Table 3, the GHSOM algorithm outperforms other clustering methods in general. In addition, the integrated NRSM–GHSOM can further improve the classification accuracy, which are shown in Table 4. Besides, most importantly, the

[1] http://archive.ics.uci.edu/ml/

Table 1
Property description of the datasets.

Datasets	Instances	Features	Classes
Wine	178	13	3
WDBC	569	32	2
Seeds	210	7	3
Ionosphere	351	34	2
Credit	690	15	2
Sonar	208	60	2
Heart	270	13	2
German	1000	24	2
Australian	690	14	2
Spect	267	44	2

Table 2
Comparison of the feature selection method.

Datasets	Classification accuracy (%)			
	Original	ReliefF	PCA	NRSM
Wine	94.94	95.51	94.94	97.75
WDBC	92.79	93.50	93.15	94.02
Seeds	89.05	89.05	89.05	90.48
Ionosphere	71.23	76.64	75.50	77.78
Credit	79.02	86.37	86.37	86.37
Sonar	53.37	62.5	61.54	66.83
Heart	79.63	81.11	79.63	79.63
German	55.8	63.7	65.6	67.8
Australian	82.32	85.51	85.51	85.51
Spect	62.55	71.54	68.91	65.54
Average	76.07	80.54	80.02	81.17

Table 3
Comparison of the clustering algorithm.

Datasets	Classification accuracy (%)				
	Kmeans	FCM	KNN	SOM	GHSOM
Wine	94.94	94.94	95.51	95.51	97.19
WDBC	92.79	92.79	96.66	92.79	95.61
Seeds	89.05	90	92.86	89.05	94.76
Ionosphere	71.23	70.94	84.33	70.94	91.17
Credit	79.02	81.16	86.22	77.95	86.68
Sonar	53.37	55.29	82.69	55.77	86.06
Heart	79.63	79.26	80.74	79.63	86.3
German	55.8	57.1	70.7	55.8	74.9
Australian	82.32	81.88	84.93	85.51	86.96
Spect	62.55	51.31	73.41	64.79	80.52
Average	76.07	75.47	84.81	76.77	88.02

Table 4
Comparison of the integrated method.

Datasets	Classification accuracy (%)			
	Kmeans	NRSM	GHSOM	NRSM–GHSOM
Wine	94.94	97.75	97.19	100
WDBC	92.79	94.02	95.61	97.54
Seeds	89.05	90.48	94.76	94.76
Ionosphere	71.23	77.78	91.17	92.88
Credit	79.02	86.37	86.68	87.29
Sonar	53.37	66.83	86.06	87.02
Heart	79.63	79.63	86.30	87.04
German	55.8	67.8	74.9	75.8
Australian	85.51	82.32	86.96	87.97
Spect	62.55	65.54	80.52	83.90
Average	76.07	81.17	88.02	89.42

GHSOM can provide visualization of the datasets. From the clustering of GHSOM, people can get to know the inherent relations of the dataset.

4.2. Experiment on a practical case

After evaluation of the proposed method on benchmark datasets, a new CBR model is developed and an empirical case study of electromotor product design is conducted to show the effectiveness of the proposed CBR model. Generally, the electromotor is an important component in industrial department and widely used in various areas. Among the development of the electromotor industry, a large number of experiences and knowledge are collected. In a customer driven electromotor design, a CBR with appropriate feature selection and case organization is put forward to assist the product development.

Table 5
List of attributes.

