Avelino J. Gonzalez / Lingli Xu / Uma M. Gupta

VALIDATION TECHNIQUES FOR CASE-BASED REASONING SYSTEMS

Abstract

Case-Based Reasoning (CaBR) systems, by their nature, have a built-in set of test cases in their case library. Effective use of this unusual feature can facilitate the validation process by minimizing the involvement of domain experts in the process. This can reduce the cost of the validation process, and eliminate the subjective component introduced by experts. This article proposes a validation technique which makes use of the case library to validate the CaBR system. Called the *Case Library Subset Test Technique* (CLST), it evaluates the correctness of the retrieval and adaptation functions of the CaBR engine with respect to the domain as represented by the case library. It is composed of two phases, 1) the Retrieval Test, and 2) the Adaptation Test. A complete description of the technique, as well as an application of the technique to validate an existing CaBR system are discussed in this paper.

1.0 INTRODUCTION

Validation of knowledge-based system has received great attention from researchers in the last several years.[Gupta, 1992] The importance of ensuring that fielded knowledge-based systems operate correctly and as intended has been recognized by developers as well as users. However, the majority of the reported validation work to date has centered around rule-based systems. This may be because in comparison to other artificial intelligence techniques, rule-based expert systems are the most mature as well as the most commercially available. In fact, the majority of the fielded systems in existence today are of this type.

Published literature that deals with validation of Case-Based Reasoning (CaBR) systems is indeed scarce. O'Leary addresses the problem of CaBR validation in his 1993 article [O'Leary, 1993], and provides a valuable insight into the problem by discussing the issues involved. We skip such discussions and refer the interested reader to that source for an in-depth analysis of these issues. However, O'Leary stops short of actually proposing and testing a detailed evaluation technique. This paper describes a technique that can be used to validate a Case-based Reasoning system with little need for involvement of an expert.

In addition to the work by O'Leary cited above, Simoudis [1990] combined simple retrieval with domain specific validation of retrieved cases to produce a tool for CaBR systems. The validation, however, was not considered as an independent phase in the system development. It is designed into the retrieval phase, and is called a validated retrieval model in CaBR.

Ram [1993] investigated systematic evaluation of the design decision in a CaBR system. He emphasized the complexity of the domain choice for a case-based system and system behavior criteria

The above methods described required the derivation of a complex mathematical model to serve as the validation criteria. Additionally, other published validation efforts for case-based systems, Protos, HYPO, and Clavier, (as discussed in [O'Leary, 1993]) made extensive use of experts. None of these systems take advantage of the unique characteristic of CaBR, which is that the expertise is built-in from explicit historical cases.

Taking advantage of this feature of CaBR systems, Yi [1995] developed a set of algorithms in her work to build and validate a case-based reasoning system to help predict software development cost. The retrieval and adjustment algorithms in the case library were implemented to meet a specified *Minimum Relative Error (MRE)*. The MRE is the percentage difference of the system estimation to real software cost. However, her work came short of actually developing a full validation technique for general use.

The developers of Battle Planner (as discussed in [O'Leary, 1993]) also make use of the case library as a source of expertise by using some of the cases for testing. We expand upon this idea in our work.

The gold standard in most knowledge-based system validation efforts is considered to be the expert's knowledge. Some problems with this criteria, however, is that it is typically quite costly to involve experts due to their general unavailability and high salaries. The research presented in this paper, however, minimizes the need for intensive domain expert involvement in the validation process. The gold standard chosen for validation will be the case library itself, as it represents the explicit collection of historical results. The technique, called the *Case Library Subset Test, (CLST)* uses Yi's Relative Error (RE) as the comparison medium. It does, however, use experts to a small degree in determining the validation criteria to be employed. However, as this is typically within the purview of the user/purchaser, and not of the development team, we can safely state that expert involvement is not necessary for this procedure. This novel technique is evaluated in the re-validation of the *Case-Based Appraiser* (CBA) [Gonzalez, 1992; Laureano-Ortiz, 1990], a CaBR system used to appraise real estate property.

