Computing and Informatics, Vol. 25, 2006, 173–193

KLEOR: A KNOWLEDGE LITE APPROACH TO EXPLANATION ORIENTED RETRIEVAL

Lisa Cummins, Derek Bridge

Department of Computer Science
University College, Cork
Ireland
e-mail: l.cummins@cs.ucc.ie, d.bridge@cs.ucc.ie

Revised manuscript received 3 February 2006

Abstract. In this paper, we describe precedent-based explanations for case-based classification systems. Previous work has shown that explanation cases that are more marginal than the query case, in the sense of lying between the query case and the decision boundary, are more convincing explanations. We show how to retrieve such explanation cases in a way that requires lower knowledge engineering overheads than previously. We evaluate our approaches empirically, finding that the explanations that our systems retrieve are often more convincing than those found by the previous approach. The paper ends with a thorough discussion of a range of factors that affect precedent-based explanations, many of which warrant further research.

Keywords: Case-based reasoning, classification, explanation, precedent-based explanation, explanation oriented retrieval

1 INTRODUCTION

You are drinking in a bar with a friend. She tells you that, in her view, you are drunk: you are *over-the-limit* and should not drive home. To explain her judgement, she likens your situation to that of someone whose recent successful prosecution for drink driving was reported in the national press. Explanations of this kind, where a judgement is supported with reference to related cases, are called *precedent-based explanations*.

What properties should the precedent possess to make your friend's explanation convincing? A precedent that is *identical* in all relevant aspects to your own situation might sway you: someone of the same gender and weight, who had eaten as little as you and consumed as many units of alcohol as you. In the absence of an identical case, your friend might settle on a *similar* instance. But what she actually needs is a *similar but more marginal* instance: for example, someone of the same gender and weight, who had eaten as little as you but had consumed fewer units of alcohol; or someone of the same gender, who had eaten as little as you and consumed the same number of units of alcohol, but whose weight is greater than yours. You would reason that such people would be less likely than you to be over-the-limit, and yet they were.

Similarly, if you wanted to convince your friend of the opposite judgement, that you are *under-the-limit*, you would seek as precedent someone lighter, who had eaten less, or who had consumed more units of alcohol and yet who was found to be under-the-limit.

In general, we can imagine cases as points in a multi-dimensional space defined by their attribute values. A hyperplane, referred to as the *decision boundary*, separates cases of different classes (e.g. over-the-limit and under-the-limit). To support a judgement that query case Q has a particular class, the ideal *explanation case*, EC, other than an identical case, will have the same class as Q but will be situated *between* Q and the decision boundary, as shown in Figure 1.

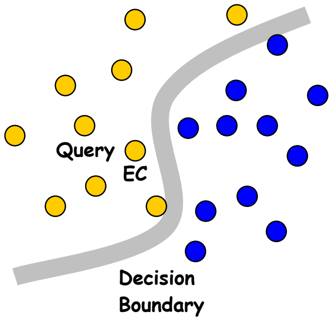

Fig. 1. Using a more marginal case as an explanation case

In this paper, we show how a case-based classifier can find explanation cases of the kind we have just described. The main contributions of the paper are: firstly we show how to find these explanation cases using an approach that has much lower knowledge engineering costs than the approach reported in [6]; and secondly we include a wide-ranging discussion of factors that affect the quality of precedent-based explanations, many of which warrant further research.

Before we do any of this, we should consider the purpose that this kind of explanation serves. The knowledge-based systems literature contains numerous categorisations of users and the different types of goal these users have when they seek a system explanation (e.g., [8, 10, 14, 16, 17]). The most recent synthesis of these categorisations is that presented by Sørmo and his co-authors [14]. They define five explanation goals:

Transparency aims to explain how the system found its answer.

Justification aims to support the answer that the system found and convince the user of its correctness.

Relevance aims to justify the strategy used by the system.

Conceptualisation aims to clarify the meaning of terms and concepts used by the system that the user may not understand.

Learning aims to increase the user's understanding of the domain.

In this paper, we are concerned with the second of these goals: *justification.* Our explanations are intended for both domain experts and for general users who may not have very much knowledge about the particular domain, but who wish to be convinced of a particular system prediction: the precedent-based explanation shows why the system's classification is the correct one. The aim is to increase the user's confidence in the system and its conclusions [20].

The rest of this paper proceeds as follows. Section 3 summarises Explanation Oriented Retrieval (EOR), which is the approach to the retrieval of marginal precedent-based explanations first described in [6]. We argue that this approach has knowledge engineering costs that can be mostly avoided. In Section 4 we describe our new approach, KLEOR, which is a knowledge-light approach to explanation oriented retrieval. We develop three variants of KLEOR, of increasing sophistication. Section 5 reports the results of an empirical comparison of EOR and the three KLEOR variants. Section 6 is a wide-ranging discussion of factors that affect precedent-based explanations in general and KLEOR explanations in particular. But first, in Section 2, we describe case-based classification, since it is these classifications that we are seeking to explain to the user.

2 CASE-BASED CLASSIFICATION

Before discussing how to find query Q's explanation case EC, we need to discuss how to predict Q's class: predicting for example whether someone is over- or under-the-limit. This is the task of *classification.*

There are many classification technologies (rules, decision trees, neural nets, etc.) but we are using a case-based, or instance-based, approach, which is more compatible than most with precedent-based explanation.

For each attribute a in set of attributes A, let case X's value for attribute a be denoted by $X.a$ and let X's class be denoted by $X.c$. Given a query Q and a case base CB, a k-NN classifier retrieves the k most similar cases to Q and predicts Q's class $Q.c$ from the k cases using a similarity-weighted vote.

