ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Classification in the Retrieval Phase of Case-Based Reasoning

Nabanita Choudhury
Dept. of Computer Science
Assam University
Silchar, India

Shahin Ara Begum
Dept. of Computer Science
Assam University
Silchar, India

Abstract: Case-based reasoning (CBR) is a problem solving technique that uses previous experiences to solve new problems. Among the four phases of CBR, Retrieval is the first and the most important phase, as it lays the foundation of the entire CBR cycle. Retrieval aims to retrieve similar cases from the case-base, given a new situation. CBR systems typically use a strategy called similarity-based retrieval for retrieving cases. One of the derivatives of similarity-based retrieval is k-nearest neighbor (k-NN) algorithm. In this paper, we compare the performances of k-NN, Fuzzy nearest neighbor (Fuzzy NN) and Genetic Programming (GP) classifiers for retrieval of cases. We evaluate these algorithms in WEKA, with benchmark data sets for classification from UCI.

Keywords: case-based reasoning; retrieval; k-nearest neighbor; fuzzy logic; genetic programming

I. INTRODUCTION

Case-based reasoning (CBR) is an Artificial Intelligence (AI) method for solving a new problem by reusing the solution of a past similar situation [1]. A new problem is known as a 'case' in CBR. CBR doesn't create a solution from scratch, rather it retrieves the best match from the case-base, and adapts the solution (if needed) to fit the current case [2]. The fact that CBR works in the way similar to the working of a human brain makes it intuitively appealing. As a result, it is useful in a large variety of problem domains, particularly in situations where the knowledge is incomplete and/or evidence is sparse or when situations (cases) recur [3], [4]. CBR has successfully been implemented in many application domains including medical diagnosis [5], help-desk service [6], product recommendation [7], and classification [8].

Case retrieval is the first phase of CBR cycle, and is often considered to be the most important phase [9], as a wrong retrieval by the CBR system would eventually produce wrong solutions. So, in order to carry out an efficient retrieval, a number of approaches have been suggested and implemented. Among these, the most commonly used approach is Similarity-based Retrieval (SBR) [10], and is implemented using k-nearest neighbor algorithm, or k-NN [11]. The other retrieval techniques include decision trees, and their derivatives [12]. Similarity metrics are developed using these techniques, which allow the distance among the cases to be measured. A major drawback of k-NN is its lazy learning approach and the biased value of k [11]. So, an alternative is Fuzzy Nearest Neighbor (Fuzzy NN) [13]. In this paper, we carry out a detailed comparative analysis of k-NN, Fuzzy NN, and Genetic Programming (GP) classifiers using experimental evaluation on data sets from UCI ML Repository.

The rest of the paper is organized as follows. Section 2 presents a brief insight into the phases of a CBR cycle, with a discussion on the retrieval phase. Section 3 briefly describes various techniques used for case retrieval. It also describes the three classifier algorithms used by us *viz*. k-NN, Fuzzy NN and GP. Section 4 describes the experimental evaluation and the results obtained. Section 5 presents concluding remarks and suggests topics for further research.

II. CASE-BASED REASONING

Given a new situation (case), CBR basically follows a four step process, known as the R^4 cycle [1].

A. *The R^4 CBR Cycle*

The CBR Cycle consists of the following four "R's":

- Retrieving the most similar cases
- Reusing the cases
- Revising the proposed solution if needed
- Retaining the new solution as a new case

Given a new case, it is matched against the cases stored in the case base. After a suitable matching is performed, similar case(s) are retrieved. If there is an exact or close match, the solution is reused, else the solution is adapted, which in turn produces a new case. Therefore, the CBR cycle mainly consists of case retrieval, case adaptation (combining the reuse and/or revise phases), and maintenance of the case base. Usually adaptation is a complicated process and it is highly domain dependent. As a result, the formulation of general case adaptation rules is very difficult [14]. So the need for an efficient retrieval mechanism arises, in order for the CBR system to succeed.

B. *Retrieval*

Case retrieval is the process of finding the cases that are closest to the current case within a case base. The success of a CBR system typically depends on its retrieval phase. As depicted in Figure 1, the closest the match is, the lesser is the retrieval distance, thereby easier it is to perform the adaptation.

