in working papers of 2nd. European Workshop on Case-Based Reasoning. Chantilly, France, November 1994.

A Comparison of Incremental Case-Based Reasoning and Inductive Learning

Barry Smyth

Hitachi Dublin Laboratory,
16 Westland Row,
Dublin 2,
Ireland.
E-Mail: barry@hdl.ie
Phone: +353-1-6798911

Pádraig Cunningham

Department of Computer Science,
Trinity College Dublin,
Dublin 2,
Ireland.
E-Mail: Padraig.Cunningham@cs.tcd.ie
Phone: +353-1-702 1765

Abstract

This paper focuses on problems where the reuse of old solutions seems appropriate but the conventional CBR methodology is not adequate because a complete description of the new problem is not available to trigger case retrieval. We describe an information theoretic technique that solves this problem by producing focused questions to fill out the case description. This use of information theoretic techniques in CBR raises the question of whether a standard inductive learning approach would not solve this problem adequately. The main contribution of this paper is an evaluation of how this incremental case-based reasoning compares with a pure inductive learning approach to the same task.

1 Introduction

The CBR strategy described in this paper is incremental in the sense that the target case description is built up during the case retrieval process. This incremental strategy is motivated by the fact that there are many potential applications for CBR for which it is expensive to determine the predictive features of the case. Our system produces focused questions for the user, only requesting features that are useful in retrieving good matches from the case-base.

This work was motivated by an attempt to reengineer a model-based fault diagnosis system as a case-based reasoning system. One of the great strengths of the model-based system was that the main part of its reasoning mechanism was goal driven so the operator was only asked to perform tests that would contribute to a particular diagnosis. In this problem domain the data used in diagnosis seems naturally to fall into two categories, free data and expensive data. The model-based system only requested such expensive data as was needed.

So our CBR implementation has free case features and expensive case features. In this paper we describe an information theoretic approach to incremental case retrieval where only a subset of the expensive features are requested from the user. In an earlier paper (Cunningham & Smyth, 1994) we have compared this incremental CBR with the model-based approach and found that less data was required for correct diagnosis than was the case with the model-based approach.

These information theoretic criteria bear the hallmarks of an inductive learning approach to the problem so in this paper we evaluate our incremental CBR by comparison with a pure inductive learning approach; this evaluation is presented in section 4. But first we start in section 2 with a look at problems that require incremental CBR. Then in section 3 we describe the incremental CBR strategy itself.

2 Problems requiring incremental CBR

Our experience with CBR applications, particularly in the area of diagnosis, convinces us that there is an important class of problems for which it is difficult to determine the predictive features of the target case in advance of case retrieval. In these situations some features may be readily available and some are more expensive to collect. In the rest of this paper we will use the following notation for these sets of features:

$\mathbf{F}$ is the set of all features that can describe a case

$\mathbf{F} = \mathbf{I} \cup \mathbf{F}$ where

$\mathbf{I} = \{I_1, ..., I_i\}$ the set of free (inexpensive) features

$\mathbf{F} = \{F_1, ..., F_f\}$ the set of features that are expensive to determine

The importance of this observation is that, of the features that are expensive to obtain, only some may actually be needed to determine a suitable case (or cases) for retrieval. Machine learning research distinguishes *characteristic descriptions* from *discriminant descriptions* (Michalski, 1983) and where features are expensive to determine a discriminant description is sufficient for case retrieval. If the case representation is a good one then $\mathbf{F}$ is a characteristic description. However some subset of $\mathbf{F}$ may be adequate as a discriminant description. In particular, **I** plus some subset of **F** may be adequate for retrieval. The challenge for incremental CBR is to generate queries that will require the user to provide a minimal subset of **F**.

Before introducing our mechanism for incremental CBR we will describe some potential CBR applications that fit this characterisation.