2.0 CASE LIBRARY SUBSET TEST METHODOLOGY

This section describes the proposed validation technique for CaBR systems called the *Case Library Subset Test* (CLST) technique.

The main concept underlying this validation method is the selection of a subset of cases from the case library and using this subset as a test set to evaluate the effectiveness of the system's retrieval and adaptation features. The comparison standards of the test set are considered to be correct because they are part of the case library. But first, the validation criteria has to be selected, as it affects the final correctness of the systems. This process is described below.

2.1 Determination of Validation Criteria

The first task is to develop a validation criteria. This consists of determining two basic parameters, the *Result Acceptability Criteria (RAC),* and the *System Validity Criteria (SVC)* The RAC serves to determine whether an individual test case has been solved correctly by the CaBR system. It mandates that the distance between the system solution to a test case, and the benchmark standard to which it is compared be calculated. If the solution is provided in numerical terms, then the Relative Error (RE) can be the percent difference between the two quantities. If, on the other hand, the output of the CaBR system is symbolic or Boolean, then *optimal*, *acceptable* and *unacceptable* solutions may be defined as the benchmark standard may allow. The SVC serves to determine whether, in light of the executed and evaluated suite of test cases, the system can be considered valid. The SVC requires that upon completion of all testing, the percent of all acceptable test cases be greater than its value before the CaBR system can be considered valid. The RAC and the SVC are typically obtained from either experts or users, and it may be defined in the requirements specification. Upon selection of the above validation criteria, the CLST technique begins as described below.

2.2 Description of the Case Library Subset Test

These are described in more detail below.

2.2.1 CaBR Retrieval Test.

Case indexing and case classification issues are intended to improve the effectiveness and efficiency of case retrieval and to reduce the complexity of similarity calculations. The correctness of the retrieval process is, therefore, one of major concerns in CaBR systems. The CaBR Retrieval Test is designed to evaluate the correctness of the retrieval function. The indexing system used, although not evaluated independently, is clearly part of the retrieval evaluation test, and deficiencies in indexing will show up as poor retrieval performance. The comparison function is also likewise validated.

Briefly, the Retrieval Test requires that each historical case in the case library "spawn" a test case identical to itself in all ways. A pointer to the historical case is maintained for the purpose of comparison later. This process generates a set of test cases, not only for the retrieval test, but also for the adaptation test as will be seen later. As part of the retrieval test, each test case is, in turn, presented to the CaBR system as the current case. The CaBR system goes through the comparison and retrieval processes, arriving at an internal list of library cases ranked in decreasing order of similarity. In order for any test case to be marked as successfully executed, the historical case which spawned the current test case should be found as the top-ranked historical case in this internal list, and the similarity distance should be the minimum allowed in the chosen measuring scheme (or very close to it).

2.2.2 CaBR Adaptation Test

The Retrieval Test ensures that the comparison and retrieval functions are correctly carried out. It is the purpose of this test to ensure that adaptations are properly made from valid retrieved cases. Therefore, the Adaptation Test should only be done after a successful Retrieval Test.

The test case set used here is the same as that of the retrieval test (e.g., spawned from each historical case in the case library). The significant difference is that in the Adaptation Test, the historical case corresponding to the test case being presented to the CaBR system is removed from the case library. Thus, if a case library has N cases in it, the modified case library will only contain N-1 cases at all times. The outputs of this test include retrieved cases, the final solution, and its RE. Although the test case is not longer in the case library, the CaBR system retrieves the most similar case from the case library and adjusts the closest matching case(s) with the adaptation strategies to obtain the final solution to this test case. Since the retrieval process has already been validated, this test isolates and evaluates the adaptation process of the CaBR system

3.0 IMPLEMENTATION AND EVALUATION OF THE CASE LIBRARY SUBSET TEST TECHNIQUE

It is important that any new concept in science and engineering be evaluated to determine its effectiveness in solving the problem it addresses. An evaluation technique should be no exception. In this section we briefly describe the steps taken to evaluate the CLST technique described in the previous section. The testbed chosen to carry out this evaluation is a CaBR prototype system for residential property appraisal called the *Cased-Based Appraiser (CBA)* [Gonzalez, 1992; Laureano-Ortiz, 1990]

3.1 The Case-Based Appraiser System

An prototype that automates property appraisal using a CaBR

approach therefore was developed by Laureano-Ortiz [1990]. Several attributes in the cases are used to calculate the price of the property. Some of these are the living area, number of bedrooms, number of bathrooms as well as others.