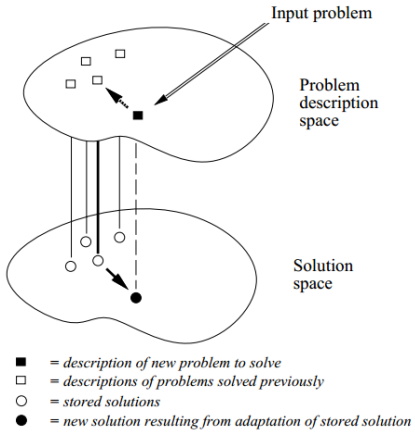

Figure 1. CBR Problem Solving [15]

Retrieval is considered to be a major research area in CBR. When the CBR system encounters a new case, retrieval is carried out using a selection criterion for determining how a case is to be chosen for retrieving from the case base.

III. RETRIEVAL TECHNIQUES

The most commonly investigated retrieval techniques include inductive approaches, nearest neighbor retrieval, knowledge guided approaches, and validated retrieval [12], [16]. Some hybrid algorithms have also been proposed in the literature, e.g. Discretised Highest Similarity with Pattern Solution Re-use algorithm [17]. CBR systems generally use a strategy called Similarity-based retrieval (SBR). In SBR, the usefulness of cases for solving a new problem is approximated by the similarity knowledge [18]. Some of the other approaches include Context guided retrieval [19], Adaptation guided retrieval [20], Diversity conscious retrieval [21]-[23], Compromise-driven retrieval [24], Order based retrieval [25], Explanation oriented retrieval [26].

Similarity measures play a very important role in CBR. The effectiveness of these measures determine the usefulness of a retrieved case for solving a new problem. Given a case base CB = $\{c_1,\ldots, c_n\}$ of objects, a similarity measure and a new problem p, the following are to be retrieved [27]:

- the object c_i that is most similar to p,
- or the k most similar objects $\{c_1,\ldots,c_k\}$
- or all objects c_i that have least a minimum similarity

SBR is typically implemented through k-nearest neighbor retrieval or simply k-NN [9]. k-NN works on the idea that to solve a new problem p, k most similar cases to p are obtained, in order to solve p.

A. *K-Nearest Neighbor*

Given an input case, the k-nearest neighbor algorithm involves searching for the k nearest cases similar to the current case using a distance measure. The class of the majority of these k cases is then selected as the retrieved class [28]. It is a lazy learning method, and the classification rate is dependent on the chosen value of k [11]. A major disadvantage of k-NN is that enormous computation is needed when there are large number of cases in the case-base, or when the number of feature dimensions is large [29].

B. *Fuzzy Nearest Neighbor*

Instead of assigning an input sample vector to a class [30], a class membership is assigned to the sample vector by the Fuzzy counterpart of the nearest neighbor algorithm called Fuzzy Nearest Neighbor (Fuzzy NN). The advantage of this is that it doesn't make any arbitrary assignments [31].

C. *Genetic Programming*

Genetic Programming (GP) [32] is an AI technique based on the evolutionary process of the naturally occurring substances. GP allows us to use complex pattern representations e.g., decision trees, classification rules, discriminant functions, and many more, thus proving itself to be a good classification technique. Moreover, GP allows the adaption of technique according to each particular problem to be solved. This makes it highly flexible [33].

IV. EVALUATION

A. *Experimental Setup*

We intend to show that the implementation of Fuzzy NN and GP improve the retrieval performance of a case-base, over the k-NN approach. The retrieval accuracy of the CBR system can be determined by the classification made, given a new case. So in this paper, classification problem is chosen as the target application task. The case-based approach for classification is defined in [34] as - 'Given a new problem (a case C), the system retrieves set of cases from a case base and classifies the new problem based on the retrieved matches.'

Based on this, we apply k-NN, Fuzzy NN and GP on the data sets obtained from UCI ML Repository [35], described in Table 1.