The retrieval of the k cases uses a global similarity measure, which is a weighted sum of (attribute-specific) local similarities:

$$\mathrm{Sim}(X, Q) =_{def} \frac{\sum_{a \in A} w_a \times \mathrm{sim}_a(X.a, Q.a)}{\sum_{a \in A} w_a}. \tag{1}$$

For the local similarity of symbolic-valued attributes, we use the following:

$$\mathrm{sim}_a(X.a, Q.a) =_{def} \begin{cases} 1 & \text{if } X.a = Q.a \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

(but see Section 6 later). For numeric-valued attributes, we use:

$$\mathrm{sim}_a(X.a, Q.a) =_{def} 1 - \left(\frac{|X.a - Q.a|}{range_a} \right) \tag{3}$$

where $range_a$ is the difference between the maximum and minimum values allowed for attribute a.

Case-based classification lends itself to precedent-based explanation: “the results of [case-based reasoning] systems are based on actual prior cases that can be presented to the user to provide compelling support for the system’s conclusions” [9]. Research has shown that precedent-based explanations are favoured by the user over rule-based explanations [4].

It should not naïvely be assumed that the best explanation case is the nearest neighbour. In Section 1, we have already argued that, in the absence of an identical case to the query, a more marginal case is a better explanation case. This has been confirmed empirically by [6]. In the next section, we describe *explanation oriented retrieval*, which implements this idea.

3 EXPLANATION ORIENTED RETRIEVAL

After Q’s class $Q.c$ has been predicted by case-based classification, Explanation Oriented Retrieval [6], henceforth EOR, retrieves an explanation case EC using a global explanation utility measure:

$$\mathrm{Exp}(X, Q, Q.c) =_{def} \begin{cases} 0 & \text{if } Q.c \neq X.c \\ \frac{\sum_{a \in A} w_a \times \mathrm{exp}_a(X.a, Q.a, Q.c)}{\sum_{a \in A} w_a} & \text{otherwise.} \end{cases} \tag{4}$$

As you can see, the local explanation utility measures, exp_a, are sensitive to Q’s class, $Q.c$. Each would be defined by a domain expert. For example, possible local explanation utility measures (one for each predicted class) for the attribute that records the number of units of alcohol someone has consumed are plotted in Figure 2.

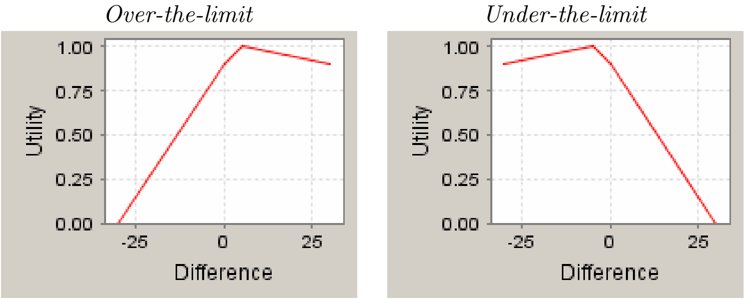

Fig. 2. Explanation-utility graphs for the units-consumed attribute

Along the x-axis is $Q.a - X.a$; the y-axis measures X's explanation utility.

Suppose Q is predicted to be over-the-limit: the left-hand plot is used. The more negative $Q.a - X.a$ is (i.e. the greater the excess X has consumed over what Q has consumed) the lower the utility of X as an explanation. When $Q.a - X.a$ is positive (i.e. X has consumed less than Q), utility is high: X and Q are both over-the-limit, but X is less likely to be over-the-limit, which makes X a good explanation case. (The graph also shows utility falling off as X becomes more remote from Q.) If Q is predicted to be under-the-limit, the right-hand plot is used and this gives higher utilities to people who have consumed more alcohol than Q.

These class-specific local explanation utility measures are inspired by the asymmetric more-is-better and less-is-better local similarity measures often used in case-based recommender systems [1].

In an experiment, in eight out of nine cases, an expert preferred as explanation an EOR explanation case (i.e. a more marginal case) over the nearest neighbour [6]. The only time that the expert preferred the nearest neighbour was when the nearest neighbour was identical to Q.

However, EOR, while billed as knowledge-light, has a not inconsiderable knowledge engineering cost. A knowledge engineer must specify the utility measures. If cases have m attributes and there are n different classes, then, in principle, $m \times n$ local utility measures are needed.

In the next section, we describe Knowledge Lite Explanation Oriented Retrieval, which seeks the same kind of explanation cases as EOR, but uses Sim and sim_a, the same similarity measures we use for case-based classification. It therefore has no additional knowledge engineering costs.

4 KNOWLEDGE LITE EXPLANATION ORIENTED RETRIEVAL

Knowledge Lite Explanation Oriented Retrieval (KLEOR) has three variants (Sim-Miss, Global-Sim and Attr-Sim). We describe each variant in turn in the following subsections; for each variant, we identify its main limitations; these limitations are then addressed by the subsequent variant.

Each variant is defined with respect to the following two cases:

- The *nearest hit NH* is the case that is *most similar* to Q that has *the same class* as Q.
- The *nearest miss NM* is the case that is *most similar* to Q that has *a different class* from Q.

(In fact, we will argue below for using a case we refer to as *NMOTB* in place of *NM*. But it aids the exposition to begin our treatment with *NM*.)

4.1 KLEOR-Sim-Miss

We reason that the decision boundary is somewhere between Q and *NM* and therefore, equally, *EC* lies between Q and *NM*, as shown in Figure 3.