Attributes	Description	Type	Example
A_1	Rated output power	Numerical	30
A_2	Rated voltage	Numerical	380
A_3	Rated speed	Numerical	2800
A_4	Rated frequency	Numerical	50
A_5	Flameproof performance	Symbolic	Y
A_6	Poles	Numerical	4
A_7	Noise	Numerical	80
A_8	Pullout torque/Rated torque	Numerical	2
A_9	Locked torque/Rated torque	Numerical	2
A_{10}	Locked current/Rated current	Numerical	7
A_{11}	Duty/Rating	Symbolic	S1
A_{12}	Motor identification symbol	Symbolic	YB2
A_{13}	Connection	Symbolic	Y
A_{14}	Power factor	Numerical	0.88
A_{15}	Manufacturer	Symbolic	N
A_{16}	Cooling method	Symbolic	IC411
A_{17}	Vibration class	Numerical	2.8
A_{18}	Protection class	Symbolic	IP55
A_{19}	Temperature rise class	Symbolic	B
A_{20}	Insulation class	Symbolic	F
A_{21}	Altitude above sea level	Numerical	1000

In this study, 1103 cases were collected from its previous product series to build a case base. Among the cases, 21 attributes are collected from its product specification, which are shown in Table 5. There are 13 numerical attributes and 8 symbolic attributes in the attributes list. Based on the proposed method, a computer-aided platform is exploited to assist product design which is illustrated in Fig. 3. Then the proposed model is described as follows:

Step 1: Data preprocessing. For the symbolic data in Table 5, it should be transformed to numerical one.

$$F_i = \frac{i-1}{n-1} \tag{13}$$

which is established on the assumption that the symbolic data has n symbolic values and each of them is denoted by an order value as its corresponding numerical one. Hence 1 is the minimum numerical value of the corresponding symbolic data while n is the maximum and i denotes the ordinal value of symbolic data of case i.

Step 2: Feature selection. After the preprocessing of the initial dataset, neighborhood based rough set algorithm stated in Section 3.2 is introduced to select principal features from the initial attributes in Table 5. Finally, four principle features are selected from the initial 21 attributes, including A_3 (Rated speed), A_1 (Rated output power), A_{12} (Motor identification symbol) and A_2 (Rated voltage), which are shown in Table 6. The others are eliminated through the reduction algorithm according to their weights. It is more convenient to use such small dimensional data to represent the cases for retrieval and management than the initial large one.

Step 3: Data normalization. After feature selection, the data to be clustered should be normalized in order to change the data between [0, 1].

$$F_i = \frac{x_i - \min(x_1, x_2, \ldots, x_n)}{\max(x_1, x_2, \ldots, x_n) - \min(x_1, x_2, \ldots, x_n)} \tag{14}$$

