Explaining the Pros and Cons of Conclusions in CBR

David McSherry

School of Computing and Information Engineering,
University of Ulster, Coleraine BT52 1SA, Northern Ireland
dmg.mcsherry@ulster.ac.uk

Abstract. We begin by examining the limitations of *precedent-based* explanations of the predicted outcome in case-based reasoning (CBR) approaches to classification and diagnosis. By failing to distinguish between features that support and oppose the predicted outcome, we argue, such explanations are not only less informative than might be expected, but also potentially misleading. To address this issue, we present an *evidential* approach to explanation in which a key role is played by techniques for the discovery of features that support or oppose the predicted outcome. Often in assessing the evidence provided by a continuous attribute, the problem is where to "draw the line" between values that support and oppose the predicted outcome. Our approach to the selection of such an *evidence threshold* is based on the *weights of evidence* provided by values above and below the threshold. Examples used to illustrate our evidential approach to explanation include a prototype CBR system for predicting whether or not a person is over the legal blood alcohol limit for driving based on attributes such as units of alcohol consumed.

1 Introduction

It is widely recognised that users are more likely to accept intelligent systems if they can see for themselves the arguments or reasoning steps on which their conclusions are based. In many problem-solving situations, the solution is not clear-cut, and it is reasonable for users to expect an intelligent system to explain the *pros* and *cons* of a suggested course of action [1-3]. In domains such as fault diagnosis, it is also reasonable for users to expect the system to explain the relevance of test results they are asked to provide, for example in the case of tests that carry high risk or cost [4-6]. Explanation is also a topic of increasing importance in areas such as intelligent tutoring and product recommendation [7-8]. However, we confine our attention in this paper to explanation of conclusions in CBR systems for classification and diagnosis.

While rule-based approaches to explanation remain an important legacy from expert systems research, a view shared by many CBR researchers is that explanations based on previous experience may be more convincing than explanations based on rules [9-10]. Recent research by Cunningham *et al.* [9] provides empirical evidence to support this hypothesis. In experiments involving human subjects, simply showing the user the most similar case in a classification task was found to be a more convincing explanation of the predicted outcome than a rule-based explanation generated from a decision tree. But what does it *mean* for an explanation to be convincing? In the case of a decision that is not clear-cut, trying to convince the user that the predicted outcome is *correct* does not make sense. Instead, the challenge is to convince the user that the predicted outcome is *justified* in spite of the evidence that opposes it. However, we argue that failure to distinguish between positive and negative evidence

P. Funk and P.A. González Calero (Eds.): ECCBR 2004, LNAI 3155, pp. 317–330, 2004.

limits the usefulness of precedent-based explanations as a basis for showing how a predicted outcome is justified by the available evidence.

We do not suggest that there is no value in showing the user the case on which the predicted outcome is based. An obvious advantage is that the user can assess for herself how closely it matches the problem description. But attempting to justify the predicted outcome simply by showing the user the most similar case ignores the possibility that some of the features it shares with the target problem may actually *oppose* rather than support the predicted outcome [2]. The result is that the explanation is not only less informative than might be expected, but also potentially misleading. In the absence of guidance to the contrary, any feature that the target problem shares with the most similar case may be interpreted by the user as evidence in favour of the predicted outcome even if it has the opposite effect.

Unfortunately, the chances of precedent-based explanations being open to misinterpretation in this way are far from remote. Given that similarity measures reward matching features whether or not they support the predicted outcome, it is not unlikely that one or more of the features that the most similar case has in common with the target problem actually provide evidence against the predicted outcome. It is worth noting that the problems we have identified are not specific to precedent-based explanations. In fact, rule-based explanations also fail to distinguish between positive and negative evidence, and as we show in Section 2, can also be misleading.

To address the need for more informative explanations, we present an *evidential* approach to explanation in which the user is shown the evidence, if any, that opposes the predicted outcome as well as the evidence that supports it. As in previous work [2], a key role in our approach is played by techniques for the discovery of features that support and oppose the predicted outcome. However, our initial approach was limited to assessing the evidence provided by nominal or discrete attributes. Here we present new techniques for explaining the pros and cons of a predicted outcome in terms of the evidence provided by continuous attributes, a requirement we consider essential to provide a realistic basis for explanation in CBR. Often in the case of a continuous attribute, the problem is where to "draw the line" between values that support and oppose the predicted outcome. Our approach to the selection of such an *evidence threshold*, which currently focuses on binary classification tasks, is based on the *weights of evidence* provided by values above and below the threshold.

