Artificial Intelligence in Medicine 52 (2011) 59–66

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Conversational case-based reasoning in medical decision making

David McSherry*

School of Computing and Information Engineering, University of Ulster, Coleraine BT52 1SA, Northern Ireland, United Kingdom

ARTICLE INFO

Article history:
Received 30 June 2010
Received in revised form 25 March 2011
Accepted 17 April 2011

Keywords:
Conversational case-based reasoning
Feature selection
Explanation of reasoning
Transparency
Medical classification and diagnosis

ABSTRACT

Objectives: Balancing the trade-offs between solution quality, problem-solving efficiency, and transparency is an important challenge in medical applications of conversational case-based reasoning (CCBR). For example, test selection in CCBR is often based on strategies in which the absence of a specific hypothesis (e.g., diagnosis) to be confirmed makes it difficult to explain the relevance of test results that users are asked to provide. In this paper, we present an approach to CCBR in medical classification and diagnosis that aims to increase transparency while also providing high levels of accuracy and efficiency.
Methods: We present an algorithm for CCBR called iNN(k) in which feature selection is driven by the goal of confirming a target class and informed by a measure of a feature's discriminating power in favor of the target class. As we demonstrate in a CCBR system called CBR-*Confirm*, this enables a CCBR system to explain the relevance of any question it asks the user. We evaluate the algorithm's accuracy and efficiency on a selection of datasets related to medicine and health care.
Results: The performance of iNN(k) on a given dataset is shown to depend on the value of k and on whether *local* or *global* feature selection is used in the algorithm. The combination of these parameters for which iNN(k) is most effective in addressing the trade-off between accuracy and efficiency is identified for each of the selected datasets. For example, only 42% and 51% on average of features in a complete problem description were needed by iNN(k) to provide accuracy levels of 86.5% and 84.3% respectively on the lymphography and SPECT heart datasets from the UCI machine learning repository.
Conclusion: Our results demonstrate the ability of iNN(k) to provide high levels of accuracy on most of the selected datasets, while often requiring the user to provide only a small subset of the features in a complete problem description, and enabling a CCBR system to explain the relevance of any question it asks the user.

1. Introduction

In case-based reasoning (CBR), a new problem is solved by retrieving a similar problem from a case base (i.e., a collection of previous problems with known solutions referred to as cases) and applying its solution, following adaptation if necessary, to the new problem [1–3]. Bichindaritz [4] provides a concise overview of CBR research in the health sciences, while Holt et al. [5] predict continued growth in medical applications of CBR as decision support systems gain wider acceptance in clinical practice.

In CBR approaches to medical classification and diagnosis, an ability to provide accurate and timely solutions is essential to build user trust and confidence. Another factor that may influence a CBR system's acceptability to users is its ability to explain its reasoning [6–8]. Explanation has been a topic of interest to CBR researchers for many years [6,9] and continues to attract significant research interest in the field (e.g., Refs. [10–16]), with most contributions related to classification and diagnosis tending to focus on the system's ability to explain or justify its *conclusions*. In this context, an important benefit of CBR is the ability to justify the solution to a given problem based on actual experience (e.g., by showing the user a previous similar case in which the same solution was successfully applied) [2–4,8,10]. In CBR systems that play an active role in the selection of tests on which conclusions are based, as in *conversational* CBR (CCBR) [17,18], users may also expect the system to explain the relevance of test results they are asked to provide.

In contrast to traditional CBR approaches, a description of the problem to be solved is not assumed to be available in advance in CCBR. Instead, a problem description (or query) is incrementally elicited by the system with the aim of minimizing the number of questions the user is asked before a solution is reached (e.g., Refs. [17–34]). As shown in CCBR applications such as interactive fault diagnosis and helpdesk support, guiding the selection of relevant tests is an important benefit in situations where it may be difficult for users to provide a complete problem description and/or it is important to avoid unnecessary tests.