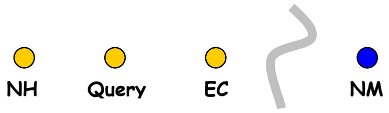

Fig. 3. The locations of *NH*, Q, *EC* and *NM*

Hence, Sim-Miss defines *EC* as follows:

Definition of *EC* for KLEOR-Sim-Miss: the case that

1. has the same class as Q;
2. is most similar to *NM*.

Here, and in all our KLEOR definitions, cases such as *NM* and thus *EC* can be retrieved using the same similarity measures used by case-based classification to predict $Q.c$.

If there are two or more cases between Q and *NM*, Sim-Miss will find the case that is most similar to *NM*, and so will find the case on the same side of the decision boundary as Q that is closest to the boundary.

4.2 KLEOR-Global-Sim

Sim-Miss makes an implicit assumption which does not always hold true. It assumes the problem space is divided neatly into two by the decision boundary, forming regions that are convex or nearly convex. However, some problem spaces are not like this. The space occupied by cases of a particular class may have concavities or be discontinuous, as we can see in Figure 4.

Applied to a space like the one depicted in Figure 4, Sim-Miss will not retrieve the best explanation case. The case found as *EC* does not lie between Q and *NM*

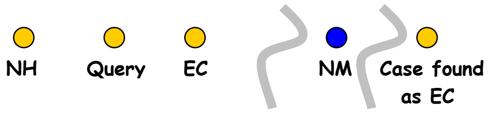

Fig. 4. A problem with Sim-Miss

but it is the case that best satisfies the Sim-Miss definition: it has the same class as Q and is closer to NM than the ideal EC is.

Global-Sim overcomes the above problem with the following definition:

Definition of EC for KLEOR-Global-Sim: the case that

1. has the same class as Q;
2. is located, according to Sim, between Q and NM:

$$\text{Sim}(Q, EC) > \text{Sim}(Q, NM)$$

3. is most similar to NM.

This rules out the problem depicted in Figure 4 by forcing Q to be closer to EC than it is to NM.

4.3 KLEOR-Attr-Sim

However, there is a naïvety in Global-Sim, fostered by our one-dimensional diagrams. Cases typically have more than one attribute; the problem space has many dimensions. Global-Sim uses global similarity to find an EC that is more similar to Q than it is to NM. But this allows one or two of the attributes to satisfy this condition strongly enough to outweigh other attributes whose values do not lie between Q and NM.

KLEOR-Attr-Sim tries to find an EC each of whose attributes has an appropriate value:

Definition of EC for KLEOR-Attr-Sim: the case that

1. has the same class as Q;
2. has the most attributes a for which:

$$\text{sim}_a(Q.a, EC.a) > \text{sim}_a(Q.a, NM.a)$$

3. is most similar to NM.

If several cases have the same number of attributes satisfying condition 2, we select from these cases the one that is most similar to NM according to global similarity.

Note that, in the case of a numeric-valued attribute whose similarity is measured using Equation 3, condition 2 is equivalent to enforcing the following:

$$\begin{array}{rcccll} Q.a & < & EC.a & < & NM.a & \text{if } Q.a < NM.a \\ NM.a & < & EC.a & < & Q.a & \text{if } Q.a > NM.a. \end{array}$$

In other words, *EC*'s value for this numeric-valued attribute lies between *Q*'s and *NM*'s values for this attribute.

In Section 6 at the end of this paper, we will discuss a variant of KLEOR-Attr-Sim that is sensitive to the attribute weights used in Equation 1.

4.4 KLEOR using the *NMOTB*

There is another naïvety in our approach. The *NM*'s position in the problem space may not be where we have been assuming that it will be. Hence, a case that lies between *Q* and *NM* may not be more marginal than *Q* and may be a poor explanation case. How can this be?

Consider using 3-NN to classify *Q*, as in Figure 5.

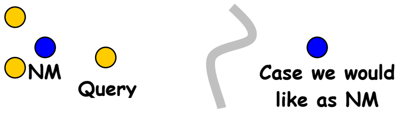

Fig. 5. The *NM* is not always what we want when finding explanations

By a weighted vote of *Q*'s 3 nearest neighbours, *Q* is predicted to belong to the light-grey class. We want *EC* to lie between *Q* and the 3-NN decision boundary. But this is not obtained by retrieving a case that lies between *Q* and *NM*.

We propose three alternative remedies, which may suit different problem domains:

Use 1-NN: Situations such as the one depicted above can only arise when using k-NN for $k > 1$. If $k = 1$, the class of *Q* is predicted from *NH* and it is impossible for a case of a different class to lie between *Q* and *NH*. While this may suit some problem domains, in others it may reduce the accuracy of the case-based classifications.

Use noise-elimination: We can run a noise elimination algorithm over the case base prior to use. These algorithms 'neaten' the decision boundary by deleting cases where a majority of the neighbours are of a different class [19]. In the situation depicted above, the case labelled *NM* would be deleted and would no longer wrong-foot KLEOR. While this may suit some problem domains, it has two problems. First, depending on the exact operation of the noise elimination algorithm, it may not eliminate all such situations. Second, in some domains it may have the undesirable consequence of deleting correct but idiosyncratic cases.

Use the *NMOTB*: Instead of seeking an explanation case *EC* that lies between *Q* and *NM*, we can seek one that lies between *Q* and the *NMOTB*, *the nearest miss that lies over the boundary.*

Definition of *NMOTB*: the case that

1. has a *different* class from *Q*;
2. is *not* located, according to Sim, between *Q* and *NH*:

$$\text{Sim}(Q, NH) > \text{Sim}(NMOTB, NH)$$

3. is *most similar* to *Q*.

Condition (2) forces *NMOTB* to be more distant from *NH* than *Q* is, and it therefore solves the problem shown in Figure 5.