In Section 2, we examine the limitations of approaches to explanation in which the user is simply shown the case or rule on which a conclusion is based. In Section 3, we describe the techniques for discovery of features that support and oppose a predicted outcome used in our evidential approach to explanation, and our approach to assessing the evidence provided by continuous attributes. In Section 4, an example case library based on Cunningham *et al.*'s [9] breathalyser dataset is used to illustrate our evidential approach to explanation as implemented in a prototype CBR system called *ProCon-2*. Related work is discussed in Section 5 and our conclusions are presented in Section 6.

2 Limitations of Existing Approaches

In this section, we examine more closely the limitations of precedent-based and rule-based explanations that motivate our evidential approach to explaining the pros and cons of the conclusions reached by a CBR system.

2.1 Rule-Based Explanations

In problem-solving based on decision trees, the standard approach to explaining how a conclusion was reached is to show the user all features on the path from the root node to the leaf node at which the conclusion was reached [5,9]. A similar approach is possible in CBR systems in which a decision tree is used to guide the retrieval process. Because any solution path in a decision tree can be regarded as a rule, the resulting explanation is often referred to as a *rule-based* explanation. In fact, explaining conclusions in this way is very similar to the standard expert systems technique of showing the user the rule on which the conclusion is based [6].

However, one of the problems associated with rule-based explanations is that some of the evidence that the user is shown may not support the conclusion [5]. Even worse, it is possible that some of the evidence presented actually *opposes* the conclusion. The example we use to illustrate this problem is based on Cendrowska's contact lenses dataset [11]. This well-known dataset is based on a simplified version of the optician's real-world problem of selecting a suitable type of contact lenses (none, soft, or hard) for an adult spectacle wearer.

Fig. 1 shows part of a decision tree induced from the contact lenses dataset with Quinlan's information gain measure [12] as the splitting criterion. The following explanation for a conclusion of no contact lenses was generated from the contact lenses decision tree.

if tear production rate = normal
and astigmatism = absent
and age = presbyopic
and spectacle prescription = myope
then conclusion = no contact lenses

However, it is clear from the contact lenses decision tree that a normal tear production rate cannot be regarded as evidence in favour of no contact lenses, since a reduced tear production rate is enough evidence on its own to reach the same conclusion. So the first condition in the explanation that the user is shown is not only redundant but also potentially misleading.

Another problem associated with decision trees is that the user may be asked for the results of tests that are not strictly necessary to reach a conclusion [11]. In recent work, we presented a *mixed-initiative* approach to classification based on decision trees in which the system does not insist on asking the questions and is capable of eliminating redundant conditions from the explanations it generates [5]. However, even if a rule-based explanation contains no redundant or opposing conditions, it remains open to the criticism of presenting only positive evidence in favour of the conclusion. The user has no way of telling whether evidence that is not mentioned in the explanation has a positive or negative effect on the conclusion.

2.2 Precedent-Based Explanations

Typically in CBR approaches to classification and diagnosis, the predicted outcome is explained by showing the user the case that is most similar to the target problem [9,13,14]. The example we use to illustrate the limitations of precedent-based explanations is based on Cunningham *et al.*'s [9] breathalyser dataset for predicting whether or not a person is over the legal blood alcohol limit for driving in Ireland.

Attributes in the dataset are the weight and sex of the subject, duration of drinking in minutes, meal consumed, and units of alcohol consumed. The outcomes to be predicted in this binary classification task are over-limit and not-over-limit.

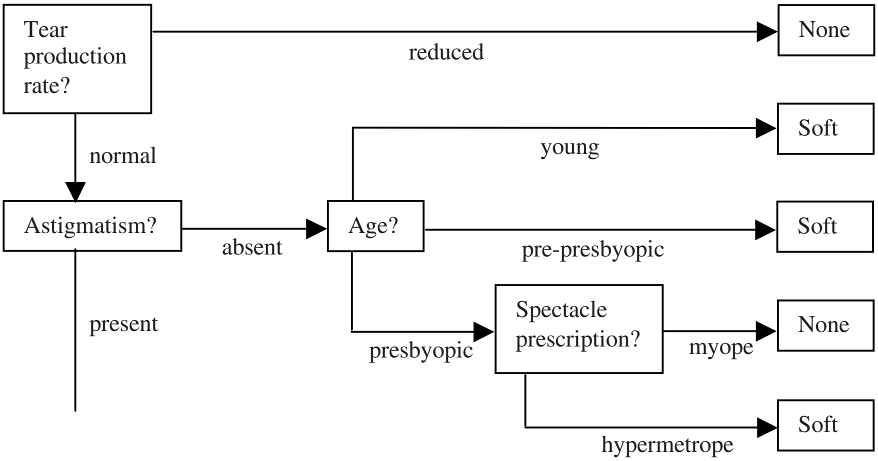

Fig. 1. Partial decision tree based on the contact lenses dataset

Target problem:	*Most similar case*:
weight = 79	weight = 79
duration = 90	duration = 240
sex = male	sex = male
meal = full	meal = full
units =10.1	units = 9.6
predicted outcome: over-limit	outcome: over-limit