However, balancing the trade-offs between accuracy, problem-solving efficiency, and transparency is an important challenge in

* Tel.: +44 028 7012 4130; fax: +44 028 7012 4916.
E-mail address: dmg.mcsherry@ulster.ac.uk

0933-3657/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2011.04.007

CCBR approaches to medical classification and diagnosis. One reason is that test selection in CCBR is often based on strategies (e.g., maximizing information gain) in which the absence of a specific goal or hypothesis makes it difficult to explain the relevance of questions the user is asked (i.e., why they are considered useful by the system). Most CCBR research has also tended to focus on application domains in which the structure of the case base differs in important ways from the traditional classification datasets that are common in medical applications of CBR. In CCBR applications such as fault diagnosis, the case base is typically heterogeneous (i.e., different attributes are used to describe different cases) and/or irreducible (i.e., each case has a unique solution) [17,18]. Moreover, measures such as precision and recall are often used in the evaluation of CCBR systems rather than classification accuracy, which cannot be assessed by traditional methods for an irreducible dataset.

Recently we proposed a new approach to CCBR in medical classification and diagnosis that aims to increase transparency while also providing high levels of accuracy and efficiency [35]. Feature selection in iNN(k), our CCBR algorithm, is driven by the goal of confirming a target class and informed by a measure of a feature's discriminating power in favor of the target class. Our approach to feature selection has the important advantage of enabling a CCBR system to explain the relevance of any question it asks the user in terms of its current hypothesis. Moreover, the idea of selecting tests to confirm a diagnostic hypothesis, as in our proposed approach to CCBR, has been widely discussed in studies of diagnostic reasoning in clinical medicine (e.g., Refs. [36–39]) and should thus be familiar to clinicians. In this paper, we extend our analysis of iNN(k) to include new theoretical and empirical results and a detailed study of the trade-off between the accuracy and efficiency of CCBR dialogues in the approach.

In Sections 2 and 3, we describe our approach to CCBR in iNN(k) and demonstrate the approach in a CCBR system called CBR-*Confirm*. In Section 4, we empirically investigate the performance of iNN(k) on a selection of datasets related to medicine and health care. Our conclusions are presented in Section 5.

2. Conversational CBR in iNN(k)

In this section, we describe the basic concepts in our approach to CCBR in medical classification and diagnosis, including: (1) the similarity measure used to construct the iNN(k) retrieval set, (2) the method used to select a target class at each stage of a CCBR dialogue, (3) the measure of discriminating power used to select features that are most useful for confirming the target class, and (4) the criteria used to decide when to terminate a CCBR dialogue. The example dataset that we use to illustrate the approach is the contact lenses dataset [40,41]. This small dataset contains only 24 cases and is based on a simplified version of the real-world problem of selecting a suitable type of contact lenses for an adult spectacle wearer. The attributes in the dataset, all of which have nominal values, are: age, spectacle prescription, astigmatism, and tear production rate. The classes to be distinguished, in order of their frequency in the dataset, are no contact lenses (63%), soft contact lenses (21%), and hard contact lenses (17%).

2.1. Dataset structure

Our approach to CCBR in iNN(k) assumes that the same attributes are used to describe each case in the dataset (or case base), and that each class is represented by several cases in the dataset. While this means that the dataset should be neither heterogeneous nor irreducible, there may be missing values in the dataset. Our current approach to feature selection in the algorithm also requires all attributes in the dataset to be nominal/discrete with limited numbers of values. We denote by A the set of attributes used to describe each case in the dataset. For each $a \in A$, we denote by $domain(a)$ the set of all values of a in the dataset. A case C consists of a case identifier, a problem description, and a solution. The problem description is a list of features $a = v$, one for each $a \in A$, such that $v \in domain(a) \cup \{\text{unknown}\}$. For each $a \in A$, we denote by $\pi_a(C)$ the value of a in C. The solution for the problem represented by C, which we denote by $class(C)$, is a diagnosis or other class label.

An important role in iNN(k) is played by the notion of the *supporting cases* of a given class.

Definition 1. A case C supports a given class G if $class(C) = G$.

2.2. Query elicitation and structure

In iNN(k), a description of the problem to be solved is called a *query*. An initially empty query is incrementally extended in a CCBR dialogue by asking the user questions that are most useful for solving the problem according to the criteria described later in this section. At each stage of a CCBR dialogue, the user is asked for the value of a selected attribute (e.g., tear production rate in the contact lenses dataset). If the user answers *unknown* to any question, then the dialogue moves on to the next most useful question. A non-empty query is represented as a list of problem features $Q = \{a_1 = v_1, \ldots, a_n = v_n\}$, where $n \leq |A|$ and $v_i \in domain(a_i) \cup \{\text{unknown}\}$ for $1 \leq i \leq n$. We denote by A_Q the set of attributes in the current query Q. For each $a \in A_Q$, we denote by $\pi_a(Q)$ the value of a in Q.