2.1 Diagnosis

Much of the work on incremental CBR described here focused on the re-engineering of a model-based system for fault diagnosis of power-supplies called NODAL (Cunningham & Brady '87; Cunningham '88). The main components of this system are shown in Figure 1.

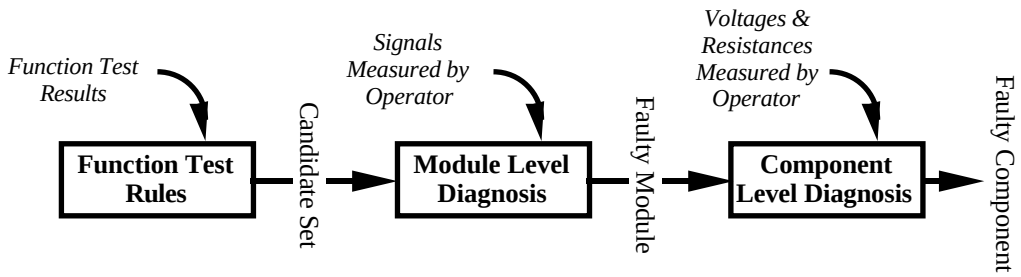

Figure 1. The main components in the NODAL fault diagnosis system.

NODAL is designed to operate in a repair shop so the first input (Figure 1) in the diagnosis are the results from the test equipment on which it was confirmed that the unit was faulty. These function tests are performed on the unit as a 'black box', and measure outputs associated with test inputs. These tests will number between twenty and forty depending on the complexity of the circuit. However, because the internals of the unit are not being examined, the amount of diagnostic information that they carry is limited. The test results are processed by the Function Test Rules (a shallow reasoning component in NODAL) and a set of candidate faulty modules is produced. In the CBR implementation of NODAL these function test results make up the free features in the case description.

In order to further isolate the fault it is necessary to perform some internal measurements on the unit. These measurements are taken at internal nodes in the circuit; evidently these are expensive features. Measurements may involve estimating the goodness of a signal, or measuring voltages and resistances. A typical circuit will have about 20 nodes at the module level and approaching 100 nodes altogether. Consequently there is a large number of measurements that can be taken during the diagnosis. The advantage of the goal-driven diagnosis is that it requests only measurements that contribute to its current hypothesis. In a typical session only about 20% of the total possible measurements are actually requested. A typical dialog with the system is as follows:-

```
Setup for Test Vector 1

What is the SIGNAL of NODE-2?  Good
What is the SIGNAL of NODE-3?  Bad

It looks like the fault is in the LOCAL-POWER-SUPPLY
Switching to considering the circuit at a component level...

   What is the VOLTAGE of NODE-2?  23.4
   What is the VOLTAGE of LPS-1?  18.79
   What is the VOLTAGE of NODE-3?  18.12
   What is the VOLTAGE of NODE-9?  0

   It looks like the fault is in CR2
```

Even by NODAL's standards this dialogue is particularly short as the first module to be examined proves to be the faulty one (for more example dialogues see Cunningham, 1988). Nevertheless, it serves to illustrate how the goal-directed reasoning focuses the requesting of measurements from the operator. This clarifies the requirements for a CBR implementation of NODAL. Since there is a cost associated with determining the inputs to the diagnosis it is important that the CBR system should not need to have all the inputs in advance. Moreover, it should be able to direct the operator as to what measurements are important just as the goal-driven system does.

2.2 Incremental CBR in Classification

In order to evaluate the merit of our incremental CBR mechanism we implemented two simple classification systems using this approach. One is a system for credit risk assessment using the Japanese credit screening data taken from the UCI repository of machine learning databases. This system contains 125 cases. The other is a property valuation system containing 65 cases gathered in the Dublin area. The object of this system is to set an asking price for the target property. Sample cases from these systems are shown in Figure 2. The free features are shown in italics. In these classification problems the imperative for the distinction between free and expensive features is not a strong as it is in diagnosis situations. Nevertheless, if case retrieval can be performed without requiring all the case features there is a benefit.