Fig. 2. A precedent-based explanation of the predicted outcome for a target problem in the breathalyser domain

A target problem and the outcome predicted by a CBR system based on the breathalyser dataset are shown in Fig. 2. The most similar case on which the predicted outcome is based in our system is also shown. The target problem and most similar case are the same as those used in [9] to illustrate a typical precedent-based explanation in the breathalyser domain.

Given that the most similar case exactly matches the target problem on three of its five features, it is not surprising that its similarity to the target problem is very high (0.97). But this is *not* the same as saying that there is strong evidence in favour of the predicted outcome. In fact, as we show in Section 3, none of the three features that the most similar case shares with the target problem supports the predicted outcome.

So how useful is showing the user the most similar case likely to be as an explanation of the predicted outcome in this example? One problem is that the user could be forgiven for thinking that the matching features sex = male, meal = full, and weight =

79 are evidence in favour of the subject being over the limit when in fact they *oppose* the predicted outcome. The explanation also fails to comment on the much shorter duration of drinking in the target problem or how this should affect our confidence in the predicted outcome.

An important message that the explanation fails to convey is that the predicted outcome, though perhaps justifiable on the basis of the number of units consumed, is far from being a clear-cut decision. As described in the following sections, showing the user the evidence that opposes the predicted outcome as well as the evidence that supports it is one of the ways in which we propose to provide more informative explanations in CBR approaches to classification and diagnosis.

3 The Pros and Cons of a Predicted Outcome

In classification and diagnosis, it is seldom the case that all the reported evidence supports the conclusion reached by an intelligent system, or indeed by a human expert. More typically in practice, some features of the target problem will provide evidence in favour of the conclusion while others provide evidence against it. In a CBR system, it is equally unlikely that all the features in the most similar case support the outcome recorded for that case. Our evidential approach to explanation in CBR aims to address the limitations of precedent-based explanations highlighted in Section 2 by showing the user the evidence, if any, that opposes the predicted outcome as well as the supporting evidence.

3.1 Criteria for Support and Opposition

An important point to be considered in assessing the evidence for and against a predicted outcome is that certain features (or test results) may sometimes increase and sometimes decrease the probability of a given outcome class, depending on the evidence provided by other features [15,16]. However, it has been shown to follow from the independence (or Naïve) version of Bayes' theorem that a given feature always increases the probability of an outcome class if it is more likely in that outcome class than in any competing outcome class [15]. We will refer to such a feature as a *supporter* of the outcome class. Conversely, a feature always decreases the probability of an outcome class if it is less likely in that outcome class than in any competing outcome class. We will refer to such a feature as an *opposer* of the outcome class. A feature that is neither a supporter nor an opposer of an outcome class may sometimes provide evidence in favour of the outcome class, and sometimes provide evidence against it.

Below we define the criteria for support and opposition of a given outcome class on which our evidential approach to explanation is based.

The Support Criterion. *A feature E is a supporter of an outcome class H_1 if there is at least one competing outcome class H_2 such that $p(E \mid H_1) > p(E \mid H_2)$ but no competing outcome class H_2 such that $p(E \mid H_1) < p(E \mid H_2)$.*

The Opposition Criterion. *A feature E is an opposer of an outcome class H_1 if there is at least one competing outcome class H_2 such that $p(E \mid H_1) < p(E \mid H_2)$ but no competing outcome class H_2 such that $p(E \mid H_1) > p(E \mid H_2)$.*

In our approach to explaining the pros and cons of a predicted outcome, a feature may be the observed value of an attribute, such as age = young in the contact lenses dataset, or a condition defined in terms of a continuous attribute, such as units ≥ 6 in the breathalyser dataset. In practice, though, the reliability of the evidence provided by a given feature is likely to depend on whether it occurs with sufficient frequency in the dataset for its conditional probability in each outcome class to be estimated with reasonable precision.

Our criteria for support and opposition of an outcome class can be expressed in simpler terms when applied to a binary classification task.

Proposition 1. *In a classification task with two possible outcomes H_1 and H_2, a given feature E is a supporter of H_1 if $p(E \mid H_1) > p(E \mid H_2)$ and an opposer of H_1 if $p(E \mid H_1) < p(E \mid H_2)$.*

It can also be seen from Proposition 1 that in a classification task with only two possible outcomes, a given feature must either support or oppose a given outcome class except in the unlikely event that its conditional probability is the same in both outcome classes. As we shall see, however, this is not the case in datasets in which there are more than two outcome classes.