2.3. Similarity measure

In contrast to CBR approaches to similarity assessment that assign varying importance weights to case attributes [3], all the attributes in a given query are equally weighted in our approach to CCBR. For any case C and non-empty query Q, we define the overall similarity between C and Q as:

$$Sim(C, Q) = \frac{\sum_{a \in A_Q} sim_a(C, Q)}{|A|} \quad (1)$$

where for each $a \in A_Q$, $sim_a(C, Q)$ is a measure of the similarity between the attribute's value in the case and its value in the query defined as:

$$sim_a(C, Q) = 1 \quad \text{if } \pi_a(Q) \neq \text{ unknown and } \pi_a(C) = \pi_a(Q) \quad (2)$$

and

$$sim_a(C, Q) = 0 \quad \text{otherwise} \quad (3)$$

Definition 2. For an empty query Q, we define $Sim(C, Q) = 0$ for every case C.

2.4. The iNN(k) retrieval set

As in other CCBR algorithms, the set of most similar cases is continually updated in iNN(k) as the user's query (i.e., problem description) is elicited. For $k \geq 1$, we refer to the set of most similar cases constructed by iNN(k) in each cycle of a CCBR dialogue as the iNN(k) retrieval set. The iNN(k) retrieval set is used to identify the target class that guides the selection of features that are most useful for solving a given problem. It is also used to monitor the progress of a CCBR dialogue and decide when to terminate the dialogue. In contrast to CCBR approaches in which a similarity threshold is used to identify the most similar cases, the iNN(k) retrieval set includes any case for which the number of more similar cases is less than k.

More formally, we define the iNN(k) retrieval set for a given query Q to be:

$$r(Q, \text{iNN}(k)) = \{C : |\textit{more-similar}(C,\ Q)| < k\} \quad (4)$$

where

$$\textit{more-similar}(C,\ Q) = \{C^* :\ \textit{Sim}(C^*, Q) > \textit{Sim}(C, Q)\} \quad (5)$$

For example, 12 cases in the contact lenses dataset are equally similar (0.25) to the query Q = {tear production rate = normal}, and all other cases in the dataset have zero similarity. Thus even the iNN(k) retrieval set for $k = 1$ must include the 12 cases that are equally good candidates for retrieval. A similar strategy used in some versions of k-NN is to include all ties for the kth most similar case in the set of cases on which the solution is based [42–45].

Theorem 1. *For the empty query Q at the start of a CCBR dialogue, $r(Q, \text{iNN}(k))$ is the set of all cases in the case base.*

Proof. It can be seen from Definition 2 that *more-similar*(C, Q) is empty for every case C, and so $r(Q, \text{iNN}(k))$ is the set of all cases in the case base as required. □

Dynamic updating of the retrieval set as a problem description is incrementally extended is a feature that iNN(k) shares with the lazy (or demand-driven) approach to inductive retrieval used in many CCBR algorithms [18]. However, in contrast to inductive retrieval based on exact matching, the elimination of a case from the iNN(k) retrieval set does not mean (in general) that it can never be readmitted to the retrieval set as the problem description is further extended.

2.5. Selecting a target class

Feature selection in iNN(k) is driven by the goal of confirming a target class. At each stage of a CCBR dialogue, the target class is the class G^* that is supported by most cases in $r(Q, \text{iNN}(k))$, the iNN(k) retrieval set for the current query Q. If there is a tie for the class supported by most cases in the iNN(k) retrieval set, then the tied class that is supported by most cases in the case base as a whole is selected as the target class. We know from Theorem 1 that the iNN(k) retrieval set for the empty query at the start of a CCBR dialogue is the set of all cases in the case base. So the target class is initially the class that is supported by most cases in the case base. However, the target class may change at any stage of the dialogue depending on the class distribution in the iNN(k) retrieval set.