Table I. Basic Information on the Data Sets from UCI

Data Sets	*Attribute Type*	*No. of Attributes*	*No. of Instances*	*No. of Classes*
Breast Cancer (Wisconsin)	Integer, Real	11	699	2
E Coli	Real	8	336	8
Hypothyroid	Integer, Real	30	3772	4
Iris	Real	5	150	3
Liver Disorders	Categorical, Integer, Real	7	345	2
Pima Indians Diabetes	Integer, Real	9	768	2

For carrying out the comparative analysis, we have chosen IBk, a k-NN algorithm available in WEKA [36], Fuzzy NN [37], and GP. For retrieving cases from a case base, there are two stages of classification – firstly, a set of similar cases is retrieved and secondly, the new problem is classified using the solutions. We focus on the first stage. We use two metrics for evaluation of the classifiers *viz.* classification accuracy and F-measure. Classification accuracy is the proportion of correctly classified instances, and is often assumed to be the best indicator of performance for classifiers, but it ignores the cost incurred in making a wrong decision. So we use F-measure as well. F-measure is defined as the harmonic mean between precision and recall.

B. Results and Analysis

IBk, Fuzzy NN and GP classifiers are compared by testing them on the five data sets as listed in Table 1. We use 10-fold cross-validation. Table 2 and Table 3 detail the results obtained. The better value for each data set is denoted in boldface.

Table II. Performance Comparison on Classification Accuracy

Data Sets	*Classification Accuracy (%)*		
	IBk	*Fuzzy NN*	*GP*
Breast Cancer (Wisconsin)	95.1359	63.3763	**95.7082**
E Coli	80.3571	**87.2024**	83.631
Hypothyroid	91.5164	**92.2853**	92.2587
Iris	95.3333	**96.6667**	94.6667
Liver Disorders	62.8986	68.4058	**71.0145**
Pima Indians Diabetes	70.1823	73.0469	**74.6094**

Table III. Performance Comparison on F-measure

Data Sets	*F-measure*		
	IBk	*Fuzzy NN*	*GP*
Breast Cancer (Wisconsin)	0.951	0.625	**0.957**
E Coli	0.801	**0.865**	0.835
Hypothyroid	**0.911**	0.886	0.886
Iris	0.953	**0.967**	0.947
Liver Disorders	0.629	0.676	**0.706**
Pima Indians Diabetes	0.698	0.725	**0.737**

Through Table 2, we find that in terms of classification accuracy, i.e., the percentage of cases correctly classified, Fuzzy Nearest Neighbor outperforms IBk and GP for E Coli, Hypothyroid and Iris data sets; and GP has the highest classification accuracy for Breast Cancer (Wisconsin), Liver Disorders and Pima Indian Diabetes data sets.

From Table 3, we observe that in terms of F-measure, IBk has higher score only for Hypothyroid data set. GP outperforms Fuzzy NN and IBk for Breast Cancer (Wisconsin), Liver Disorders and Pima Indian Diabetes data sets. Fuzzy NN has higher F-measure for E Coli and Iris data sets.