3.2 Nominal and Discrete Attributes

Values of a nominal or discrete attribute that support and oppose the outcome classes in a given case library can easily be identified from the conditional probabilities of the attribute's values. Table 1 shows the conditional probabilities for two of the attributes, age and tear production rate, in the contact lenses dataset. For example, it can be seen that:

$$p(\text{age} = \text{young} \mid \text{none}) = 0.27$$
$$p(\text{age} = \text{young} \mid \text{soft}) = 0.40$$
$$p(\text{age} = \text{young} \mid \text{hard}) = 0.50$$

So according to our criteria for support and opposition, age = young is a *supporter* of hard contact lenses and an *opposer* of no contact lenses. On the other hand, age = young is neither a supporter nor an opposer of soft contact lenses.

Table 1. Conditional probabilities for two of the attributes in the contact lenses dataset

Type of contact lenses:	None	Soft	Hard
Age:			
young	0.27	0.40	0.50
pre-presbyopic	0.33	0.40	0.25
presbyopic	0.40	0.20	0.25
Tear production rate:			
normal	0.20	1.00	1.00
reduced	0.80	0.00	0.00

It can also be seen from Table 1 that tear production rate = normal is an *opposer* of no contact lenses, a finding that is consistent with our impression from the contact lenses decision tree in Fig. 1.

Table 2 shows the conditional probabilities for sex and meal consumed in the breathalyser dataset [9]. According to our criteria for support and opposition, sex = female is a supporter of over-limit as it is more likely in this outcome class than in the only competing outcome class. Interestingly, meal = full is a supporter of not-over-limit while meal = lunch is a supporter of over-limit. The estimated conditional probabilities for meal = snack and meal = none should perhaps be regarded more cautiously as neither of these values is well represented in the dataset.

Table 2. Conditional probabilities for two of the attributes in the breathalyser dataset

		over-limit	not-over-limit
Sex:	female	0.23	0.19
	male	0.77	0.81
Meal:	full	0.33	0.51
	lunch	0.43	0.19
	snack	0.10	0.14
	none	0.13	0.16

3.3 Explanation in ProCon

In previous work we presented a CBR system for classification and diagnosis called ProCon that can explain the pros and cons of a predicted outcome in terms of the evidence provided by nominal or discrete attributes [2]. As often in practice, the predicted outcome for a target problem in ProCon is the outcome associated with the most similar case. An example case library based on the contact lenses dataset was used in [2] to illustrate ProCon's ability to construct a structured explanation of a predicted outcome in which the user is shown any evidence that opposes the conclusion as well as the evidence that supports it. When explaining a predicted outcome of no contact lenses, for example, ProCon recognises a normal tear production rate, if reported by the user, as evidence *against* the predicted outcome.

In the case of a feature that is neither a supporter nor an opposer of the predicted outcome, ProCon *abstains* from commenting on the impact of this feature in its explanation of the predicted outcome. When explaining a predicted outcome of soft contact lenses, for example, ProCon would abstain from commenting on age = young while recognising age = pre-presbyopic as evidence in favour of the predicted outcome and age = presbyopic as evidence against it.

In Section 4 we present a new version of ProCon called ProCon-2 that can also explain the pros and cons of a predicted outcome in terms of the evidence provided by continuous attributes such as units of alcohol in the breathalyser dataset.

3.4 Continuous Attributes

Focusing now on classification tasks in which there are only two possible outcomes, we present the techniques used in ProCon-2 to explain the pros and cons of a predicted outcome in terms of the evidence provided by continuous attributes. Often in assessing the evidence provided by a continuous attribute, the problem is where to "draw the line" between values that support the predicted outcome and values that oppose it. Choosing a realistic evidence threshold for a continuous attribute can be more difficult than it might seem at first sight. In the case of units of alcohol in the breathalyser dataset, the problem is that for any value x of units apart from the minimum value in the dataset:

$$p(\text{units} \geq x \mid \text{over-limit}) > p(\text{units} \geq x \mid \text{not-over-limit})$$

Thus according to our criteria for opposition and support, units ≥ x is a supporter of over-limit for *any* value x of units apart from the minimum value in the dataset. If we choose a high value of units such as 15 as the evidence threshold, it is intuitive that values above the threshold will provide strong evidence in favour of over-limit. However, it is equally intuitive that values below the threshold will provide little evidence *against* over-limit. A system that attempts to justify a conclusion that the subject is not over the limit on the basis that units < 15 is unlikely to inspire user confidence in its explanation capabilities.