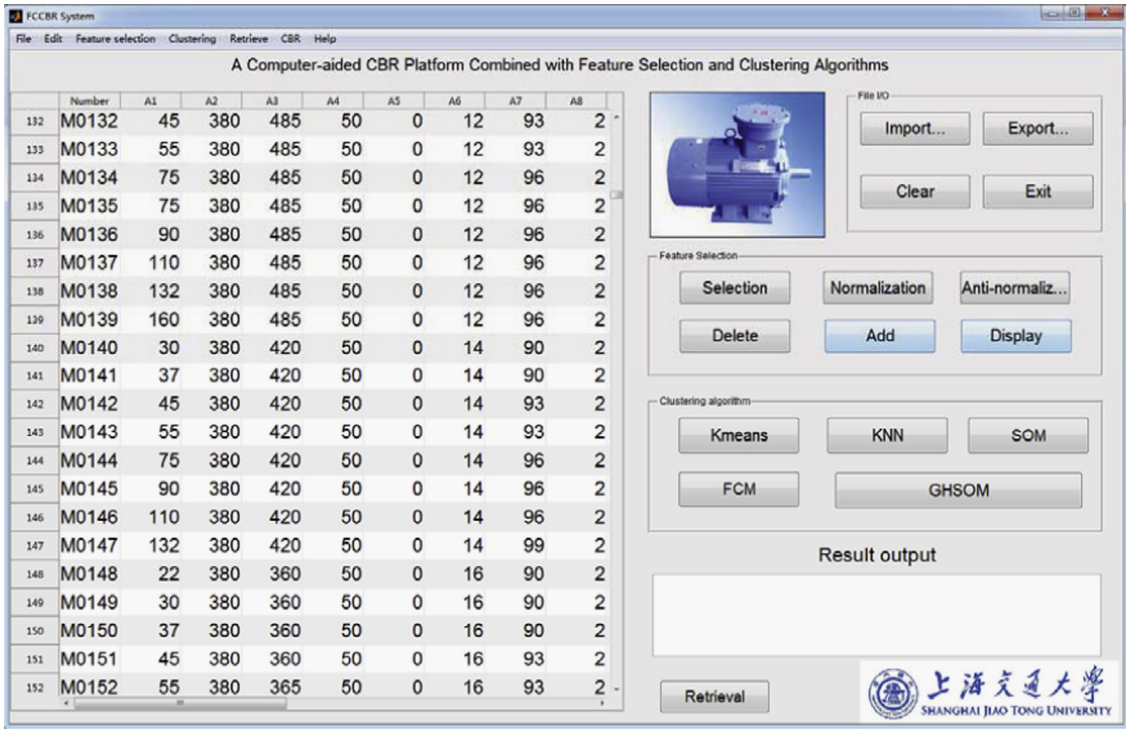

Fig. 3. The computer-aided CBR system.

where F_i is the normalized value of each variable, n denotes the sum of cases and x_i is the value of X of case i.

Table 6
List of selected features.

Attributes	Description	Weights
A_3	Rated speed	0.0037
A_1	Rated output power	0.2653
A_{12}	Motor identification symbol	0.3501
A_2	Rated voltage	0.3809
Total		1.0000

Table 7
Input dataset after feature selection.

Number	A_1	A_2	A_3	A_{12}
M0001	2.7E−5	0	0.9285	0
M0002	5.8E−5	0	0.9285	0
M0003	1.1E−4	0	0.9285	0
M0004	1.9E−4	0	0.9285	0
M0005	2.8E−4	0	0.9380	0
M0006	4.4E−4	0	0.9380	0
M0007	6.2E−4	0	0.9437	0
M0008	9.3E−4	0	0.9437	0
M0009	0.0013	0	0.9589	0
M0010	0.0017	0	0.9627	0
M0011	0.0024	0	0.9665	0
M0012	0.0033	0	0.9665	0
M0013	0.0049	0	0.9779	0
M0014	0.0066	0	0.9779	0
M0015	0.0082	0	0.9779	0
M0016	0.0098	0	0.9817	0
M0017	0.0133	0	0.9855	0
M0018	0.0165	0	0.9855	0
M0019	0.0200	0	0.9932	0
M0020	0.0245	0	0.9932	0
M0021	0.0334	0	0.9932	0
…	…	…	…	…

Step 4: Construction of the database. After the preprocessing of the original data and selection of the principal features, the initial database can be built by these data which is shown in Table 7.

Step 5: Case base organization. Once the database is created, GHSOM is adopted as a clustering tool to partition the whole case base into hierarchical small ones, which composite the structure of the case memory. Fig. 4 shows the clustering result of the cases. Because there are too many cases, it is not possible to show all the cases in one figure, so Fig. 4 only shows the sum of cases in each cluster. According to Fig. 4, there are only 26 cases in the biggest cluster, not the whole 1103 cases; it is very convenient to retrieve similar cases in such small clusters. Besides, it is very simple to adjust the amount of cases in the cluster.

Step 6: Case retrieval. When a new problem happens, it will be represented by the problem feature vectors and then put together with the past cases to conduct clustering. With the help of GHSOM, the retrieval data is led into corresponding small case library and the most similar case can be obtained within the small case base by similarity measurement. Fig. 5 shows the retrieval result. It should be noted that the most similar cluster is shown with cases, and the others still only show the sum of cases. It is very convenient to know the whole cases in the small group and which is the most similar one to the retrieved case. From Fig. 5 it is obviously that there are 3 cases which are most similar with the retrieved case. Then the similarity within the cluster is measured by cosine similarity, and 'M0168' has the biggest similarity 0.99986 with the retrieved cases 'D0001'.