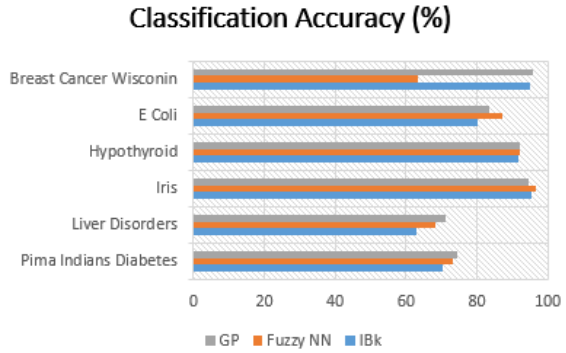

Figure 2. Comparison of Classification Accuracy

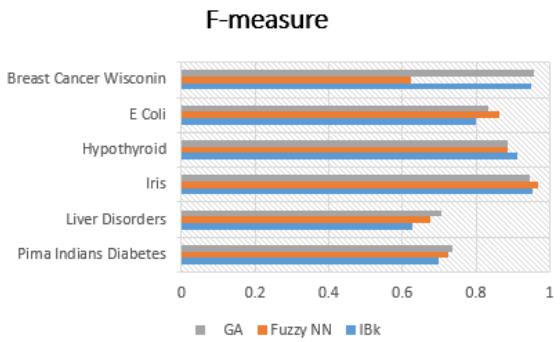

Figure 3. Comparison of Classification Accuracy

The comparative results are depicted in Figure 2 and Figure 3. From this, it can easily be construed that for classification purpose, though Fuzzy NN and GP outperform the classical k-NN (IBk in our implementation), none of the classifiers perform well for a variety of data sets. The performance of the classifiers depend much on the number of attributes and other properties of the data sets.

V. Conclusion and Future Scope

Retrieval is often considered to be the most importance phase of the CBR cycle, as the performance of a CBR system depends on the accuracy of retrieval. Retrieval is most commonly implemented through similarity measurements, often realized through the Nearest Neighbor algorithm. Classification is the first step of retrieval, wherein a new case is classified based on the retrieved cases. This paper compares the performances of three well known classifiers *viz.* k Nearest Neighbor (k-NN), a Fuzzy derivative of k-NN called Fuzzy NN, and Genetic Programming (GP). We evaluate the algorithms in WEKA, using data sets from UCI ML Repository. The experimental results show that though IBk is outperformed by Fuzzy NN and GP, the performance of Fuzzy NN and GP depend on the properties of the data set. Neither of the two algorithms perform consistently well for all the data sets.

In CBR, most of the systems built are retrieval only, as automatic adaptation is not achieved for all the application fields. So the retrieval mechanism must be strong enough to correctly identify the nearest match of a new case from the case-base. From our study, we construe that a single algorithm doesn't perform well for a variety of data sets. As classification forms the backbone of retrieval, further research can be carried out on a hybrid combination of the classification algorithms, to be implemented in retrieval, which is the first and the most important phase of CBR.

VI. References

[1] A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," AI communications, vol. 7, no. 1, pp.39-59, 1994.

[2] L. D. Xu, "Case based reasoning," IEEE potentials, vol. 13, no. 5, pp.10-13, 1994.

[3] J. L. Kolodner, "An introduction to case-based reasoning", Artificial Intelligence Review, vol. 6, no. 1, Kluwer Academic Publishers, pp. 3–34, 1992.

[4] P. Cunningham, "CBR: Strengths and weaknesses", Tasks and Methods in Applied Artificial Intelligence,

Lecture Notes in Computer Science, vol. 1416, Springer Berlin Heidelberg, pp. 517–524, 1998.

[5] H. Ahn and K. Kim, "Global optimization of case-based reasoning for breast cytology diagnosis," Expert Syst. Appl., vol. 36, pp. 724–734, 2009.

[6] Y. B. Kang, A. Zaslavsky, S. Krishnaswamy, and C. Bartolini, "A knowledge-rich similarity measure for improving it incident resolution process," Proceedings of the 2010 ACM SAC, pp. 1781–1788, 2010.

[7] K. Bradley and B. Smyth, "Personalized information ordering: a case study in online recruitment," Knowledge-Based Systems, vol. 16, pp. 269–275, 2003.

[8] M. Nilsson, P. Funk, and M. Sollenborn, "Complex Measurement Classification in Medical Applications Using a Case-Based Approach," Ashley, K.D., Bridge, D.G. (eds.) ICCBR 2003. LNCS, vol. 2689, pp. 63–73. Springer, Heidelberg, 2003.

[9] R. L. De Mantaras et al., "Retrieval, reuse, revision and retention in case-based reasoning," The Knowledge Engineering Review, vol. 20, no. 03, pp. 215-240, 2005.

[10] B. Smyth and M T. Keane, "Adaptation-guided retrieval: questioning the similarity assumption in reasoning," Artificial intelligence, vol. 102, no. 2, pp.249-293, 1998.

[11] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," OTM Confederated International Conferences, On the Move to Meaningful Internet Systems, pp. 986-996, Springer Berlin Heidelberg, 2003.