The CBA System works by determining the most similar cases to the current property, adjusting these cases to account for any remaining similarities, and then obtaining the appraised value using one of two widely accepted methods in property appraisal. Refer to [Laureano-Ortiz 1990] and to [Gonzalez, 1992]for more details on this system.

3.2 Determination of Validation Criteria

The task of determining the validation criteria was the first undertaken. A questionnaire was sent to individuals knowledgeable in the field of appraisal with the following questions:

Question 1: "What is the maximum acceptable Relative Error of a CBA system?" The maximum acceptable error range refers to the percent difference between appraised price and the actual real sold price. This corresponds to the Result Acceptability Criteria (RAC) defined in Section 2 above

Question 2: "What percentage of the correct appraised cases in a CBA system is considered reasonable"? This question asks for the Correctness Ratio (CR) in all appraised cases. This criteria corresponds to the System Validity Criteria (SVC) seen in section 2 above

In regards to the RE, the majority of the responders felt that 20% was appropriate, based on the limited set of attributes considered. Human bargaining, the seller's economic situation, and the various marketing factors are not considered in the CBA analysis. Yet, the actual price is strongly affected by those factors. Therefore, it was decided that 20% RE would be set as the RAC (the threshold between acceptable and unacceptable results from the CBA).

Likewise, the total CR was determined to be a minimum of 80% SVC for a valid system. That is, 80% of the subject properties to be appraised were valued at a price less than 20% different than the actual sale price. It is necessary to note here, however, that the CBA system sometimes displays a dummy "-1" as the result when the subject case does not have any similar cases in the case library (e.g., little similarity between the most similar historical case and the current case). This can be quite a normal occurrence in CaBR systems, and these subject properties should not be considered when determining the CR.

4.0 EVALUATION OF RESULTS FOR THE CASE LIBRARY SUBSET TEST TECHNIQUE

This section evaluates a rewritten version of the CBA prototype (in C/C++) with the Case Library Subset Test validation model. Since the new C/C++ prototype used exactly the same algorithms and data types as the original lisp-based version by Laureano-Ortiz, we shall assume that they functionally identical. The CLST validation model is designed to empirically demonstrate that the CBA system works correctly for property appraisal. However, the true purpose of this section is to evaluate the validity and usefulness of the Case Library Subset Test Technique itself, as its effectiveness in validating the CBA will be compared with Laureano-Ortiz's [1990] original expert-based validation of the lisp-based CBA. We refer the reader to [Xu, 1995] for all the raw data pertaining to the results shown in this section.

4.1 Retrieval Test

The results of the Retrieval test on the CBA case library showed that in

100% of the test cases, the case in the case library corresponding to the test case was chosen as the closest match. This indicates that retrieval was done properly. Furthermore, the average RE for the Retrieval Test was 8.239%. The reader should note that the appraisal value is computed by averaging the several most similar cases in the library, rather than exclusively using the most similar one, even if it is identical. Thus, the RE for this test with the appraisal testbed should not be expected to be 0%.

4.2 Adaptation Test

In the Adaptation test, the retrieval technique is also inherently evaluated, as the RE is also calculated. However, the emphasis here is on the adaptation aspect. The average RE for the Adaptation test was computed to be 13.2057%. A higher RE than for the Retrieval Test is to be expected, as the case base is somewhat less similar to the test cases by virtue of removing the case corresponding to the test case.

4.3 Original, Expert-based Validation of CBA

Laureano-Ortiz [1990] evaluated his original CBA system according to the traditional method of comparing the CBA's output to the domain expert's appraisal results for the same set of test cases. In his validation exercise, seventy (70) test cases were presented to the domain experts, and their appraised values for those test cases were recorded. The same test cases were presented to the original CBA, and its results were recorded. Using the original data, the RE for each of the 70 test cases was computed as part of the present investigation.