4.3. Comparison and discussion

The purpose of this study is to enhance CBR performance by introducing feature selection and cluster analysis. To evaluate the proposed hybrid CBR (HCBR), traditional CBR (TCBR) (Kolodner, 1993) and CBR combined with ReliefF (CBR-ReliefF) (Robnik-Šikonja and Kononenko, 2003), are conducted in the experiment for comparison. In this experiment, relative error rate (*RER*) and computational time are employed as the performance

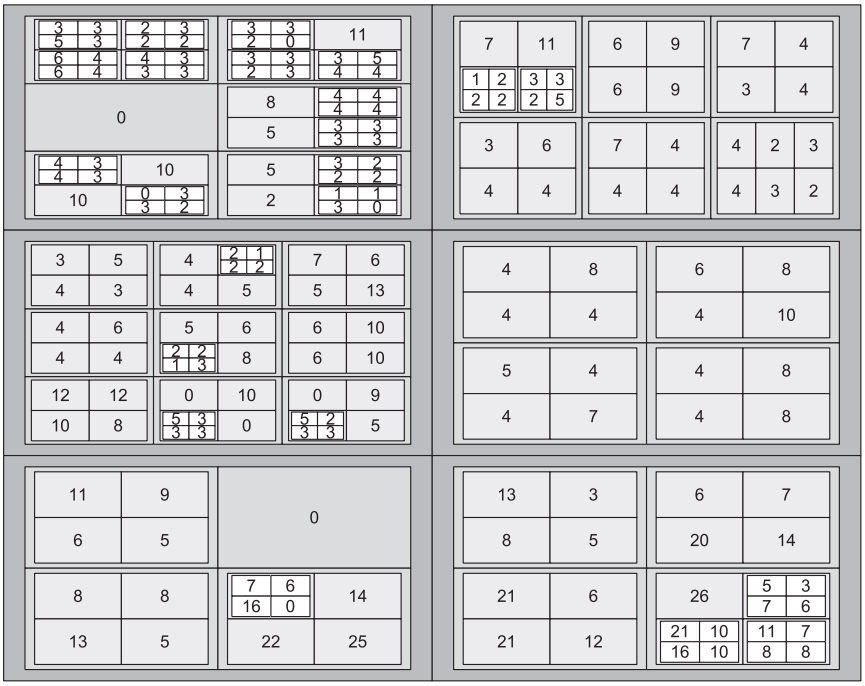

Fig. 4. The result of case clustering.

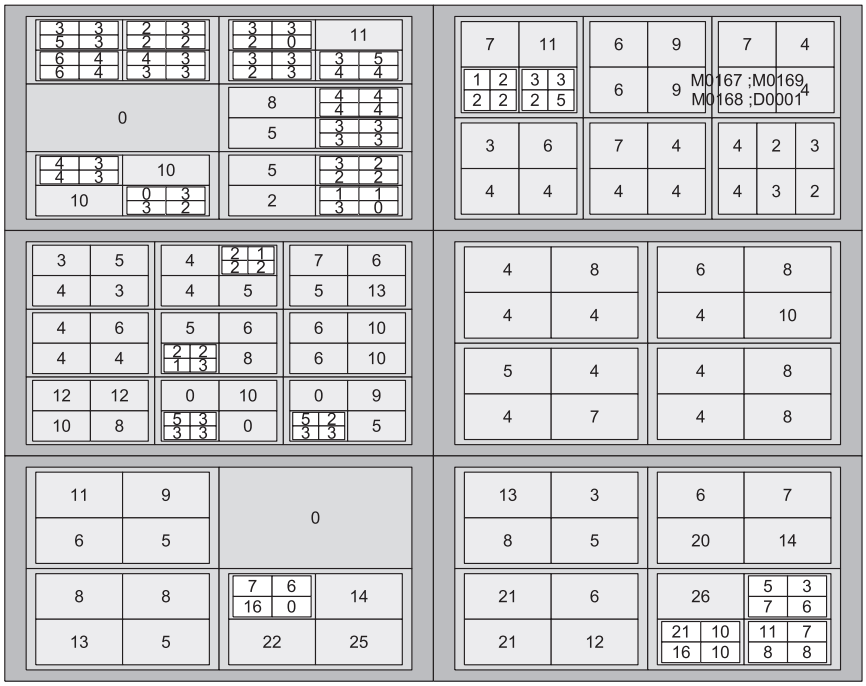

Fig. 5. The result of case retrieval.