2.6. Measure of discriminating power

In iNN(k), the selection of features (and thus questions) that are most useful for confirming a target class is based on a simple measure of a feature's *discriminating power* in favor of the target class. For any class G, attribute a, and $v \in \textit{domain}(a)$, the discriminating power of $a = v$ in favor of G is:

$$d(a = v, G) = \frac{p(a = v|G) - p(a = v|\neg G)}{|\textit{domain}(a)|} \quad (6)$$

In the contact lenses dataset, for example, the feature astigmatism = yes occurs in 8 of the 15 cases that support no contact lenses, and in 4 of the 9 cases that support soft or hard contact lenses. As astigmatism has two values (yes, no) in the dataset, the discriminating power of astigmatism = yes in favor of no contact lenses is:

$$d(\text{astigmatism} = \text{yes, no contact lenses}) = \frac{1}{2} \times \left(\frac{8}{15} - \frac{4}{9}\right) = 0.04 \quad (7)$$

However, the feature with most discriminating power (0.40) in favor of no contact lenses is tear production rate = reduced.

In Theorems 2 and 3, we identify some basic properties of the measure of discriminating power used in iNN(k).

Theorem 2. *For any class G, attribute a, and $v \in \textit{domain}(a)$, $-0.5 \le d(a = v, G) \le 0.5$.*

Proof. Since $0 \le p(a = v|G) \le 1$, $0 \le p(a = v \mid \neg G) \le 1$, and $|\textit{domain}(a)| \ge 2$, it follows that $d(a = v,\ G) \le (1 - 0)/2 = 0.5$ and $d(a = v, G) \ge (0 - 1)/2 = -0.5$ as required. □

Theorem 3. *For any class G and attribute a with values $v_1, \ldots, v_r$, $\sum_1^r d(a = v_i, G) = 0$.*

Proof.

$$\sum_1^r d(a = v_i.G) = \sum_1^r \frac{p(a = v_i|G) - p(a = v_i|\neg G)}{r} = \frac{1}{r} \times \left(\sum_1^r p(a = v_i|G) - \sum_1^r p(a = v_i|\neg G)\right) = 0$$

□

Corollary 1. *For any class G and binary attribute a with values v_1 and v_2, $d(a, v_1, G) = -1 \times d(a, v_2, G)$.*

2.7. Local and global feature selection

The integer $k \ge 1$ used to construct the retrieval set is one important parameter in iNN(k). Another is whether *local* or *global* feature selection is used in the algorithm. In local feature selection, only features that appear in one or more cases in the iNN(k) retrieval set that support the target class are considered for selection. The assessment of a feature's discriminating power is also local (i.e., based only on cases in the iNN(k) retrieval set). In global feature selection, any feature that appears in at least one case that supports the target class, whether or not the supporting case is in the iNN(k) retrieval set, may be selected. Also in contrast to local feature selection, the assessment of a feature's discriminating power is based on all cases in the case base.

We will refer to the local and global versions of iNN(k), when necessary to distinguish between them, as iNN(k)-L and iNN(k)-G respectively. In Section 4, we present empirical results which show that the optimal choice of parameters in the algorithm depends on the dataset, for example with iNN(2)-L giving the best performance on the contact lenses dataset in our experiments.

2.8. Deciding when to stop asking questions

At each stage of a CCBR dialogue in iNN(k), the user is asked for the value of a^*, where $a^* = v^*$ is the feature selected by the algorithm as most useful for confirming the target class G^*. Typically, a CCBR dialogue continues until all cases in the iNN(k) retrieval set have the same class label G. At this point, G is selected as the solution to the current problem, and the dialogue ends with the solution being presented to the user.

Alternatively, the dialogue may reach a stage where all possible questions have been asked and there are still cases with different class labels in the iNN(k) retrieval set. In this situation, the solution presented to the user is the class supported by most cases in the iNN(1) retrieval set, whether or not $k = 1$. If there is a tie for the class supported by most cases in the iNN(1) retrieval set, then the tied class that is supported by most cases in the case base as a whole is selected as the solution. It is also possible, though unlikely, for a point in the dialogue to be reached where a most useful question cannot be identified by iNN(k). In iNN(k)-L, for example, this occurs when all cases in the iNN(k) retrieval set that support the target class have missing values for all remaining attributes. When this

happens, the solution is again the class supported by most cases in the iNN(1) retrieval set.

2.9. Overview of iNN(k)-L

Algorithm 1 is an informal description of iNN(k)-L, the version of iNN(k) that we use to demonstrate our approach to CCBR in Section 3. Conditions for termination of the CCBR dialogue are tested in Lines 5, 7, and 11 of the algorithm. The class G^* supported by most cases in the iNN(k) retrieval set is selected as the target class in Line 10. The feature $a^* = v^*$ with most discriminating power in favor of G^* is selected in Lines 15–18. As noted in Section 2.7, the assessment of discriminating power in iNN(k)-L is based only on cases in the current retrieval set. The user is asked for the value of a^* in Line 19. Finally, the user's answer is used to extend the current query in Line 20 before the CCBR cycle is repeated.