Similarly, if we choose a low value of units as the evidence threshold, then values below the threshold will provide strong evidence in favour of not-over-limit, whereas values above the threshold will provide only weak evidence in favour of over-limit. Our solution to this dilemma is based on the concept of weights of evidence [4,15,17].

Definition 1. *If H_1 and H_2 are the possible outcomes in a binary classification task, then for any feature E such that* $0 < p(E \mid H_2) \leq p(E \mid H_1) \leq 1$, *we define the weight of evidence of E in favour of H_1 to be*:

$$we(E, H_1) = \frac{p(E \mid H_1)}{p(E \mid H_2)}$$

For example, a weight of evidence of two in favour of H_1 means that E is twice as likely in H_1 as it is in H_2. The usefulness of weight of evidence as a measure of the impact of reported evidence on the probabilities of the competing outcome classes can be seen from the following proposition, which follows easily from the independence (or Naïve) form of Bayes' theorem [15].

Proposition 2. *If H_1 and H_2 are the possible outcomes in a binary classification task, then for any feature E such that* $0 < p(E \mid H_2) \leq p(E \mid H_1) \leq 1$,

$$\frac{p(H_1 \mid E)}{p(H_2 \mid E)} = we(E, H_1) \times \frac{p(H_1)}{p(H_2)}$$

Our intuitions regarding the trade-off associated with 15 as an evidence threshold for units consumed are borne out by the following calculations based on the breathalyser dataset [9].

$p(\text{units} \geq 15 \mid \text{over-limit}) = 0.33$ $\quad p(\text{units} < 15 \mid \text{over-limit}) = 0.67$

$p(\text{units} \geq 15 \mid \text{not-over-limit}) = 0.07$ $\quad p(\text{units} < 15 \mid \text{not-over-limit}) = 0.93$

$$we(\text{units} \geq 15, \text{over-limit}) = \frac{0.33}{0.07} = 4.7 \qquad we(\text{units} < 15, \text{not-over-limit}) = \frac{0.93}{0.67} = 1.4$$

As might be expected, *we*(units ≥ 15, over-limit) is relatively high whereas *we*(units < 15, not-over-limit) is close to its minimum possible value.

In the case of units, our approach to the selection of a realistic evidence threshold is to select the threshold x that maximises the minimum of the weights of evidence that units $\geq x$ provides in favour of over-limit and units $< x$ provides in favour of not-over-limit. That is, we choose the value x of units that maximises:

$$\text{MIN}(we(\text{units} \geq x, \text{over-limit}), we(\text{units} < x, \text{not-over-limit}))$$

In general for a continuous attribute, the evidence threshold selected in our approach is the value that maximises the minimum of the weights of evidence provided by values above and below the threshold.

3.5 Experimental Results

Fig. 3 shows the results of an empirical evaluation of possible evidence thresholds for units consumed in the breathalyser dataset [9]. We confine our attention here to values x of units for which we(units $\geq x$, over-limit) and we(units $< x$, not-over-limit) are both defined. For example, we exclude units $\geq$ 17 because p(units $\geq$ 17 | not-over-limit) = 0.

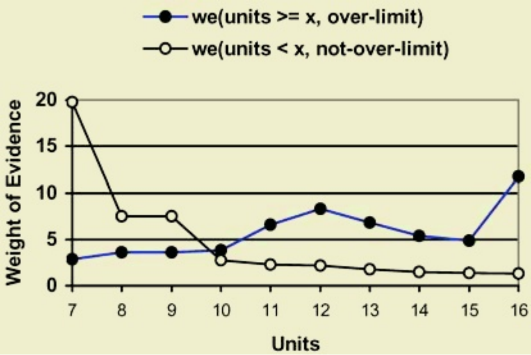

Fig. 3. Weights of evidence for units of alcohol consumed in the breathalyser dataset

Two of the possible thresholds, 8 and 9, can be seen to maximise the minimum of the weights of evidence in favour of over-limit and not-over-limit. For example, we(units $\geq$ 8, over-limit) = 3.6 and we(units < 8, not-over-limit) = 7.5. The fact that the weights of evidence are the same for units = 8 and units = 9 can be explained by the fact that there are no cases in the dataset with values of units between 8 and 8.9. On this basis, a reasonable choice of evidence threshold would be 8 or 9 (though of course it is possible for a person who has consumed much less than 8 units to be over the limit). In practice, the selected evidence threshold is based in our approach on actual values that occur in the dataset rather than equally spaced values as in this analysis. The threshold selected by ProCon-2, based on a minimum weight of evidence of 4.0 in favour of over-limit, is 9.1.