Credit Assessment Case

Gender	*Female*
Purchase Item	*PC*
Jobless	No
Unmarried	Yes
Problematic Region	No
Age	18
Deposit	20
Monthly Payment	2
No. of Months	15
No. of Years in Company	1
Credit Screening	Pos

Property Valuation Case

Location	*loc4*
No-bed-rooms	*5-bed-rooms*
No-rec-rooms	
Kitchen	
Structure	detached
No-floors	2-floors
Condition	excellent-condition
Age	mature-age
Facilities	facilities-near
Price	88000

Figure 2. Sample cases for credit risk assessment and property valuation; free features are shown in italics.

3 Incremental CBR

An incremental CBR system solves a target problem in two stages. First, the free features of the target are used to rapidly identify a set of candidate case compatible with the target problem situation; the so called *base filtering stage*. The following stage is responsible for selecting, from the candidate cases, the best case to retrieve. To do this additional features of the target situation must be gathered in order to identify the actual candidate case to use. This involves choosing a set of features from the candidate cases and obtaining their values in the target situation. Since we are assuming that this data gathering process is expensive it is crucial that this set of features is minimal. The choice of these features is controlled by an information theoretic criterion that selects features according to their information content or *discriminatory power*. In this section an example incremental CBR system ($NODAL_{CBR}$) is briefly introduced and the critical feature selection procedure is described.

3.1 $Nodal_{CBR}$: An Incremental CBR System for Diagnosis

$NODAL_{CBR}$ was developed to investigate whether the desirable, goal-directed behaviour of the model-based diagnosis system, NODAL, could be transferred to an equivalent CBR system. The result is a CBR system that can begin operation without a complete target case description and can generate minimal sets of queries to help it to home in on a particular solution.

$NODAL_{CBR}$ solves a target problem in two stages. The first stage, base filtering, uses the function test results (free features) to select cases that exhibit similar test result characteristics. The result of base filtering is a reduced set of candidate cases, one of which should contain the desired solution (diagnosis). The second stage must then formulate a set of queries which call for the measurement of signals in the internal nodes of the circuit. The information theoretic criterion described in the next section forms the core of this data gathering process, ensuring that a minimal set of questions are formulated by choosing those features (internal circuit nodes) which maximally discriminate between the candidate cases. So, at each stage in the reduction of the set of candidates, the most discriminating feature is selected and the user is requested to determine its value in the target case. Cases that cannot match on this feature (cases which are known to exhibit a differnt signal value for the specified internal node) are removed from the retrieved set. This process is repeated until the set reduces to one diagnosis or the target case proves to be dissimilar to all the retrieved cases. It is important to emphasise that a full discrimination tree for the case-base is not being produced, instead local discriminations for the candidate cases are determined at run-time.

In fact, the $NODAL_{CBR}$ system performs better than the original model-based reasoning system because it only requires enough information to uniquely classify the target problem in the case-base. In comparison, NODAL requires enough information to *verify* a hypothesis in its knowledge base. Thus, strictly speaking, the CBR system requires a further validation phase where it informs the user of remaining information that will confirm that the cases match the target. However, the importance of this validation depends on the coverage of the case-base and for well populated case-bases validation can be safely ignored. When we compared the CBR system with the old system on a sample set of faults for particular DC/DC circuits we found that it reduced the number of question needed to form a diagnosis from between 33% to 85%; that is performance gains from 67% to 15%. So, from a situation where our initial aspiration was to produce a CBR system that would have the informational parsimony of a goal-driven system we find that the CBR system is actually *better* than the old NODAL system.