Algorithm 1. iNN(k)-L

Input: An integer $k \geq 1$ and a case base with attributes A
Output: A solution class S

```
1   Q ← {}
2   S ← undecided
3   while S = undecided do
4   begin
5     if all cases in r(Q, iNN(k)) have the same class label G
6     then S ← G
7     else if A_Q = A
8       then S ← class supported by most cases in r(Q, iNN(1))
9       else begin
10          G* ← class supported by most cases in r(Q, iNN(k));
11          if all cases in r(Q, iNN(k)) that support G* have missing values for
12          all a ∈ A − A_Q
13          then S ← class supported by most cases in r(Q, iNN(1))
14          else begin
15            select the feature a* = v* with most discriminating power
16            in favor of G* over all features a = v such that a ∈ A − A_Q
17            and π_a(C) = v for at least one C ∈ r(Q, iNN(k)) such that
18            class(C) = G*;
19            v ← askuser(a*);
20            Q ← Q ∪ {a* = v}
21            end
22          end
23  end
24  return S
```

3. CBR-*Confirm*

CBR-*Confirm* is a CCBR system for classification and diagnosis tasks based on iNN(k). As described in Section 2, an initially empty query (i.e., problem description) is incrementally extended in iNN(k) by asking the user questions selected with the goal of confirming a target class, and the CCBR dialogue continues until the target class is confirmed or another solution is reached. In this section, a brief discussion of our approach to explanation in CBR-*Confirm* is followed by an example dialogue based on the contact lenses dataset [40,41] in which iNN(k) is used with $k = 1$ and local feature selection (Algorithm 1).

3.1. Explanation in CBR-Confirm

Before answering any question in CBR-*Confirm*, the user can ask the system to explain why the question is relevant. As described in Section 2, feature selection in iNN(k) is informed by a measure of discriminating power in favor of the target class. However, CBR-*Confirm* does not attempt to explain the relevance of a selected feature $a^* = v^*$ in terms of its discriminating power, as the meaning of this measure may not be apparent to the user. Instead, it explains the relevance of a selected feature (as far as possible) by "looking ahead" one step to determine its effects on the class distribution in the iNN(k) retrieval set. For example, if G^* is the target class, Q is the current query, and all cases in the iNN(k) retrieval set for $Q \cup \{a^* = v^*\}$ support G^*, then the effect of $a^* = v^*$ will be to confirm the target class.

Alternatively, the effect of a selected feature may be to eliminate all cases that support a competing class G from the iNN(k) retrieval set. In contrast to inductive retrieval approaches based on exact matching, this does not mean that such cases can never be readmitted to the iNN(k) retrieval set as the user's query is further extended. However, it can be explained to the user that the selected feature, if present, will provide evidence against the competing class. If a selected feature has neither of these effects, then CBR-*Confirm*'s explanation of its relevance is simply that it may help to confirm the target class.

At each stage of a CCBR dialogue, CBR-*Confirm* also displays a graph showing the percentage of cases in the iNN(k) retrieval set that support each class. This provides a visualization of the reasoning process that enables the user to see immediately the effects of a reported finding on the class distribution in the iNN(k) retrieval set. Changes in the target class that the system is attempting to confirm (i.e., the class currently supported by most cases in the retrieval set) are also immediately visible to the user. A similar approach to visualization of the reasoning process is used in CBR Strategist [7], a CCBR system based on inductive retrieval.

At the end of a CCBR dialogue, CBR-*Confirm* displays the class G it has selected (based on the criteria described in Section 2) as the solution to the problem described by the user. The system also explains its conclusion by showing the user the most similar case that supports G. If two or more cases that support G are equally similar to the problem described by the user (and more similar than any other supporting case) then the first such case in the case base is presented as the solution case. Also with the aim of increasing transparency, features that match the problem described in the user's query are highlighted in the solution case.

3.2. Example dialogue in CBR-Confirm

Table 1 shows the questions asked, and the user's answers, in a CCBR dialogue in CBR-*Confirm* based on the contact lenses dataset [40,41]. Feature selection is based on iNN(1)-L, and the target class in each cycle of the example dialogue is shown. The table also shows the explanation provided by CBR-*Confirm*, if requested by the user, in each cycle of the CCBR dialogue. (As the task of contact lenses selection is highly simplified in the dataset, the example dialogue should not be regarded as a realistic example of decision making in the domain.)