Using *NMOTB* in place of *NM* is a generic solution that, unlike using 1-NN or noise elimination, works in all problem domains. (A positive side-effect that we have confirmed empirically is that it increases the number of times an *EC* can be found: sometimes no case can be found that lies between *Q* and *NM*, but a case can be found that lies between *Q* and *NMOTB*.) Henceforth, when we refer to KLEOR in any of its three variants, we will be using *NMOTB*.

5 EMPIRICAL EVALUATION

We used data collected in Dublin public houses one evening in February 2005. For each of 127 people, this dataset records five descriptive attributes: the person's weight and gender; the duration of their drinking and the number of units consumed; and the nature of their most recent meal (none, snack, lunch or full meal). On the basis of breath alcohol content (BAC), measured by a breathalyzer, each person is classified into one of two classes: over-the-limit (BAC of 36 or above) or under-the-limit (BAC below 36). We ignored a further nine instances whose data was incomplete.

We implemented the EOR system from [6] and the three KLEOR systems described in this paper (Sim-Miss, Global-Sim and Attr-Sim), all using *NMOTB*. To give the systems a default strategy, if, by their definitions, they were unable to find an *EC*, they returned *NH*, the nearest hit, as explanation.

The first question that we set out to answer is: how often does each system resort to the default strategy? We took 67 % of the data (85 cases) as a case base (training set) and we treated the remaining 33 % (42 cases) as queries (test set). For each query, we used 3-NN classification to predict the class and then we retrieved an explanation case using each of the explanation systems. We used 100 iterations, with different training and test sets in each, and recorded how often (of the total $100 \times 42 = 4200$ queries) a system had to resort to the default explanation. The results are given in Table 1.

	Defaults
EOR	5 %
Sim-Miss	20 %
Global-Sim	41 %
Attr-Sim	24 %

Table 1. The percentage of times a strategy had to resort to the default explanation

As Table 1 shows, EOR defaults least often. This is because it places no conditions on EC except that it be of the same class as Q. Sim-Miss defaults least of the KLEOR systems because it places the least restrictive conditions on EC. Counter-intuitively, perhaps, Global-Sim defaults more often than Attr-Sim. Attr-Sim seeks an EC that has attribute values that lie between those of Q and $NMOTB$. Provided a case has at least one such attribute, then it is a candidate EC; the system defaults only if it finds no cases with attributes between Q and $NMOTB$, and this is a less stringent requirement than the one imposed on global similarity by Global-Sim.

In the same experiments, we also recorded how often pairs of systems' explanations coincided. The results are shown in Table 2. (The figures are, of course, symmetrical.)

	Sim-Miss	Global-Sim	Attr-Sim
EOR	6 %	4 %	6 %
Sim-Miss	—	59 %	64 %
Global-Sim	—	—	53 %

Table 2. The percentage of times the strategies find the same explanation cases

The table shows that pairs of KLEOR variants often find the same explanations (between 53 % and 64 % of the time). In fact, we found that 43 % of the time all three KLEOR systems produced the same explanation. Rarely does EOR agree with the KLEOR systems (4 % to 6 % of the time). We found that the explanations produced by all four systems coincided only 4 % of the time.

Since the explanations produced by EOR have already been shown to be good ones [6], we would probably like it if more of the KLEOR explanations coincided with the EOR ones. However, even though they do not coincide often, KLEOR systems might be finding explanations that are as good as, or better than, EOR's. We set about investigating this by running an experiment in which we asked people to choose between pairs of explanations.

We used 30 informants, who were staff and students of our Department. We showed the informants a query case for which 3-NN classification had made a correct prediction (under-the-limit or over-the-limit). We showed the correctly-predicted class and two explanation cases retrieved by two different systems. We ensured that explanations were never default explanations and we ensured that not only had the explanations been produced by different systems but they were also different explanation cases. We asked the informants to decide which of the two explanation cases better convinced them of the prediction.

We asked each informant to make six comparisons. Although they did not know it, a person's pack of six contained an explanation from each system paired with each other system. The ordering of the pairs and the ordering of the members within each pair was random. For 30 informants, this gave us a total of 180 comparisons.

The outcomes of these 180 comparisons are shown in Table 3.

		Loser				
		EOR	Sim-Miss	Global-Sim	Attr-Sim	Total Wins
Winner	EOR	—	2	6	6	14
	Sim-Miss	27	—	20	14	61
	Global-Sim	21	8	—	7	36
	Attr-Sim	22	14	21	—	57
	Total Loses	70	24	47	27	168

Table 3. User preferences when comparing pairs of explanations

The totals do not sum to 180 because in 12 comparisons the informants found neither explanation to be better than the other.

We read the results in Table 3 as follows. Each of the 30 informants saw an EOR explanation paired with a Sim-Miss explanation, for example. Two of the 30 preferred the EOR explanation; 27 of the 30 preferred the Sim-Miss explanation; in one case neither explanation was preferred. Similarly, six informants preferred an EOR explanation over a Global-Sim explanation; 21 preferred the Global-Sim explanations; three preferred neither. EOR explanations were preferred a total of 14 of the 90 times they appeared in these experiments (row total) and were not preferred 70 times (column total); in six out of the 90 times that an EOR explanation was shown, the informant was unable to choose.