Table 8
Comparison of different CBR methods.

Ten-fold	Performance					
	TCBR		CBR-ReliefF		HCBR	
	RER (%)	Time (s)	*RER* (%)	Time (s)	*RER* (%)	Time (s)
1	13.95	6.11	10.82	5.99	7.09	4.93
2	18.43	6.12	10.54	5.91	6.42	4.99
3	15.23	6.05	10.25	5.98	6.26	4.91
4	16.95	6.11	13.45	5.90	6.26	4.91
5	16.15	6.11	9.22	5.91	5.35	4.91
6	17.04	6.09	12.64	5.95	5.52	4.89
7	16.69	6.08	10.63	5.99	6.51	4.98
8	15.15	6.04	10.47	6.04	6.49	4.87
9	16.71	6.04	13.12	6	5.22	4.88
10	15.57	6.08	12.06	6	6.9	4.92
Average	16.19	6.08	11.32	5.97	6.31	4.93

indices of each method, which are expressed as follows:

$$RER = \left| \frac{Actual\ value - Predict\ value}{Actual\ value} \right| \times 100\% \tag{15}$$

where *Actual value* is the actual value of the case and *Predict value* is the retrieved one.

To evaluate the performance of different CBR models, ten-fold cross-validation is conducted based on the practical example in Section 4.2. The results of the experiments are illustrated in Table 8. Among them, the proposed HCBR has the least *RER* and computational time comparing with the other two methods. In *RER*, the proposed model is 6.31% while TCBR is 16.19% and CBR-ReliefF is 11.32%. In computational time, the proposed method employs 4.93 s while the TCBR is 6.08 s and CBR-ReliefF is 5.97 s. Obviously, by integrating with neighborhood rough set and GHSOM, the proposed CBR model can effectively promote the retrieval accuracy and efficiency.

Moreover, comparing with other two CBR models, the proposed CBR can also lower the burden of retrieving by measuring the similarity only between the principal features rather the whole ones with the help of feature selection technique. It is helpful even reduce a small group of attributes especially in the environment of large case base or complex case which is represented by hundreds attributes. Besides, due to the outstanding ability in the visualization of topological relationship among the high-dimensional inputs and the low-dimensional representation of GHSOM approach, the proposed CBR can visually display the inherent similar relations within the case cluster. The GHSOM can automatically determine the number of clusters according to the training process rather than predefined. It adopts a flexible and hierarchical structure to represent the results of case clustering. When a new problem happens, the proposed method can easily and visually show the similar cluster of the retrieved case through GHSOM clustering. The most suitable case is obtained within the similar cluster by similarity measurement as well as the closeness between different cases.

5. Conclusion

Due to the important roles of feature selection and case organization, this paper proposes reduction technique in feature selection and cluster analysis in case organization in CBR system. In feature selection, the paper introduces neighborhood rough set method as a reduction technique to calculate the weights of all the features and irrelevant ones are removed according to their weights. Then a minimal subset of features is selected from the problem domain. Once the feature selection is accomplished, the paper conducts cluster analysis to organize those cases. In this phase, GHSOM is introduced as a clustering tool to divide the initial case memory into several small ones, in which the cases are more similar to each other in the same group than those from different clusters. Thus the large case library is transformed into several hierarchical small ones. New cases are guided into corresponding small clusters and case retrieval is done within the cluster. By assigning a suitable threshold, an appropriate amount of similar cases are obtained for future revision and reuse. Finally, the paper evaluates the proposed algorithms on benchmark

datasets and adopts the new developed CBR model in a practical case study to assist electromotor product design. After comparing with several other methods, the conclusion that the proposed CBR model is applicable for problem solving with higher performance is draw.