At the start of the example dialogue, CBR-*Confirm* selects the majority class in the dataset (no contact lenses) as the target class. The feature with most discriminating power in favor of the target class (Section 2.6) is tear production rate = reduced, and so CBR-*Confirm* asks the user for the tear production rate. In light of the user's answer (tear production rate = normal), the class now supported by most cases in the iNN(1) retrieval set is soft contact lenses. CBR-*Confirm* therefore changes the target class to soft contact lenses, and identifies astigmatism = no as the feature with most discriminating power in favor of the new target class. There are no further changes in the target class as the dialogue continues, and soft contact lenses is finally confirmed as the solution even though the spectacle prescription is unknown to the user in the third cycle.

At the end of the example CCBR dialogue, there are two cases in the iNN(1) retrieval set, namely Cases 2 and 6. Both of these cases have the same solution, and they are equally similar (0.75) to the final query Q = {tear production rate = normal, astigmatism = no, spectacle prescription = unknown, age = young}. The first of the two most similar cases is thus presented to the user as the solution case. The solution case (Case 2) is shown in Table 2. Matching features in

Table 1
Questions asked by CBR-*Confirm*, and the user's answers, in a CCBR dialogue based on the contact lenses dataset with question selection guided by iNN(1)-L. The target class in each cycle of the CCBR dialogue is also shown, together with the explanations provided by CBR-*Confirm* if requested by the user.

Cycle no.	Target class	Question	Explanation	User's answer
1	No contact lenses	Tear production rate?	If tear production rate = reduced this will confirm no contact lenses	Normal
2	Soft contact lenses	Astigmatism?	If astigmatism = no this will be evidence against hard contact lenses	No
3	Soft contact lenses	Spectacle prescription?	If spectacle prescription = hypermetrope this will confirm soft contact lenses	Unknown
4	Soft contact lenses	Age?	If age = young this will confirm soft contact lenses	Young

the solution case are indicated by a plus sign (+), and are similarly highlighted in CBR-*Confirm*.

4. Experiments

In this section, we evaluate the performance of iNN(k) on a selection of datasets related to medicine and health care. The performance measures of interest in the evaluation are classification accuracy and problem-solving efficiency as measured by the average number of questions required to reach a solution. We expect to find that the algorithm's accuracy and efficiency on a given dataset depends on the value of k and also on whether local or global feature selection (Section 2.7) is used in the algorithm.

4.1. Selected datasets

The datasets used in our experiments (all of which are available from the UCI machine learning repository [41]) are described in Table 3. All attributes in the selected datasets are nominal or discrete, and there are missing values only in the breast cancer and primary tumor datasets. The number of attributes shown for each dataset does not include the class attribute. The SPECT heart dataset [46] includes both the training and testing data provided in the UCI repository.

4.2. Experimental methodology

Leave-one-out cross validation [47] is used to evaluate each algorithm on the selected datasets. For this purpose, each case is temporarily removed from the dataset and the problem features in the left-out case are used to provide the description of a problem to be solved by a CCBR system based on iNN(k). During the problem-solving process, features in the left-out case are revealed by a simulated user in answer to questions selected by the CCBR system. When asked for an attribute value that is missing in the left-out case, the simulated user answers *unknown*. At the end of each dialogue, the number of questions asked by the system is recorded as well as whether or not the solution is correct (i.e., the same as in the left-out case). The problem description from the left-out case is also presented to a basic k-NN classifier in which ties for the kth most similar case are broken by selecting the tied cases that occur first in the dataset.

4.3. Classification accuracy

Table 4 shows the accuracy of k-NN for $k = 1, 3$, and 5 and iNN(k)-L/G for $k = 1$, 2, and 3 on each of the selected datasets. The best accuracy results are shown in bold for each dataset. Maximum levels of accuracy in iNN(k) can be seen to exceed those achieved by k-NN on all datasets except primary tumor. For example, the highest accuracy on breast cancer (75.2%) was achieved by iNN(2)-G and iNN(3)-G. The highest levels of accuracy on lymphography (86.5%) and SPECT heart (84.3%) were also achieved by iNN(2)-G, while accuracy on contact lenses was highest (83.3%) in iNN(1)-L and iNN(2)-L. The best accuracy on primary tumor (41.3%) was achieved by k-NN with $k = 5$.