Encouragingly, KLEOR explanations are often preferred to EOR ones. Among the KLEOR systems, Global-Sim gives the least convincing explanations. Sim-Miss and Attr-Sim are barely distinguishable: their explanations are preferred over Global-Sim ones by 20 and 21 informants respectively; Sim-Miss explanations are preferred over Attr-Sim ones 14 times but Attr-Sim explanations are preferred over Sim-Miss ones 14 times also. Sim-Miss does slightly better overall by beating EOR more often than Attr-Sim does. On the one hand, there is no clear-cut winner. On the other hand, it has to be remembered that all three KLEOR systems produce the same explanation 43 % of the time (see earlier). But queries where the three systems agree on the explanation case do not figure in these experiments with human informants: we need the explanations to be different if users are to choose between them, so the results are confined to queries on which the systems disagree.

We were aware that not all our informants had the same backgrounds. We distinguished between *non-initiates*, who knew nothing of the research, and *initiates*, to whom we had at some time presented the idea of marginal precedent-based explanations prior to the experiment. There were 14 non-initiates and 16 initiates. (Some of the 14 non-initiates went on to repeat the experiment as initiates). The results are shown in Tables 4 and 5.

		Loser				
14 non-initiates		EOR	Sim-Miss	Global-Sim	Attr-Sim	Total Wins
Winner	EOR	—	1	4	4	9
	Sim-Miss	13	—	10	8	31
	Global-Sim	8	4	—	2	14
	Attr-Sim	10	6	11	—	27
	Total Loses	31	11	25	14	81

Table 4. The preferences of non-initiates

		Loser				
16 initiates		EOR	Sim-Miss	Global-Sim	Attr-Sim	Total Wins
Winner	EOR	—	1	2	2	5
	Sim-Miss	14	—	10	6	30
	Global-Sim	13	4	—	5	22
	Attr-Sim	12	8	10	—	30
	Total Loses	39	13	22	13	87

Table 5. The preferences of initiates

We do not see any major differences between the two kinds of informant. A small difference is that for non-initiates in only three of 84 comparisons was no preference expressed whereas for initiates the figure was nine of 96 comparisons (6 % more).

Looking over all of the results, we were surprised that KLEOR explanations were preferred over EOR ones. This raises an objection to our experimental set-up: it can be argued that the human-engineered local explanation utility metrics that we have used in EOR are not correct for this domain. If they were correct, they would surely be finding the better explanation cases. However, finding these 'correct' utility measures implies an even higher knowledge engineering effort.

The quality of EOR explanation cases is likely to be improved had we used richer similarity/utility measures on the symbolic-valued attributes (see next section). But the richer similarity measures are also likely to improve the quality of Attr-Sim's explanations, which, contrary to expectations, we found to be slightly worse than Sim-Miss explanations. In practice, it might also be desirable in condition 2 of the definition of Attr-Sim to set a lower limit on how many attribute values must lie between those of Q and $NMOTB$. This will make Attr-Sim default more often but, when it does not default, its explanation cases will more convincingly lie in the desired region of the problem space. In the next section, we discuss making this condition sensitive to attribute weights and even degrees of noise.

6 DISCUSSION

In this paper, we have presented an approach, KLEOR, that is even more knowledge light than Explanation Oriented Retrieval [6]. In all its variants, to find explanation cases KLEOR uses only the similarity measures that a system would be equipped with for the purposes of case-based classification. In this section, we discuss some broader issues that affect precedent-based explanations.

Symbolic-valued attributes: In the experiments we have described in this paper, we used an equality measure (Equation 2) to calculate the similarities for symbolic-valued attributes. This can be satisfactory for attributes which have only two values, such as gender, or where the values have no relationship to each other. However, sometimes domain knowledge renders some of the values more similar to others. In this case, alternative, richer similarity measures are possible. For example, for some symbolic attributes, there may be an underlying order on the values. This is the case in the breathalyzer domain for the attribute that records the person's most recent meal, where the ordering is based on the amount of food likely to have been ingested:

$$\mathit{None} < \mathit{Snack} < \mathit{Lunch} < \mathit{Full}$$

The similarity measure should show *None* and *Snack* as more similar than *None* and *Lunch*. However, using Equation 2, this will not be the case. Using the ordering, we can instead define similarity as the inverse of the distance within the ordering. So, for example, *None* and *Snack* are similar to degree $\frac{2}{3}$ (their distance is 1 out of a maximum distance of 3, and this is subtracted from 1 to give the degree of similarity), whereas *None* and *Lunch* are similar to degree $\frac{1}{3}$ (distance of 2 out of 3, subtracted from 1). For other attributes, distances in trees or graphs that represent domain-specific knowledge (e.g. taxonomic relationships) can also be used [13]. Occasionally, domain experts explicitly define similarity values for all pairs of values [1]. Recent work takes an evolutionary approach to learning such similarity measures [15].

The failure to use these richer similarity measures may be affecting some of our empirical results, especially in the case of EOR and KLEOR-Attr-Sim. Had we used these richer measures, explanation cases that more plausibly lie between the query and the decision boundary might be retrieved. Reassuringly, these richer measures are compatible with EOR and all variants of KLEOR.

Missing attribute values: In both queries and cases from the case base, some attributes may be missing their values. In the breathalyzer dataset, this was the case for 9 of 136 instances. For simplicity, in our experiments we removed these 9 and worked with just the 127 complete cases. At one level, this was not necessary; similarity or distance functions that handle missing values have been devised, e.g. [18], and these can be extended to the retrieval of EOR and KLEOR explanation cases.

However, what is more difficult is to ensure that, in the face of missing values, explanation cases will be interpretable by, and convincing to, the users. For example, suppose the query case records no value for the person's weight. Is it equally convincing to show explanation cases that are also missing the weight; ones that have an arbitrary value for weight; and ones that have a below-average value for weight for an under-the-limit prediction and an above-average value for an over-the-limit prediction? Or consider the inverse: the query has a value for weight. Is an explanation case with a missing value for weight as understandable and convincing as one without? These questions can only be answered through further detailed empirical investigation. The answers may be domain-specific and may even differ from attribute to attribute in the same domain.