3.2 Discriminatory Power

The selection of a minimal set of features with which to discriminate against a set of cases amounts to building a decision tree with internal nodes corresponding to feature queries and leaf nodes corresponding to the different solution classes; the leaf nodes contain the cases that fall into these classes. A good rule of thumb in building an efficient decision tree is that features that provide the most information should appear high up in the tree (close to the root). Information theory provides us with just such means of measuring the information content or discriminatory power (DP) of case features with respect to a set of solution classes; this is similar to ID3 (Quinlan, 1986) and also related to the work of Wess, Altoff and Derwand, 1993.

The mechanics of selecting the most discriminatory feature with which to classify a set of cases can be easily understood by using $NODAL_{CBR}$ as an example. In $NODAL_{CBR}$ the selection of discriminating features amounts to building a decision tree that will have leaf nodes corresponding to the different diagnoses **D**, and the set of cases **C** will be located, or classified, on these nodes. The process used in $NODAL_{CBR}$ differs from that of ID3 in that the semantics of the branching in the decision tree is slightly different because of the large number of unknowns in the case features. A brief explanation of how the discrimination works is as follows:-

D={D_1,...,D_d} the set of possible classes or diagnoses

C={C_1,...,C_c} the set of cases to classify

F={F_1,...,F_f} the set of descriptive features that will form the nodes of the decision tree.

We can view the decision tree as an information source producing one of *d* messages from the set **D**. Let $|D_i|$ represent the number of cases with diagnosis D_i. Then the expected information needed to generate the appropriate message, for some case, using the tree is:-

$$I(|D_1|,\ldots,|D_d|) = -\sum_{i=1}^{d}\left(\frac{|D_i|}{|D_1|+\ldots+|D_d|}\bullet \log_2\left[\frac{|D_i|}{|D_1|+\ldots+|D_d|}\right]\right) \quad (1)$$

Consider the root decision node of the tree (see Figure 3). Assume this node tests the feature $F \in \mathbf{F}$ and this feature has possible values **V**={V_1,...,V_n}. Then **V** partitions **C** into n groups of cases, G_1,...,G_n; where G_i contains those cases that have value V_i for feature F.

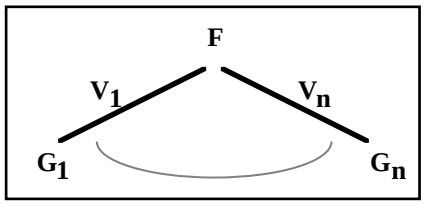

Figure 3. The root classification of the cases in **C**.

Let G_i contain $|D^i_j|$ cases with diagnosis D_j, that is $|D^i_j|$ instances of class D_j. Then the expected information required for the sub-tree of G_i is $I(|D^i_1|,...,|D^i_d|)$. We can obtain the expected information for the tree with F as root by computing the weighted average over all value branches of F as follows:-

$$E(F) = \sum_{i=1}^{n}\left(\frac{|D^i_1|+\ldots+|D^i_d|}{|D_1|+\ldots+|D_d|}\right)\bullet I\left(|D^i_1|,\ldots,|D^i_d|\right) \quad (2)$$

The weight of the *i*th branch is the proportion of cases in C that belong to G_i. The information gained from using F, or the discriminatory power of F, is:-

$$DP(F) = I(|D_1|,\ldots,|D_d|) - E(F) \quad (3)$$

So, at each step in the reduction of a set of candidate cases the user is asked to supply the value of that feature with the highest DP score; that is, the most discriminatory feature. Again, it is important to recognise that whereas the more common approach is to build a global decision tree over the entire training set of cases, in incremental CBR partial, local decision trees are constructed when needed over reduced sets of relevant cases.

4 Incremental CBR *vs.* Pure Inductive Learning: *An Experimental Comparison*

In this section we investigate the application of incremental CBR to more traditional inductive learning problems such as risk assessment and classification.We provide experimental evidence to support the hypothesis that, even for such tasks, incremental CBR does exhibit performance advantages over inductive learning. In particular, these advantages are readily manifest in domains where the accumulation of relevent data is costly and where some initial data is provided 'free' in the form of a set of free features[1]. The first set of experiments is designed to compare the performance characteristics of inductive learning to incremental CBR and demonstrate that a CBR approach benifits from a very definite performance gain, while the second set investigate the sensitivity of this gain to variations in the available free features.