The relative error (RE) was computed for the results obtained in the original CBA validation. We found that 11.4% of the test cases had a computed RE of more than 20%. This calculation excluded cases that resulted in a "-1" as indicated above. We also calculated the average RE of the original CBA validation test (excluding the "-1" answers) to be 9.91123%.

The average RE of the Test Case set used by Laureano-Ortiz (computed to be 9.9%) was deemed to be acceptable by the experts that were involved in the original validation process. The case library (107 cases) and the test case set (70 cases) used in that evaluation did not have any cases in common. However, they were obtained from real-world data, and thus realistic in their makeup. Laureano-Ortiz [1990] concluded his CBA system test by stating that the "CBA does a good job given the limitations it has in its condition of prototype in its early stage: small case base, small number of represented figures and lack of better sources of cases."

In the CLST, the same case library used in the original evaluation (107 cases) was employed. However, the test case set was the case library itself. Since the case library is real-world data, it can be said that these test case sets were generally of equivalent makeup as the 70 case test set used in the original evaluation. Thus, the use of the case library itself as the source of test cases can be considered to be equally realistic as the original (70 case) test case set, and thus acceptable. The average RE was computed to be 8.2% for the Retrieval Test, and 13.2% for the Adaptation Test. Since the Adaptation Test is more similar to the original test run by Laureano-Ortiz than the Retrieval Test, that is the number to which we compared the performance of the CLST technique. However, it should be noted that the Adaptation Test always removes the most similar of cases in the case library, thus introducing a slight disadvantage. Nevertheless, regardless of which number is used for comparison, the numbers are quite close to each other, suggesting strongly that the CLST technique is an effective way of validating CaBR systems.

5.0 SUMMARY AND CONCLUSION

In this investigation, a new method called Case Library Subset Test is presented for the purposes of validating Case-Based Reasoning Systems without the need for involving a domain expert. A prototype system which carries out the CLST testing technique automatically was designed, built and successfully demonstrated on a testbed CaBR system, the Case-Based Appraiser (CBA) for appraising single family residential property.

The implementation of Case Library Subset Test techniques presented here is a realization of a new validation idea. We believe that the methodology presented here is not only applicable to small CaBR system like the CBA, but also to validation of more complex systems.

6.0 REFERENCES

[Gonzalez, 1992] Gonzalez, A. J. and Laureano-Ortiz, R., (1992) "A Case-Based Reasoning Approach to Real Estate Property Appraisal", *Expert Systems with Applications*, Vol. 4, No. 2, pp. 229-246.

[Gupta, 1992] Gupta, U. G. (ed), (1992) Validating and verifying knowledge-based systems, Las Alamedas, CA: IEEE Computer Society Press.

[Laureano-Ortiz, 1990] Laureano-Ortiz, R., (1990) "Application of case-based reasoning techniques to the automation of single-family residential property appraisals", Master's Thesis, Department of Electrical and Computer Engineering, University of Central Florida, December.

[O'Leary, 1993] O'Leary, D. E., (1993) "Verification and validation of case-based systems" *Expert Systems with Applications*, Vol. 6, pp. 57-66.

[Ram, 1993] Ram, A., "Indexing, Elaboration and Refinement: Incremental Learning of Explanatory Cases", Machine Learning, 1993, pp. 7-54.

[Simoudis, 1990] Simoudis, E. and Miller, J., (1990) "Validated retrieval in case-based reasoning", *Proceedings of the 1990 National Conference on Artificial Intelligence*, AAAI Press, pp. 310-317.

[Xu, 1995] Xu, L., (1995) "Validation techniques for case-based reasoning systems", Master's Thesis, Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL, Fall Term.

[Yi, 1995] Yi, Y., (1995) "CASHEEW", Master's thesis, University of Central Florida of Department of Electrical and Computer Engineering, Spring Term.

Avelino J. Gonzalez / Lingli Xu
Electrical and Computer Engineering Dept.
University of Central Florida
Orlando, FL 32816-2450

Uma M. Gupta
Decision Sciences Department
East Carolina University
Greenville, NC 27858-4353