Fig. 4 shows the results of a similar evaluation of evidence thresholds for duration of drinking in the breathalyser dataset. In this case the choice is more clear-cut. The evidence threshold that maximises the minimum weight of evidence is duration $\geq$ 150. The evidence threshold selected by ProCon-2 from the actual values that occur in the dataset is also duration $\geq$ 150.

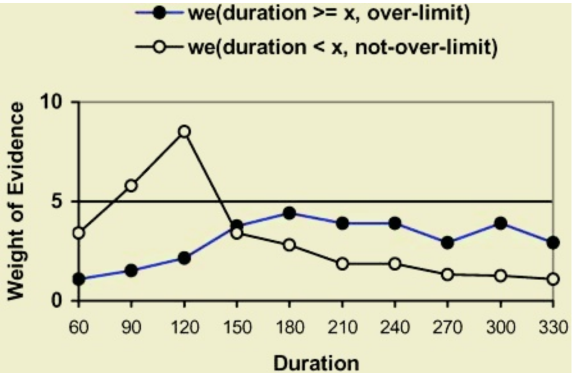

Fig. 4. Weights of evidence for duration of drinking in the breathalyser dataset

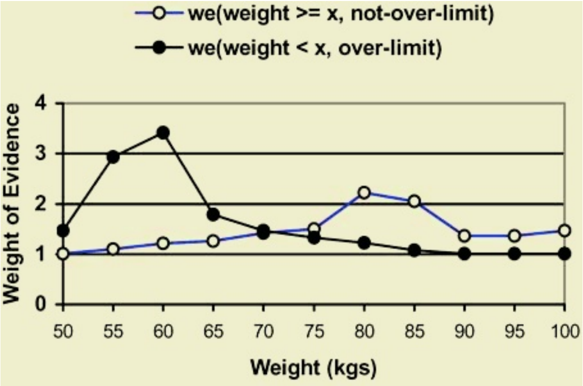

Fig. 5. Weights of evidence for weight in kgs in the breathalyser dataset

Finally, Fig. 5 shows the results of an evaluation of evidence thresholds for weight in the breathalyser dataset. This attribute differs from units and duration in that values *below* a given threshold tend to support the conclusion that the subject is over the limit. The evidence threshold that maximises the minimum weight of evidence in this case is weight ≥ 70. The threshold selected by ProCon-2 from the actual values that occur in the dataset is weight ≥ 73.

Table 3. Evidence thresholds in the breathalyser dataset selected by two different strategies and their minimum weights of evidence (WE)

	Units (WE)	Duration (WE)	Weight (WE)
Mid-Point	16.2 (1.3)	220 (1.9)	74 (1.3)
Max-Min	9.1 (4.0)	150 (3.4)	73 (1.5)

In Table 3 we summarise the results of an empirical comparison of our "Max-Min" strategy of maximising the minimum weight of evidence with an alternative "Mid-

Point" strategy of selecting the mid-point of the attribute's range of values in the dataset as the evidence threshold. For each continuous attribute in the breathalyser dataset, the evidence threshold selected in each strategy is shown together with the minimum of the weights of evidence provided by values above and below the threshold. While the evidence thresholds selected in the two strategies differ only slightly in the case of weight, there are marked differences in the evidence thresholds selected for units and duration. In the case of units, the evidence threshold selected by our Max-Min strategy can be seen to *treble* the minimum weight of evidence provided by the mid-point value. In the case of duration, the minimum weight of evidence is increased by nearly 80% from 1.9 to 3.4.

4 Explanation in ProCon-2

We now present an implementation of our evidential approach to explanation in ProCon-2, a CBR system for classification and diagnosis in which the predicted outcome is based on nearest-neighbour retrieval. Initially on entering the description of a target problem, the user is shown the *three* most similar cases and the predicted outcome from the most similar case. If the outcome for the three most similar cases is the same, this may help to increase the user's confidence in the predicted outcome. On requesting an explanation of the predicted outcome, the user is also shown:

- Features in the target problem that *support* and *oppose* the predicted outcome
- Features in the most similar case, if any, that *oppose* the predicted outcome

Showing the user features in the most similar case that oppose the predicted outcome may help to convince the user that the predicted outcome is *justified* if one or more of the opposing features are also present in the target problem. Thus the fact that the predicted outcome was the correct outcome for the most similar case, in spite of any opposing evidence it shares with the target problem, is used in our approach as additional justification of the predicted outcome.