The results support our hypothesis that the accuracy of iNN(k) on a given dataset depends on the value of k and on whether local or global feature selection is used in the algorithm. It is also worth noting that accuracy does not always increase as k increases in iNN(k). In iNN(k)-G, for example, accuracy can be seen to decrease or remain the same for all datasets as k increases from 2 to 3.

4.4. Dialogue efficiency

Average lengths of iNN(k) dialogues on the selected datasets are shown in Table 5. The number of attributes in each dataset (i.e., the maximum possible length of a CCBR dialogue) is also shown in the table. The results support our hypothesis that the efficiency of iNN(k) on a given dataset depends on the value of k, and on whether local or global feature selection is used in the algorithm. They also reveal some interesting patterns in the algorithm's performance in terms of dialogue efficiency. For example, the average length of CCBR dialogues in iNN(k)-L/G can be seen to increase or remain unchanged for all five datasets as k increases from 1 to 3. There is also a clear tendency for average dialogue length to increase from iNN(k)-L to iNN(k)-G for $k = 1$, 2, and 3.

The efficiency of CCBR dialogues in iNN(k) is most apparent in the results for the two datasets with the largest numbers of attributes, namely lymphography (18) and SPECT heart (22). In lymphography, for example, less than 50% of features in a complete problem description are required on average to reach a solution in all six versions of iNN(k). However, lower levels of dialogue efficiency were achieved on some of the other datasets, for example with average dialogue lengths for breast cancer ranging from 70% to 91% of the number of features (9) in a complete problem description.

4.5. Accuracy vs. efficiency

A trade-off between accuracy and efficiency can be seen, for example, in the iNN(k)-G results for lymphography. An increase in accuracy from 79.1% in iNN(1)-G to 86.5% in iNN(2)-G (Table 4) is gained at the expense of an increase in average dialogue length from 5.5 to 7.5 (Table 5). Nevertheless, the average number of features (7.5) required for 86.5% accuracy in iNN(2)-G is much smaller than the number of features (18) in a complete problem description. Average dialogue length required for maximum accuracy in iNN(k) ranges from 42% to 84% of the numbers of attributes in the five datasets, with an overall average of 62%.

Among the six versions of iNN(k) evaluated in our experiments, iNN(2)-G was most effective in addressing the trade-off between

Table 2
Solution case presented to the user at the end of the example CCBR dialogue in CBR-*Confirm*.

	Age	Spectacle prescription	Astigmatism	Tear production rate	Class
Case 2:	Young (+)	myope	No (+)	Normal (+)	Soft contact lenses

Table 3
Datasets used in the experiments.

	No. of attributes	No. of cases	No. of classes	Missing values
Contact lenses	4	24	3	No
Breast cancer	9	286	2	Yes
Primary tumor	17	339	21	Yes
Lymphography	18	148	4	No
SPECT heart	22	267	2	No

Table 4
Accuracy of k-NN and iNN(k)-L/G on the selected datasets. The best accuracy results for each dataset are shown in bold.

Dataset	k-NN			iNN(k)-L			iNN(k)-G		
	$k=1$	$k=3$	$k=5$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
Contact lenses	75.0	62.5	70.8	**83.3**	**83.3**	70.8	70.8	70.8	70.8
Breast cancer	72.7	73.4	73.4	70.3	73.1	74.5	71.7	**75.2**	**75.2**
Primary tumor	33.0	36.6	**41.3**	34.8	40.4	40.4	36.6	39.5	39.5
Lymphography	78.4	79.7	81.8	74.3	79.1	83.1	79.1	**86.5**	85.8
SPECT heart	73.0	74.9	78.7	77.5	82.4	82.0	79.0	**84.3**	83.5

Table 5
Average lengths of iNN(k) dialogues (i.e., average numbers of questions asked) on the selected datasets. The number of attributes in each dataset is also shown.