Attribute weights: In making classifications, some attributes are more *important* than others. This is allowed for by attribute weights in the global similarity measure (Equation 1), although all are set to 1 in our experiments. These same weights are used by EOR in its global explanation utility measure (Equation 4) and they will also be used by Sim-Miss and Global-Sim because both use *Sim*, the global similarity measure (Equation 1). However, the weights will not be used in condition 2 of the definition of Attr-Sim because this condition uses the local similarity measures, sim_a. The definition of Attr-Sim is easily modified to take these weights into account: instead of *counting* the number of attributes that satisfy condition 2 and choosing the explanation case with the highest count, we could *sum the weights* of the attributes that satisfy condition 2 and choose the case with the highest total.

Of course, all of this assumes that the same weights should be used for both classification and explanation: see the discussion of fidelity below.

Noise: Noise in classification tasks manifests itself in the training data as incorrect mappings between attribute values and classes. Values and class labels may be incorrectly reported for any number of reasons, e.g. where measuring equipment is unreliable or where values are subjective. Noise may affect values in queries and in cases in the case base. It may lead to unreliable classification (see the discussion of uncertainty below). But here we discuss how it directly affects explanation.

To the extent that users are aware of relative uncertainties in values due to noise, it can affect the extent to which they find an explanation convincing. On an epidemiology dataset, for example, we have informally observed how an expert was sensitive to what he knew about the reliability of certain attribute values when he judged explanations that we showed him. If the values in an explanation case that make it more marginal than a query case are ones known by the user to be unreliable, then the explanation case (and the classification that it aims to support) may be treated with skepticism by the user. If the unreliability can be quantified, e.g. probabilistically, then this could be taken into account in EOR and KLEOR. In particular, KLEOR-Attr-Sim could take these probabilities (as well as attribute weights, above) into account in condition 2 of its definition.

If not taken into account, noise may be especially damaging to the kinds of explanation cases that EOR and KLEOR seek to find. Noisy cases are among the most marginal cases of their class. Noise-free cases will form class-specific clusters in the problem space; noisy cases will tend to be on the fringes of these clusters. Indeed, this is why, as discussed earlier, noise-elimination algorithms delete cases close to decision boundaries. EOR and KLEOR run the risk that the kinds of cases they retrieve as explanation cases will be noisy. This risk is greater for the KLEOR systems as currently defined because, among the eligible cases, each selects the one that is most similar to the NMOTB, i.e. the most marginal eligible case. It is possible to modify these definitions to lessen this behaviour. We discuss this further below under the heading *The role of knowledge engineering.*

Uncertainty: Occasionally, a case-based classifier may not be certain that it has classified a query correctly. This may happen, for example, if the k nearest neighbours (who vote on the class of the query) are noisy (discussed above). But it can happen even when the neighbours are noise-free. If the votes for competing classes are close, then the classification is uncertain. For example, suppose a 3-NN classifier retrieves two cases that predict that the query case is over-the-limit, and suppose that both of these cases are 0.4 similar to the query case. Suppose the third of the retrieved cases predicts that the query case is under-the-limit and it has 0.75 similarity to the query case. The votes for over-the-limit sum to 0.8; the vote for over-the-limit is 0.75: the query is classified as under-the-limit. But the vote is close; the classification is uncertain. Under these kinds of circumstances, an explanation for the classification that fails to reveal the uncertainty of the classification is, arguably, at least misleading and may, in some domains, be dangerous.

Detecting, quantifying and reporting uncertainty is a topic that has received recent attention in case-based reasoning, e.g. [2, 5]. But making the explanation reflect the uncertainty is only now being seriously tackled, e.g. [7, 12]. One possibility that is compatible with EOR and KLEOR, that we may investigate in the future, is to retrieve explanation cases for each of the closely competing outcomes. Ideally, multiple explanation cases should be presented to the user in a way that highlights the strengths and weaknesses of each outcome.

Intelligibility: Showing a whole case as an explanation might overwhelm users to the point where they are unable to appreciate why and how it explains the classification. This is particularly so if cases are made up of many attributes, because it becomes difficult for the user to appreciate the similarity between the two [14]. We have informally observed the difficulties an expert had when judging EOR and KLEOR explanation cases comprising 14 attributes; in some domains, cases may have thousands of attributes. It might be appropriate to highlight important attribute values or even to eliminate attribute values of low importance from the explanation. There is a risk, however, that this will undermine the credibility of the classification to the user: users might fear that the wool is being pulled over their eyes.

The more attributes there are, the more likely it is that only a subset of the attributes in an explanation case will support the classification; others will, singly or in combination, support a conflicting classification. For example, the explanation case for someone predicted to be over-the-limit might describe a person who weighs more and who consumed fewer units of alcohol than the query case (both of which support the classification) but who ate a smaller meal (which does not support the classification). It might be appropriate to distinguish between the attribute values in the explanation case that support and oppose the classification and to find a way of showing why the values that oppose the classification do not matter; see, e.g., [11, 7, 12].

The role of knowledge engineering: We have shown that KLEOR can retrieve convincing precedent-based explanations using the same similarity measures used for case-based classification. The question is whether there are situations in which engineered EOR-style explanation utility measures should be used instead.

One candidate can be seen if we look again at either of the EOR local explanation utility graphs (one of which is repeated here for ease of reference as Figure 6).