The example dialogue in Fig. 6 is based on the breathalyser dataset [9]. Attributes in the case library and weights assigned to them in ProCon-2 are weight (3), duration (2), sex (2), meal (3) and units (4). The similarity of each case to the target problem is shown in brackets.

The example dialogue shows how ProCon-2 attempts to justify the predicted outcome in terms of the features that support and oppose the predicted outcome. When the user asks for an explanation, she is shown one feature in the target problem that supports the predicted outcome and three features that oppose the predicted outcome. ProCon-2 also shows the user the most similar case and points out that the outcome in that case was over-limit in spite of three opposing features that it shares with the target problem. However, it should be obvious to the user from the evidence presented that the decision is not clear-cut. The fact that the outcome in the second and third most similar cases differs from the predicted outcome may also help to convey this important point.

Discovery of the supporters and opposers of the predicted outcome is *demand driven* in ProCon-2. The selection of evidence thresholds for continuous attributes is also based on the available cases at problem-solving time. In the example dialogue, the user is not shown the evidence thresholds used by ProCon-2 to assess the evidence provided by continuous attributes. The possible benefit of providing this additional information is one of the issues to be addressed by further research.

ProCon-2: Please describe the target problem.

User: weight = 79, duration = 90, sex = male, meal = full, units =10.1

ProCon-2: The 3 most similar cases are:

Case 13: 79, 240, male, full, 9.6, over-limit (0.97)
Case 26: 73, 120, male, full, 9, not-over-limit (0.96)
Case 56: 79, 120, male, full, 7.2, not-over-limit (0.96)

The predicted outcome is: over-limit.

User: **why**

ProCon-2: Features of the target problem that SUPPORT the prediction:

units = 10.1

Features of the target problem that OPPOSE the prediction:

weight = 79
duration = 90
sex = male
meal = full

The most similar case is:

Case 13: 79, 240, male, full, 9.6, over-limit

The outcome in this case was over-limit in spite of the opposing evidence that:

weight = 79
sex = male
meal = full

Fig. 6. Example ProCon-2 dialogue based on the breathalyser dataset

5 Related Work

Existing approaches to explaining the pros and cons of a suggested action or hypothesis in CBR tend to be more knowledge intensive than our approach. Murdock *et al.* [3] describe an interpretive CBR approach to assisting intelligence analysts in the evaluation of hypothesised asymmetric threats such as an attempt by an organised crime group to take over a commercial industry. Given such an hypothesis, their system generates a structured summary of the arguments for and against the hypothesis by comparing the available evidence with the retrieved model that most closely matches the hypothesised activity. Currently, the decision as to whether or not the hypothesis is valid given the evidence presented is left to the user.

Brüninghaus and Ashley [1] describe an approach to predicting the outcome of legal cases in which the predicted outcome is accompanied by an annotated, *issue-based* analysis of factors that favour the defendant and factors that favour the plaintiff. Relying on a weak domain model to identify issues raised in a case, the system also uses cases to reason about conflicting evidence related to each issue.

An interesting example of an approach to explanation in which precedent-based and rule-based explanations play complementary roles is Evans-Romaine and Marling's [13] prototype system for teaching students in sports medicine to prescribe exercise regimes for patients with cardiac or pulmonary diseases. On entering the description of a patient in terms of attributes such as age, sex, weight and diagnosis, students are shown both a recommendation based on rules and a possibly conflicting solution based on CBR. The former solution is supported by a rule-based explanation, and the latter by showing the student the most similar case. In this way, it is argued, students learn not only to apply the standard rules but also how experienced prescribers look beyond the rules to the needs of individual patients.

6 Conclusions

Our evidential approach to explaining the pros and cons of conclusions in CBR aims to address the limitations of approaches to explanation in which the user is simply shown the case (or rule) on which a predicted outcome is based. An important role in our approach is played by techniques for the discovery of features that support or oppose the outcome predicted by the system. We have also presented a principled approach to the selection of evidence thresholds for assessing the evidence provided by continuous attributes. Initial results suggest that our strategy of maximising the minimum of the weights of evidence provided by values above and below the threshold produces more realistic evidence thresholds than simply selecting the mid-point of the attribute's range of values in the dataset. Finally, we have presented an implementation of our evidential approach to explanation in a CBR system called ProCon-2 and demonstrated its ability to provide explanations that are more informative than is possible by simply showing the user the most similar case.

Currently our approach is best suited to binary classification tasks. One reason is that our techniques for assessing the evidence provided by continuous attributes are currently limited to binary classification tasks. Another is that in a dataset with several outcome classes, some of the features in the target problem may be neither supporters nor opposers of the predicted outcome according to our criteria for support and opposition. A possible approach to addressing both issues that we propose to investigate in future research is to treat the problem as a binary classification task for the purpose of explaining the predicted outcome.