[12] S. K. Pal and S. C. Shiu, "Foundations of soft case-based reasoning," vol. 8, John Wiley & Sons, 2004.

[13] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," IEEE transactions on systems, man, and cybernetics, vol. 4, pp.580-585, 1985.

[14] R. Mitra and J. Basak, "Methods of case adaptation: a survey," International journal of intelligent systems," vol. 20, no. 6, 627-645, 2005.

[15] D. B. Leake, "CBR in context: The present and future," Case-Based Reasoning, Experiences, Lessons & Future Directions, pp.1-30, 1996.

[16] E. Simoudis and J. Miller, "Validated retrieval in case-based reasoning," AAAI, pp. 310-315, 1990.

[17] D. W. Patterson, N. Rooney, and M. Galushka, "Efficient Retrieval for Case-Based Reasoning," FLAIRS Conference, pp. 144-149, 2003.

[18] Y. B. Kang, S. Krishnaswamy, and A. Zaslavsky, "A case retrieval approach using similarity and association knowledge. In OTM Confederated International Conferences", On the Move to Meaningful Internet Systems, pp. 218-235, Springer Berlin Heidelberg, 2011.

[19] I. Watson and S. Perera, "A hierarchical case representation using context guided retrieval," Knowledge-Based Systems, vol. 11, no. 5, pp.285-292, 1998.

[20] B. Smyth and M. T. Keane, "Adaptation-guided retrieval: questioning the similarity assumption in reasoning," Artificial intelligence, vol. 102, no. 2, pp.249-293, 1998.

[21] B. Smyth and P. McClave, "Similarity vs. diversity," International Conference on Case-Based Reasoning, pp. 347-361, Springer Berlin Heidelberg, 2001.

[22] D. McSherry, "Diversity-conscious retrieval," European Conference on Case-Based Reasoning, pp. 219-233, Springer Berlin Heidelberg, 2002.

[23] B. Mougouie, M. Richter, and R. Bergmann, "Diversity-conscious retrieval from generalized cases: A branch and bound algorithm," Case-Based Reasoning Research and Development, pp.1064-1064, 2003.

[24] D. McSherry, "Similarity and compromise," International Conference on Case-Based Reasoning, pp. 291-305, Springer Berlin Heidelberg, 2003.

[25] D. Bridge and A. Ferguson, "Diverse product recommendations using an expressive language for case retrieval," European Conference on Case-Based Reasoning, pp. 43-57, Springer Berlin Heidelberg, 2002.

[26] D. Doyle, P. Cunningham, D. Bridge, and Y. Rahman, "Explanation oriented retrieval," European Conference on Case-Based Reasoning, pp. 157-168, Springer Berlin Heidelberg, 2004.

[27] M. M. Richter and R. O. Weber, "Case-Based Reasoning – A Textbook," Springer-Verlag Berlin Heidelberg, 2013.

[28] L. E. Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no.2, pp.1883, 2009.

[29] S. K. Pal and A. Pal (eds.), Pattern Recognition: From Classical to Modern Approaches, World Scientific, Singapore, 2001.

[30] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inform. Theory, vol. IT-13, pp. 21-27, Jan. 1967.

[31] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," IEEE transactions on systems, man, and cybernetics, vol.4, pp. 580-585, 1985.

[32] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone, "Genetic Programming—An Introduction", On the Automatic Evolution of Computer Programs and its Applications. San Mateo, CA/Heidelberg, Germany: Morgan Kaufmann/dpunkt.verlag, 1998

[33] P. G. Espejo, S. Ventura, F. Herrera, "A survey on the application of genetic programming to classification," IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 40, no. 2, pp.121-144, 2010.

[34] I. Jurisica and J. Glasgow, "Case-Based Classification Using Similarity-Based Retrieval," International Conference on Tools with Artificial Intelligence, pp. 410, 1996.

[35] M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[36] E. Frank, M. A. Hall, and I. H. Witten, The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Fourth Edition, 2016.

[37] R. Jensen and C. Cornelis, "Fuzzy-rough nearest neighbour classification," Transactions on rough sets XIII, pp.56-72, 2011.