4.1 Performance Experiments

These experiments compared the performance of a pure inductive learning approach and an incremental CBR approach over a set of problems from the property valuation and credit risk assessment domains. The performance measure used is the number of queries made in solving a particular target problem; our assumption being that performance is primarily influenced by the cost of gathering expensive data during the solution process.

To test the pure inductive learning approach a decision tree was built from the available training cases. Each case then served as a target problem and the total number of questions asked during their solution was accumulated; of course questions relating to the free features were not counted during these tests. Similarily, to test the incremental CBR approach each available case served as a target problem. For each run the free features were used during base filtering and a local, partial decision tree was constructed over the selected candidates. Again the number of questions asked during the traversal of these local trees was accumulated.

The results of these experiments can be seen in Figure 4(a & b). Figure 4(a) plots the results for a set of 52 target problems from the property valuation domain, and Figure 4(b) shows the results for a set of 125 problems from the credit risk assessment domain. In each graph the cumulative number of questions asked is plotted against the number of target problems

[1]For example, in the credit risk assessment domain the free features may be the purchase item whereas the costly unknown data may be such things as the employement status or social history of the applicant.

solved. It is clear from each that the incremental CBR approach provides a consistent reduction in the number of questions needed to solve the set of targets. In fact for each domain we can obtain an average value of the *performance gain* (the percentage reduction in questions asked) for a chosen index set; 11% for the property valuation tests and 38% for the credit risk assessment tests.

The incremental CBR gain is due to the fact that the constructing a local decision tree specific to some target problem tends to result in a considerably more efficient (compact) tree that if a global decision tree was constructed overall the entire training set; not only are the free features absent from this more compact tree but also other features which were needed in the global tree may also be unneccessary.