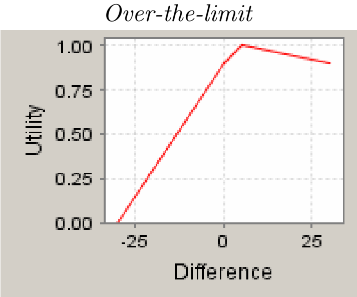
Over-the-limit

Fig. 6. Explanation-utility graphs for the units-consumed attribute

We see that the more negative $Q.a - X.a$, the lower the explanation utility of case X (left-hand half of the diagram). But we also see that utility, while high, falls off as $Q.a - X.a$ becomes more positive (right-hand half of the diagram). The utility measure falls off gently in this part of the diagram because it has been engineered to respect the judgements of human domain experts. (An added advantage relates to the point we made above that explanation cases that are closer to the decision boundary are more likely to be noisy cases. Making explanation utility fall off in this way lowers the utility of the most extreme cases.)

In an altogether different domain, that of bronchiolitis treatment, human experts required the explanation utility graph for a child's age to fall off even more sharply [7]. Here again, the EOR approach easily allowed knowledge engineers to define utility measures to respect the judgements of human experts.

Although this is a candidate for the need for engineered utility measures, it is easy to devise variants of KLEOR-Global-Sim and KLEOR-Attr-Sim that achieve similar effects. In the definitions of Global-Sim and Attr-Sim, condition 1 ensures explanation cases have the same class as the query; condition 2 tries to ensure that the explanation case lies between the query and the decision boundary; condition 3 selects from among the cases that satisfy conditions 1 and 2 the case that is closest to the boundary. Condition 3 can easily be replaced, e.g., by one that selects the eligible case that is closest to the query case or by a condition that is based on an analysis of the distribution of the eligible cases (e.g. of the eligible cases, the one whose similarity to $NMOTB$ is closest to their median similarity). Such tweaks to the definitions (if desirable at all, given that, in the domain of our empirical results, KLEOR's explanations were often preferred to EOR's) do not offer the same easy fine-tuning that EOR's engineered approach offers.

If other candidates emerge, it would be interesting to see whether a hybrid EOR/KLEOR system could be devised in which EOR's fine-tuned measures would be layered on top of KLEOR, which would act as the base strategy.

Non-binary classification: As already noted, for m attributes and n classes, EOR requires $m \times n$ local utility measures. So far, EOR has only been demonstrated for binary classification, where the number of classes $n = 2$. Admittedly, this is by far the most common scenario. However, KLEOR has the advantage that, with no additional knowledge engineering, it can find explanation cases for arbitrary n. In particular, having predicted that the query belongs to class $Q.c$, we can retrieve an explanation case with respect to the decision boundary for each other class. There remain challenges, however, in presenting such a set of explanation cases in an intelligible way to the user.

Fidelity: Explanations that seek to *justify* a decision should, in general, be true to the reasoning process that lead to that decision. However, fidelity may sometimes be traded for intelligibility. (Recall the weaknesses of the reasoning traces used to explain early rule-based expert systems [3].) It could be argued that EOR and KLEOR explanations are not true to the classifications they seek to explain: the classifications are made using the k nearest neighbours to the query; the classifications are explained using other cases, ones that lie between the query and the decision boundary. (We raised the possibility above that the weights used for classification and for explanation could be different: the desire for fidelity suggests otherwise.) However, this is not a major decrease in fidelity and informants have found EOR and KLEOR explanations to be quite convincing.

User learning: In Section 1, we listed some of the goals that explanations can serve, including transparency, justification, etc. We indicated that in this paper we have been considering the goal of justification: how to convince the user of the correctness of the system's prediction. We now suggest that the kinds of explanations produced by EOR and KLEOR could also be used to achieve the *learning* goal.

By showing a person who does not have knowledge of the domain both a query case and an explanation case that is a more marginal exemplar of the class to which the query belongs, the person can learn how the different classes in the domain relate to each other, and how different attribute values contribute to the classification.

For example, suppose the query case was classified to be *under-the-limit* and the explanation case, also classified to be *under-the-limit*, has a lower value for weight in kgs than the value in the query case. The user would know that the explanation case is more marginal than the query, and so could learn the fact that, in this domain, the less people weigh, the less likely they are to be *under-the-limit*. A user who is exposed to enough queries and explanations could learn how each of the attributes influences the classification. Further work could study the effectiveness of this form of learning.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we have reviewed the idea that precedent-based explanations in classification tasks should be more marginal exemplars of their class than the query is. We showed that the existing method for retrieving such explanation cases requires that the knowledge engineer furnish the system with separate explanation utility measures. We have proposed three new ways in which these explanation cases can be retrieved. Our new approaches go under the name of Knowledge Lite Explanation Oriented Retrieval (KLEOR) reflecting their lower knowledge engineering requirements. In particular, they retrieve explanation cases using the same similarity measures as are used for classification.

We reported the results of an extensive empirical evaluation. First, we carried out automatic experiments to determine how often different systems produce the same explanation case. Second, for queries where pairs of systems produce different explanation cases, we asked human informants to choose the better of the two explanations. Two of the new systems in particular, KLEOR-Sim-Miss and KLEOR-Attr-Sim, perform very well in our empirical investigation. However, there is no clear-cut winner.

In a wide-ranging discussion, we have mentioned a number of issues that affect precedent-based explanation, many of which give us an agenda for future research.

In particular, we need a deeper analysis of why some explanations are preferred over others, so that we might design an approach that more often retrieves the best explanation. We also need much more research into the effects of noise and ways of making explanations more intelligible.

Acknowledgement

We thank Dónal Doyle of Trinity College Dublin for sharing his breathalyzer data with us and for the helpful comments he has made to us during this research.