An interesting question is whether it is possible to explain the pros and cons of the conclusions reached by a CBR system without relying on probabilistic criteria for support and opposition. We are currently investigating an approach to explanation that closely resembles our evidential approach but in which the criteria for support and opposition of an outcome class are defined in terms of the underlying similarity measure on which retrieval is based rather than in probabilistic terms.

Acknowledgements. The author would like to thank Pádraig Cunningham, Dónal Doyle and John Loughrey for providing the breathalyser dataset used here to illustrate our evidential approach to explanation in CBR.

References

1. Brüninghaus, S., Ashley, K.D.: Combining Case-Based and Model-Based Reasoning for Predicting the Outcome of Legal Cases. In: Ashley, K.D., Bridge, D.G. (eds.) Case-Based Reasoning Research and Development. LNAI, Vol. 2689. Springer-Verlag, Berlin Heidelberg New York (2003) 65-79
2. McSherry, D.: Explanation in Case-Based Reasoning: an Evidential Approach. In: Lees, B. (ed.) Proceedings of the 8th UK Workshop on Case-Based Reasoning (2003) 47-55
3. Murdock, J.W., Aha, D.W., Breslow, L.A.: Assessing Elaborated Hypotheses: An Interpretive Case-Based Reasoning Approach. In: Ashley, K.D., Bridge, D.G. (eds.) Case-Based Reasoning Research and Development. LNAI, Vol. 2689. Springer-Verlag, Berlin Heidelberg New York (2003) 332-346
4. McSherry, D.: Interactive Case-Based Reasoning in Sequential Diagnosis. Applied Intelligence **14** (2001) 65-76
5. McSherry, D.: Mixed-Initiative Intelligent Systems for Classification and Diagnosis. Proceedings of the 14th Irish Conference on Artificial Intelligence and Cognitive Science (2003) 146-151
6. Southwick, R.W.: Explaining Reasoning: an Overview of Explanation in Knowledge-Based Systems. Knowledge Engineering Review **6** (1991) 1-19
7. Sørmo, F., Aamodt, A.: Knowledge Communication and CBR. In: González-Calero, P. (ed.) Proceedings of the ECCBR-02 Workshop on Case-Based Reasoning for Education and Training (2002) 47-59
8. McSherry, D.: Similarity and Compromise. In: Ashley, K.D., Bridge, D.G. (eds.) Case-Based Reasoning Research and Development. LNAI, Vol. 2689. Springer-Verlag, Berlin Heidelberg New York (2003) 291-305
9. Cunningham, P., Doyle, D., Loughrey, J.: An Evaluation of the Usefulness of Case-Based Explanation. In: Ashley, K.D., Bridge, D.G. (eds.) Case-Based Reasoning Research and Development. LNAI, Vol. 2689. Springer-Verlag, Berlin Heidelberg New York (2003) 122-130
10. Leake, D.B.: CBR in Context: the Present and Future. In Leake, D.B. (ed.) Case-Based Reasoning: Experiences, Lessons & Future Directions. AAAI Press/MIT Press (1996) 3-30
11. Cendrowska, J.: PRISM: an Algorithm for Inducing Modular Rules. International Journal of Man-Machine Studies **27** (1987) 349-370
12. Quinlan, J.R.: Induction of Decision Trees. Machine Learning **1** (1986) 81-106
13. Evans-Romaine, K., Marling, C.: Prescribing Exercise Regimens for Cardiac and Pulmonary Disease Patients with CBR. In: Bichindaritz, I., Marling, C. (eds.) Proceedings of the ICCBR-03 Workshop on Case-Based Reasoning in the Health Sciences (2003) 45-52
14. Ong, L.S., Shepherd, B., Tong, L.C, Seow-Cheon, F., Ho, Y.H., Tang, C.L., Ho, Y.S., Tan, K.: The Colorectal Cancer Recurrence Support (CARES) System. Artificial Intelligence in Medicine **11** (1997) 175-188
15. McSherry, D.: Dynamic and Static Approaches to Clinical Data Mining. Artificial Intelligence in Medicine **16** (1999) 97-115
16. Szolovits, P., Pauker, S.G.: Categorical and Probabilistic Reasoning in Medical Diagnosis. Artificial Intelligence **11** (1978) 115-144
17. Spiegelhalter, D.J., Knill-Jones, R.P.: Statistical and Knowledge-Based Approaches to Clinical Decision-Support Systems with an Application in Gastroenterology. Journal of the Royal Statistical Society Series A **147** (1984) 35-77