Case Mining from Raw Data for Case Library Construction

Chien-Chang Hsu Ye-Hong Huang

Department of Computer Science and Information Engineering,
Fu-Jen Catholic University
510 Chung Cheng Rd., Hsinchuang, Taipei, TAIWAN 242

Abstract

Case-based reasoning systems usually use prior experiences and examples to solve problems. The successfulness of the systems depends on the completeness of case library. It may generate contradictory solutions or increase adaptation cost if the case library contains irrelevant and disordered cases. This work proposes a case mining system to extract representative cases from raw data. The system constructs a case library by feature mining and case mining. Feature mining evaluates relevance between feature and class by fuzzy measurement. The system then uses relevant features to divide raw data into different clusters. Case mining selects cases from each cluster by genetic algorithm. Finally, the system verifies completeness of case library by covering test and utilization statistics. The experimental results show that the system can select representative cases from the data correctly.

Keywords: Case library, Feature extraction, Case mining, Genetic algorithm

1. Introduction

Case-based reasoning (CBR) is a reasoning methodology that uses past experiences to solve problems. CBR systems have been applied into many applications successfully, such as diagnosis, design, planning, and decision support [1, 4, 10, 11, 15, 16]. For example, R. Amen uses CBR to help mechanical designers to find appropriate raw materials and heat treatment of steel [2]. Y. Fu proposes CBR to improve traditional questions-and-answers system [6]. It uses correlation between keywords and cases to rank candidate cases from case library.

The reasoning process of CBR contains four phases, that is, retrieval, reuse, revision, and retain. Case retrieval is first step to select most similar cases for further reasoning. However, the successfulness of CBR depends on completeness of case library in case retrieval. Most CBR systems assume case library is born as the system constructed deservedly. It is possible to generate incorrect solutions or fail to solve problems if the case library contains deficient cases. On the other hand, the system may also spend too much effort on selecting and adapting inappropriate cases from disordered case library. Case coverage and case utilization of CBR systems become important factors. Moreover, it is hard to construct a complete case library from raw data without any assistant of intelligent technology in initial phase of CBR. Some systems use data mining and case elimination approach to construct the case library. For example, MOE4CBR system uses data mining for case-based reasoning in biological domains. The system uses spectral clustering to cluster the data set into k groups and logistic regression model to select descriptive features in each cluster. Each cluster is considered as a case-based for the k CBR experts [3]. Y. Li proposes a CBR classifier by combining feature reduction and case selection. Feature reduction uses a fast rough-set approach for computing the approximate reduction and feature importance. Case selection uses feature similarly measurement to select the most similar cases [12]. Z. W. Ni uses outlier data mining to delete deviation cases in the case library. The system extracts outliers from existing case base, eliminates erroneous outlier cases, and sieving cases from non-outliers [14].

This work proposes a case mining system for case library construction. The system contains two modules, namely, feature extractor and case miner. Feature extractor uses fuzzy measurement to find dominant features by evaluating feature correlation, data appearance, and gain ration of features. Moreover, case miner uses k-means algorithm and genetic algorithm to select representative cases from each cluster. Finally, the system is verified by covering test and utilization statistics.

The rest of this paper is organized as follows. Section 2 introduces the architecture of a case mining system for case library construction. Section 3 and 4 explore the feature extractor and case miner. Section 5 demonstrates the application of the case mining

system in medical database. Finally, section 6 concludes the work.

2. System Architecture

Figure 1 shows the architecture of a case mining system for case library construction. The system contains two main modules, that is, feature extractor and data miner. Basically, feature extractor uses fuzzy measurement to select dominant features from raw data. It uses Pearson product moment correlation, data appearance, and gain ration to evaluate correlation between feature and class. The features with higher fuzzy measurement are selected from each class as case features.

Case miner uses dominant features and k-means algorithm to partition raw data into different clusters. It then uses genetic algorithm to select representative cases from each cluster for case library construction. The chromosome of genetic algorithm uses binary value to represent cases. Genetic algorithm uses case coverage as evaluation factors of fitness function. Finally, the system is evaluated by using covering test and utilization statistics to verify the robustness of case library.

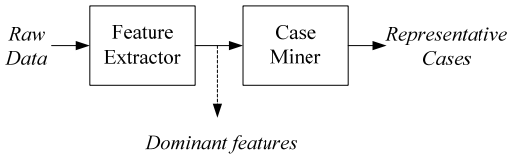

Figure 1 System Architecture

3. Feature Extractor

The main idea of feature extractor is to reduce the dimension of raw data. Specifically, feature extractor uses fuzzy measurement, FM, to evaluate correlation between feature and class.

$$FM = PC * DA * GR \tag{2}$$

where PC, DA, and GR represent feature correlation, data appearance, and gain ration correspondingly. The Pearson product moment correlation, PC, evaluates relationship between feature and class.

$$PC = \left| \frac{\sum_{i=1}^{n}(A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \overline{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \overline{B})^2}} \right| \tag{3}$$

where A, B, $\overline{A}$, and $\overline{B}$ are feature value, class value, average feature value, and average class value.