This publication has emanated from research conducted with financial support from Science Foundation Ireland.

REFERENCES

[1] Bergmann, R.—Breen, S.—Göker, M.—Manago, M.—Wess, S.: Developing Industrial Case-Based Reasoning Applications: The INRECA Methodology. Springer, 1998.

[2] Cheetham, W.—Price, J.: Measures of Solution Accuracy in Case-Based Reasoning Systems. In: P. Funk and P. González-Calero (Eds.): Proceedings of the Seventh European Conference on Case-Based Reasoning, pp. 106–118, Springer, 2004.

[3] Clancey, W. J.: The Epistemology of a Rule-Based Expert System — A Framework for Explanation. Artificial Intelligence, Vol. 20, 1983, pp. 215–251.

[4] Cunningham, P.—Doyle, D.—Loughrey, J.: An Evaluation of the Usefulness of Case-Based Explanation. In: K. D. Ashley and D. G. Bridge (Eds.): Proceedings of the Fifth International Conference on Case-Based Reasoning, pp. 65–79, Springer, 2003.

[5] Delany, S. J.—Cunningham, P.—Doyle, D.—Zamolotskikh, A.: Generating Estimates of Classification Confidence for a Case-Based Spam Filter. In: H. Mufioz-Avila and P. Funk (Eds.): Proceedings of the Sixth International Conference on Case-Based Reasoning, pp. 177–190, Springer, 2005.

[6] Doyle, D.—Cunningham, P. —Bridge, D. —Rahman, Y.: Explanation Oriented Retrieval. In: P. Funk and P. González-Calero (Eds.): Proceedings of the Seventh European Conference on Case-Based Reasoning, pp. 157–168, Springer, 2004.

[7] Doyle, D.—Cunningham, P.—Walsh, P.: An Evaluation of the Usefulness of Explanation in a CBR System for Decision Support in Bronchiolitis Treatment. In: I. Bichindaritz and C. Marling (Eds.): Proceedings of the Workshop on Case-Based Reasoning in the Health Sciences, Workshop Programme at the Sixth International Conference on Case-Based Reasoning, 2005, pp. 32–41.

[8] Gregor, S.—Benbesat, I.: Explanations From Intelligent Systems: Theoretical Foundations and Implications for Practice. MIS Quarterly, Vol. 23, 1999, No. 4, pp. 497–530.

[9] Leake, D.: CBR In Context: The Present and Future. In: D. Leake (Ed.): Case-Based Reasoning: Experiences, Lessons & Future Directions, pp. 3–30, MIT Press, 1996.

[10] Mao, J. Y.—Benbasat, Y.: The Use of Explanations in Knowledge-Based Systems: Cognitive Perspectives and a Process-Tracing Analysis. Journal of Management Information Systems, Vol. 17, 2000, No. 2, pp. 153–179.

[11] McSherry, D.: Explaining the Pros and Cons of Conclusions in CBR. In: P. Funk and P. González-Calero (Eds.): Proceedings of the Seventh European Conference on Case-Based Reasoning, pp. 317–330, Springer, 2004.

[12] Nugent, C.—Cunningham, P.—Doyle, D.: The Best Way to Instil Confidence is by Being Right. In: H. Muńoz-Avila and P. Funk (Eds.): Proceedings of the Sixth International Conference on Case-Based Reasoning, pp. 368–381, Springer, 2005.

[13] Osborne, H.—Bridge, D.: A Case Base Similarity Framework. In: I. Smith and B. Faltings (Eds.): Proceedings of the Third European Workshop on Case-Based Reasoning, pp. 309–323, Springer, 1996.

[14] Sørmo, F.—Cassens, J. —Aamodt, A.: Explanation in Case-Based Reasoning – Perspectives and Goals. Artificial Intelligence Review, Vol. 24, 2005, No. 2, pp. 109–143, Springer.

[15] Stahl, A.—Gabel, T.: Using Evolution Programs to Learn Local Similarity Measures. In: K. D. Ashley and D. G. Bridge (Eds.): Proceedings of the Fifth International Conference on Case-Based Reasoning, pp. 537–551, Springer, 2003.

[16] Swartout, W.—Moore, J.: Explanation in Second Generation Expert Systems. In: J.-M. David, J.-P. Krivine and R. Simmons (Eds.): Second Generation Expert Systems, pp. 543–585, Springer, 1993.

[17] Wick, M. R.—Thompson, W. B.: Reconstructive Expert System Explanation. Artificial Intelligence, Vol. 54, 1992, Nos. 1–2, pp. 33-70.

[18] Wilson, D. R.—Martinez, T. R.: Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research, Vol. 6, 1997, pp. 1–34.

[19] Wilson, D. R.—Martinez, T. R.: Reduction Techniques for Exemplar-Based Learning Algorithms. Machine Learning, Vol. 38, 2000, No. 3, pp. 257–286.

[20] Ye, L. R.—Johnson, P.: The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice. MIS Quarterly, Vol. 19, 1995, No. 2, pp. 157–172.

Lisa Cummins is a Ph. D. student in the Knowledge Engineering Group within the Department of Computer Science at University College Cork. She graduated from University College Cork in 2005 and holds a B. Sc. in Computer Science. Her research interests include case-based reasoning, especially case-based maintenance, and explanation in intelligent systems.

Derek Bridge is a senior lecturer in the Department of Computer Science at University College Cork, where he leads the Knowledge Engineering Group. His Ph. D. from the University of Cambridge was in the area of computational linguistics. His subsequent research has covered topics such as machine learning of natural language grammars, collaborative recommender systems, analyses of case-based learning, similarity measures for case-based reasoning, and applications of case-based reasoning to product recommendation and software engineering.