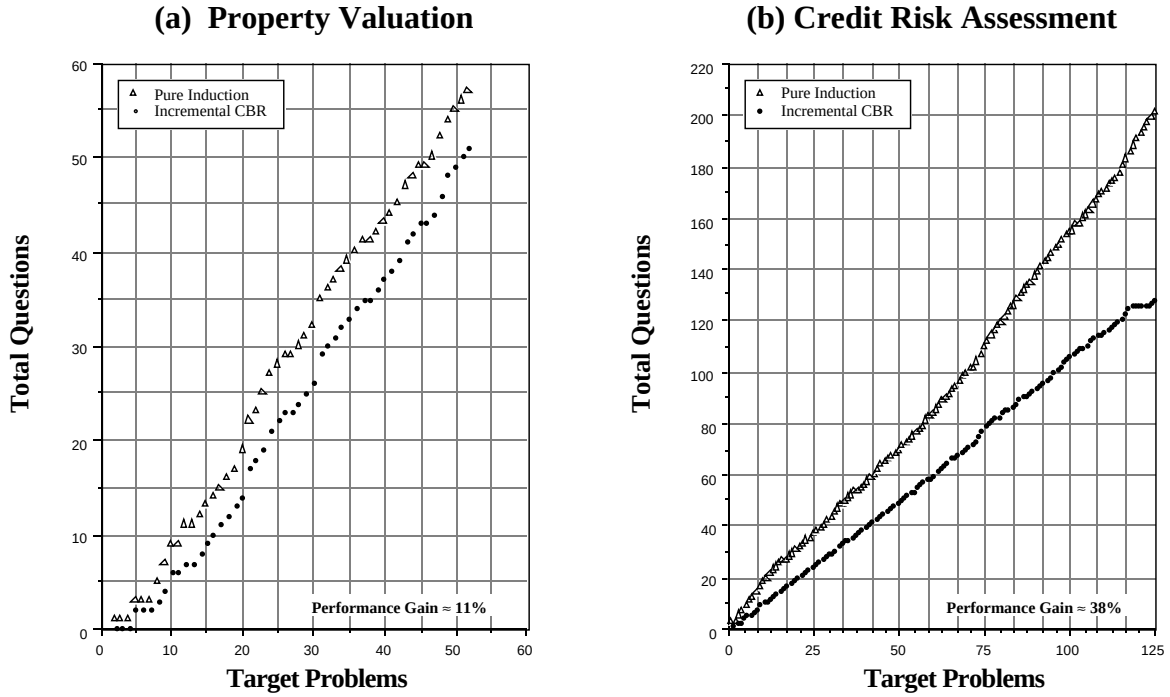

Figure 4. Performance Experiments
(a) Free Features = {Location, Num-Bedrooms}, (b) Free Features = {Purchase, Sex}

4.2 Sensitivity Experiments

In this set of experiments we set about investigating the relationship between the performance gain and the available free features. Modeling the precise nature of this relationship is beyond the scope of this paper but we can at least gain an impression of the range of potential performance gains for different sets of free features in different types of problem domain.

Again the property valuation and credit risk assessment domains were investigated. For each domain the experiments of the previous section were re-run 25 times, each time with a different set of nominated free features[2], and each time the average performance gain was noted.

Figure 5(a) and 5(b) show the results for the property valuation and credit risk assessment experiments respectively, plotting the performance gain against the different sets of free features. The property valuation domain exhibits a fairly stable performance gain over the different free features ranging from 26% to 10% and averaging at about 17%. In comparison, the gain in credit risk assessment domain is somewhat more sensitive to the choice of free feature, ranging from only 1% to 52% and averaging at about 23%. Clearly, the available features are important; the best case scenario for incremental CBR can result in substantial performance gains over pure induction methods (witness the 52% maximum gain in the credit risk assessment domain), but even the worst case situation above resulted in marginal improvements.

Basically, the discriminatory power of the free features is important in the resulting performance gain. In particular, free features with very low DP values result in very low gains as the impact of base-filtering is minimal and the resulting local decision tree is essentially the same as the global one produced by a pure inductive learning approach. At the other extreme, free features with high DP values tend to result in much greater performance gains.

[2]For these experiments each free feature set comprised of two randomly selected features from the domain feature list.