Data appearance, DA, measures the times of feature region appearance. Data appearance partitions the feature value into different discrete regions and computes the region appearance times in different class.

$$DA = \frac{\sum_{i=1}^{k} R_i}{N} \tag{4}$$

where k, Ri and N represent number of related discrete region for the class, number of the same region of feature value in different class, and total number of class.

Gain ratio, GR, evaluates ration of information gain for each feature.

$$GR = \frac{\sum_{i=1}^{c} -P_i \log_2 P_i - \sum_{v \in Value(A)} \frac{S_v}{S} \sum_{j=1}^{n} -P_j \log_2 P_j}{\sum_{v \in Value(A)} \frac{S_v}{S} \log_2 \frac{S_v}{S}} \tag{5}$$

where S, c, and P represent the set, class, and value probability belonged to the class.

Feature extractor then uses linguistic variables, high (H), mid (M), and low (L) to represent the value of above factors. Fig. 2 and Table 1 show fuzzy partition and fuzzy association matrix of linguistic variables. Feature extractor selects the features with higher correlation, that is, H, M, and ML, as the meaningful features.

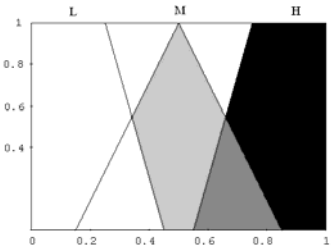

Fig. 2 Fuzzy partition of feature evaluation

Table 1 Fuzzy association matrix

	L	LM	M	HM	H
L	L	L	LM	LM	M
LM	L	LM	M	M	M
M	LM	M	M	M	HM
HM	LM	M	M	HM	H
H	M	M	HM	H	H

4. Case Miner

Case miner is responsible for extracting representative cases from raw data. First, it uses meaningful features and k-means algorithm [7. 9, 13] to partition raw data into different clusters. Each cluster contains different number of cases. Cases of each cluster represent the cases with similar feature values. Notably, the initial number of k is decided by average standard deviation (ASD) of each class.

$$ASD = \frac{\sum_{i=1}^{k} SD_i}{k} \tag{6}$$

where k and SD are the cluster number and standard deviation of each cluster. Case miner chooses the value of k in the curve slope of ASD with the largest descend value.

Case miner then uses genetic algorithm [7] to search representative cases from each cluster. Cases are represented by binary value regarding the chromosome. The initial population of genetic algorithm selects one case from each cluster randomly as candidate cases. Genetic algorithm then uses cross over, mutation, and reproduction operators to find representative cases. It uses case coverage, CV, to evaluate case difference for finding the similar cases by weighted distance function.

$$CV(X,Y) = (\sum_{i=1}^{n} w_i \times \|x_i - y_i\|)^{-1} \quad (7)$$

where x_i, y_i, n, and w_i are the i^{th} feature of case X, i^{th} feature of case Y, feature number, and i^{th} feature weight. Notably, *w* is the defuzzification of feature evaluation, that is, centroid value of FM. The higher case coverage means the more representative ability of the case. If case X is the representative case, case Y is covered by case X when CV is smaller than threshold θ. Finally, genetic algorithm will stop when all selected cases can cover 90% cases in each cluster.

5. Case Mining Application

The system uses new-thyroid and breast cancer data of UCI machine learning repository [8] to examine the case mining system. Table 2 shows raw data name, feature number, class number, data number, and cluster number of the above two raw data. Feature extractor evaluates FM value to find meaningful features. Figure 3 and Table 3 show evaluated value of PC, DA, and GR of new-thyroid data. It selects three features from raw data as dominant features, that is, T3-resin, Serum thyroxin, and Serum triiodothyronine.

Case miner then uses genetic algorithm to select representative cases. The initial population of candidate cases in each data set is 7 and 29 corresponding to cluster number. It uses 215 and 682 digitals to represent the chromosome. The crossover and mutation rate are 0.9 and 0.1. The experimental weights of data features are presented in Tables 4 and 5. Fig. 4 shows case coverage rate of new-thyroid. Finally, the system uses covering test and utilization statistics of representative case in each cluster to verify the completeness of case library. Table 6 summarizes the experimental results on the number of covered cases and utilization statistics in new thyroid data.

Table 2 Example raw data

Raw data name	New-thyroid	Breast cancer
Feature number	5	9
Class number	3	2
Data number	215	682
Cluster number	9	20

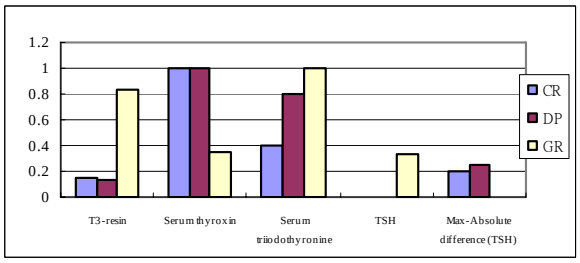

Figure 3 Fuzzy measurement value of new-thyroid

Table 3 Fuzzy measurement of new-thyroid

	T3-resin	Serum thyroxin	Serum triiodothyronine	TSH	Max-Absolute difference (TSH)
CR	Low	High	Medium	Low	Low
DP	Low	High	High	Low	Low
GR	High	Medium	High	Medium	Low