Although the proposed method is a progress to some extent, there are still many places need to improve. In the future, efforts should be placed on the simultaneously optimization of feature selection, case selection and case organization under complex environment. Besides, researchers should pay more attention to the combination of CBR and database technology as well as data mining techniques to lower the burden of the case memory and enhance the case retrieval efficiency.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (no. 51475288, 51305260, 51275293), National Key Scientific Instruments and Equipment Development Program of China (no. 2013YQ03065105, 2011YQ030114), "Shu Guang" project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation (12SG14), Shanghai Committee of Science and Technology (no. 11JC1406100, 13111102800), Natural Science Foundation of Shanghai (no. 13ZR1421400).

References

Ahmed, M.U., Begum, S., Funk, P., Xiong, N., von Scheele, B., 2011. A multi-module case-based biofeedback system for stress treatment. Artif. Intell. Med. 51 (2), 107–115.

Ahn, H., Kim, K.-j., 2009a. Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. Appl. Soft Comput. 9 (2), 599–607.

Ahn, H., Kim, K.-j., 2009b. Global optimization of case-based reasoning for breast cytology diagnosis. Expert Syst. Appl. 36 (1), 724–734.

Bajo, J., De Paz, J.F., Rodríguez, S., González, A., 2010. A new clustering algorithm applying a hierarchical method neural network. Logic J. IGPL 19 (2), 304–314.

Cao, G., Shiu, S., Wang, X., 2001. A fuzzy-rough approach for case base maintenance. In: Case-Based Reasoning Research and Development. Springer, pp. 118–130.

Cao, G., Shiu, S.C., Wang, X., 2003. A fuzzy-rough approach for the maintenance of distributed case-based reasoning systems. Soft Comput. 7 (8), 491–499.

Chakraborti, S., Beresi, U.C., Wiratunga, N., Massie, S., Lothian, R., Khemani, D., 2008. Visualizing and evaluating complexity of textual case bases. In: Advances in Case-Based Reasoning. Springer, pp. 104–119.

Chuang, C.-L., 2013. Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction. Inf. Sci. 236, 174–185.

Fan, C.-Y., Chang, P.-C., Lin, J.-J., Hsieh, J., 2011. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl. Soft Comput. 11 (1), 632–644.

Fernandez-Riverola, F., Diaz, F., Corchado, J.M., 2007. Reducing the memory size of a fuzzy case-based reasoning system applying rough set techniques. IEEE Trans. Syst. Man Cybern. C Appl. Rev. 37 (1), 138–146.

Freyne, J., Smyth, B., 2010. Visualization for the masses: learning from the experts. In: Case-Based Reasoning Research and Development. Springer, pp. 111–125.

Guo, Y., Hu, J., Peng, Y., 2011. Research on CBR system based on data mining. Appl. Soft Comput. 11 (8), 5006–5014.

Guo, Y., Hu, J., Peng, Y., 2012. A CBR system for injection mould design based on ontology: a case study. Comput.-Aided Des. 44 (6), 496–508.

Hu, Q., Yu, D., Liu, J., Wu, C., 2008. Neighborhood rough set based heterogeneous feature subset selection. Inf. Sci. 178 (18), 3577–3594.

Jiang, Y.-j., Chen, J., Ruan, X.-y., 2006. Fuzzy similarity-based rough set method for case-based reasoning and its application in tool selection. Int. J. Mach. Tools Manuf. 46 (2), 107–113.

Jolliffe, I., 2005. Principal Component Analysis. Wiley Online Library.

Jung, S., Lim, T., Kim, D., 2009. Integrating radial basis function networks with case-based reasoning for product design. Expert Syst. Appl. 36 (3), 5695–5701.

Kang, Y.-B., Krishnaswamy, S., Zaslavsky, A., 2014. A retrieval strategy for case-based reasoning using similarity and association knowledge. IEEE Trans. Cybern. 44 (4), 473–487.