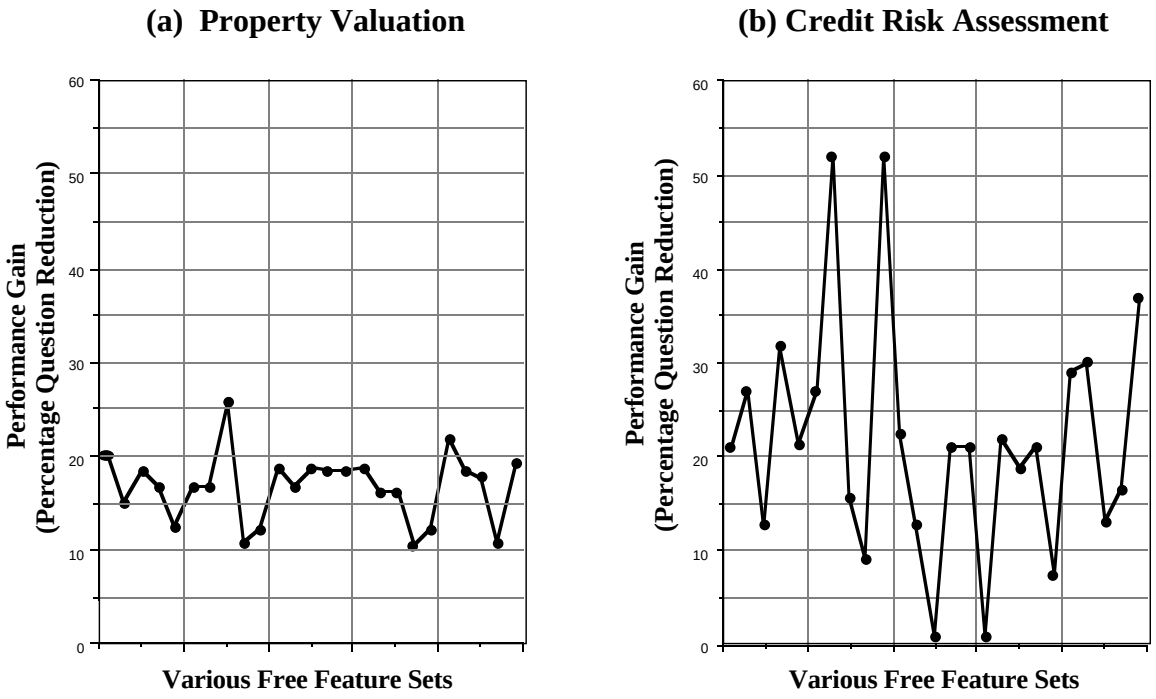

Figure 5. Performance Gain Sensitivity Experiments

5 Conclusion

We have described a technique, inceremental CBR, that borrows from inductive learning research to solve a problem in case-based reasoning. In addition, we have common class of problem tasks that can benifit from an incremental CBR approach instead of a more traditional inductive learning approach; the class is characterised by the fact that many relevant problem features are expensive to procure with the exception of a small set of initial features which are freely available. The key components of the incremental CBR approach are: firstly, the identification of a subset of relevant candidate cases by using the free features to eliminate many incompatible cases; secondly the identification of a minimal set of features from the remaining cases which, when instantiated in the target problem situation, will lead to the appropriate candidate case with which to solve the target problem.

While the immediate result is a CBR technique that benifits from the informational parsimony of goal-directed questioning, the main thrust of this paper was to investigate whether or not incremental CBR would actually exhibit performance improvements over pure inductive learning across a number of problem domains that are traditionally associated with inductive learning methods. In our experiments, carried out over two such domains, incremental CBR consistently outperformed traditional induction techniques, exhibiting performance gains (depending on the nature of the free features) from as little as 1% to as much as 52%, and averaging at about 20%. In conclusion then, we submit that incremental CBR represents a model of case-based reasoning which is particularly useful for tasks which involve sparse target problem descriptions and where the accumulation of additional features is expensive.

References

Cunningham P., Brady M. (1987) "Qualitative reasoning in electronic fault diagnosis", in Proceedings of Tenth International Joint Conference on Artificial Intelligence IJCAI-'87, ed. J. McDermott, Morgan Kaufmann, Milan Italy, pp443-445.

Cunningham P. (1988) Knowledge Representation in Electronic Fault Diagnosis, Ph. D. Thesis, Department of Computer Science, Dublin University, Trinity College, Ireland.

Cunningham P., Smyth B. (1994) "A comparison of model-based and incremental case-based approaches to electronic fault diagnosis", submitted to the Case-Based Reasoning Workshop, AAAI-1994.

Michalski R.S. (1983) "A theory and methodology of inductive learning", in Machine Learning An Artificial Intelligence Approach, R.S. Michalski, J.G. Carbonnell, T.M. Mitchell eds., Morgan Kaufmann.

Quinlan J.R. (1986) "Induction of Decision Trees", Machine Learning, Vol. 1, No. 1, pp81-106.

Smyth B., Cunningham P. (1992) "Déjà Vu: A Hierarchical Case-Based Reasoning System for Software Design", in Proceedings of European Conference on Artificial Intelligence, ed. Bernd Neumann, John Wiley, pp587-589, Vienna Austria.

Wess S., Althoff K-D., Derwand G. (1993) "Improving the Retrieval Step in Case-Based Reasoning", in Proceedings of the First European Workshop on Case-Based Reasoning, pp83-88, Germany.