Table 4 Feature weight of new-thyroid

T3-resin	Serum thyroxin	Serum triiodothyronine	TSH	Max-Absolute difference (TSH)
0.66	0.5	0.9	0.42859718	0.388382994

Table 5 Feature weight of breast cancer

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion
0.515675375	0.9	0.9	0.690744528
Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli
0.716	0.681	0.683	0.669

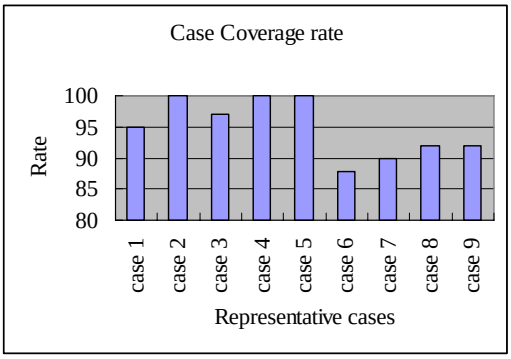

Fig. 4 Case coverage rate of new-thyroid

Table 6 Covering test and utilization statistics

Class name	Representative case	Covered cases/Total cases	Utilization statistics
Nrmal	Case 1	65/150	0.43
	Case 2	47/150	0.32
	Case 3	38/150	0.25
Hyper	Case 4	11/35	0.45
	Case 5	10/35	0.31
	Case 6	14/35	0.25
Hypo	Case 7	11/30	0.2
	Case 8	8/30	0.29
	Case 9	11/30	0.5

6. Conclusion

This paper proposes a case mining system for case library construction. The system uses fuzzy measurement to extract significant features of raw data. Fuzzy measurement computes the correlation, appearance times, and entropy of features in the class. The system then uses k-means algorithm to cluster raw data into different clusters. Each cluster is used to conduct gene evolution and case coverage evaluation. The system uses genetic algorithm to perform case mining. The fitness function of genetic algorithm uses weighted distance function to evaluate case coverage of representative cases.

The experimented results show that the system can construct case library correctly. The system can uses 3% to 5% representative cases to represent the raw data. It not only reduces the scale of case library dramatically but also improves the performance in case retrieval. A further study will be done on feature coverage to enhance the robustness of case library.

7. Acknowledgements

This work was partially supported by NSC of R. O. C. under grant 94-2745-E-030-004-URD.

8. References

[1] A. Aamodt and E. Plazaz, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approach," *AI Communications*, vol. 7, no. 1, pp. 39-52, 1994.

[2] R. Amen and P. Vomacka, "Case-Based Reasoning as a Tool for Materials Selection," *Materials and Design*, vol. 22, no. 5, pp. 353-358, 2001.

[3] N. Arshadi and I. Jurisica, "Data Mining for Case-Based Reasoning in High-Dimensional Biological Domains," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 8, pp. 1127-1137, 2005.

[4] F. T. S. Chan, "Application of a Hybrid Case-Based Reasoning Approach in Electroplating Industry," *Expert Systems with Applications*, vol. 29, no. 1, pp. 121-130, 2005.

[5] E. Cox, *Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration*, Morgan Kaufmann, San Francisco, 2005.

[6] Y. Fu and R. Shen," GA-Based CBR approach in Q&A system," *Expert Systems with Applications*, vol. 26, no. 2, pp. 167-170, 2004.

[7] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, New York, 2000.

[8] S. Hettich, C. L. Blake, and C. J. Merz, *UCI Repository of Machine Learning Databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science. (1998)

[9] J. Z. Huang, M. K. Ng, H. Rong and Z. Li, "Automated Variable Weighting in k-means Type Clustering," *IEEE Transactions on Pattern Analysis An Machine Intelligence*, vol. 27, no. 5, pp. 657-668, 2005.

[10] J. Kolodner, *Case-Based Reasoning*, Morgan Kaufmann, San Mateo, 1993.

[11] M. Lenz, S. B. Bartsch, and S. Wess, *Case-Based Reasoning Technology, from Foundations to Applications*, Springer, Berlin, 1998.

[12] Y. Li, S. C. K. Shiu, S. K. Pal, "Combining Feature Reduction and Case Selection in Building CBR Classifiers," *IEEE Transaction on Knowledge and Data Engineering*, vol. 18, no. 3, pp. 415-429, 2006.

[13] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

[14] Z. W. Ni, F. G. Li, and S. L. Yang, "Case-Based Maintenance on Outlier Data Mining," *Proc. of the International Conference on Machine Learning and Cybernetics*, pp. 2861-2864, 2005.

[15] S. K. Pal and C. K. Shiu, *Foundations of Soft Case-Based Reasoning*, Wiley, New Jersey, 2004.

[16] E. L. Rissland, "AI and Similarity," *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 39-49, 2006.