Kar, D., Chakraborti, S., Ravindran, B., 2012. Feature weighting and confidence based prediction for case based reasoning systems. In: Case-Based Reasoning Research and Development. Springer, pp. 211–225.

Kim, K.-S., Han, I., 2001. The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases. Expert Syst. Appl. 21 (3), 147–156.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43 (1), 59–69.

Kolodner, J., 1993. Case-based Reasoning. Morgan Kaufmann Publishers, San Mateo, CA.

Li, S.-T., Ho, H.-F., 2009. Predicting financial activity with evolutionary fuzzy case-based reasoning. Expert Syst. Appl. 36 (1), 411–422.

Li, Y., Shiu, S.C.-K., Pal, S.K., Liu, J.N.-K., 2006. A rough set-based case-based reasoner for text categorization. Int. J. Approx. Reason. 41 (2), 229–255.

Li, Y., Xie, M., Goh, T., 2009. A study of mutual information based feature selection for case based reasoning in software cost estimation. Expert Syst. Appl. 36 (3), 5921–5931.

Lin, S.-W., Chen, S.-C., 2011. Parameter tuning, feature selection and weight assignment of features for case-based reasoning by artificial immune system. Appl. Soft Comput. 11 (8), 5042–5052.

Liu, C.-H., Chen, L.-S., Hsu, C.-C., 2008. An association-based case reduction technique for case-based reasoning. Inf. Sci. 178 (17), 3347–3355.

Massie, S., Wiratunga, N., Craw, S., Donati, A., Vicari, E., 2007. From anomaly reports to cases. In: Case-Based Reasoning Research and Development. Springer, pp. 359–373.

Qi, J., Hu, J., Peng, Y., Wang, W., Zhan, Z., 2011. AGFSM: an new FSM based on adapted Gaussian membership in case retrieval model for customer-driven design. Expert Syst. Appl. 38 (1), 894–905.

Qi, J., Hu, J., Peng, Y.-H., Wang, W., Zhang, Z., 2009. A case retrieval method combined with similarity measurement and multi-criteria decision making for concurrent design. Expert Syst. Appl. 36 (7), 10357–10366.

Rauber, A., Merkl, D., Dittenbach, M., 2002. The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. IEEE Trans. Neural Netw. 13 (6), 1331–1341.

Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. 53 (1–2), 23–69.

Salamó, M., Golobardes, E., 2001. Rough sets reduction techniques for case-based reasoning. In: Case-Based Reasoning Research and Development. Springer, pp. 467–482.

Salamó, M., López-Sánchez, M., 2011. Rough set based approaches to feature selection for case-based reasoning classifiers. Pattern Recognit. Lett. 32 (2), 280–292.

Tseng, H.-E., Chang, C.-C., Chang, S.-H., 2005. Applying case-based reasoning for product configuration in mass customization environments. Expert Syst. Appl. 29 (4), 913–925.

Walczak, B., Massart, D., 1999. Rough sets theory. Chemom. Intell. Lab. Syst. 47 (1), 1–16.

Xie, X., Lin, L., Zhong, S., 2013. Handling missing values and unmatched features in a CBR system for hydro-generator design. Comput.-Aided Des. 45 (6), 963–976.

Xiong, N., Funk, P., 2006. Construction of fuzzy knowledge bases incorporating feature selection. Soft Comput. 10 (9), 796–804.

Xiong, N., Funk, P., 2010. Combined feature selection and similarity modelling in case-based reasoning using hierarchical memetic algorithm. In: IEEE Congress on Evolutionary Computation (CEC), 2010, pp. 1–6.

Yang, Q., Wu, J., 2001. Enhancing the effectiveness of interactive case-based reasoning with clustering and decision forests. Appl. Intell. 14 (1), 49–64.

Zhuang, Z.Y., Churilov, L., Burstein, F., Sikaris, K., 2009. Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. Eur. J. Oper. Res. 195 (3), 662–675.