Dataset	No. of attributes	iNN(k)-L			iNN(k)-G		
		$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
Contact lenses	4	2.1	2.1	2.4	2.1	2.3	2.4
Breast cancer	9	6.3	6.9	7.5	6.8	7.6	8.2
Primary tumor	17	9.9	13.4	14.4	11.6	14.9	15.6
Lymphography	18	5.0	6.3	7.3	5.5	7.5	8.6
SPECT heart	22	8.1	8.8	9.8	9.7	11.2	12.3

accuracy and efficiency on breast cancer, lymphography, and SPECT heart, while iNN(2)-L gave the best iNN(k) results on contact lenses and primary tumor. We now look more closely at the characteristics of CCBR dialogues in the versions of iNN(k) that gave the best results on the respective datasets. Fig. 1 shows the minimum, average, and maximum lengths of CCBR dialogues in iNN(2)-G on breast cancer, lymphography, and SPECT heart, and the corresponding results for iNN(2)-L on contact lenses and primary tumor.

An interesting feature of the results is that the maximum possible dialogue length (e.g., 18 on lymphography) is reached on all datasets. This means that in some CCBR dialogues, the simulated user was asked to provide a complete description of the problem (or as near a complete description as possible). This applies equally to the datasets (contact lenses, lymphography, and SPECT heart) in which the absence of missing values means that the simulated user is never forced to answer *unknown* to any question. On the other hand, minimum dialogue length is less than 5 for all datasets.

Fig. 2 shows the cumulative frequencies, in percentages, of the lengths of CCBR dialogues in iNN(2)-G on breast cancer, lymphography, and SPECT heart, and the corresponding results for iNN(2)-L on contact lenses and primary tumor. The results clearly show the efficiency of iNN(2)-G on lymphography, with at most 5 of the 18 features in a complete problem description being used in more than 50% of CCBR dialogues. A high level of dialogue efficiency can also be seen in the iNN(2)-G results for SPECT heart, with at most 7 of the 22 features in a complete problem description being used in more than 50% of dialogues. However, the steep rise in cumulative frequency from 21 to 22 questions in SPECT heart shows that more than 25% of dialogues extend to full length (22) on this dataset. It can also be seen that CCBR dialogues in iNN(2)-G are much less efficient on the breast cancer dataset, with the simulated user being asked at least 8 of the 9 possible questions in more than 50% of dialogues.

In iNN(2)-L, there is a marked contrast in the efficiency of CCBR dialogues on contact lenses and primary tumor. For example, a solution is reached after only one question has been asked in 50% of contact lenses dialogues, while more than 50% of primary tumor dialogues extend to the maximum length of 17 questions. On the

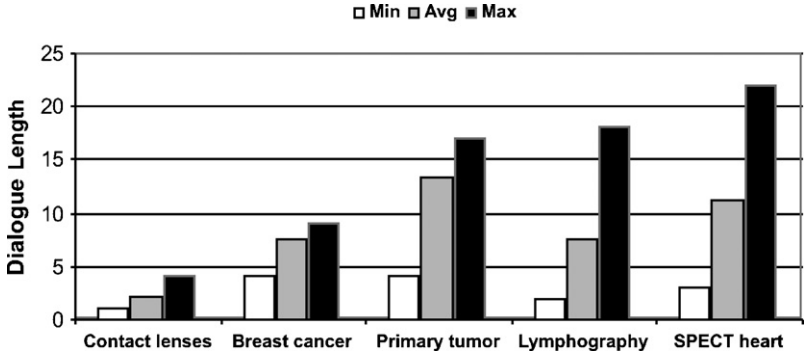

Fig. 1. Minimum, average, and maximum lengths of CCBR dialogues in iNN(2)-G on breast cancer, lymphography, and SPECT heart, and in iNN(2)-L on contact lenses and primary tumor.

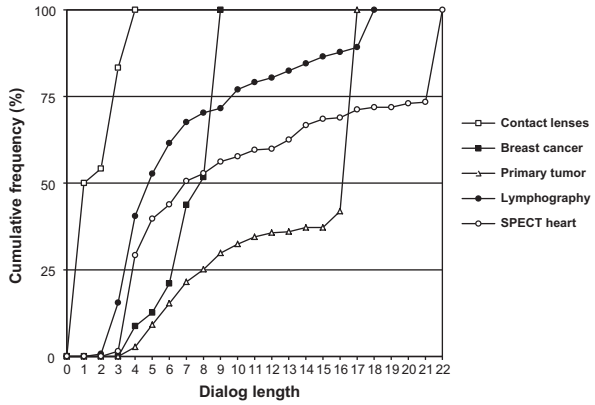

Fig. 2. Cumulative frequencies, in percentages, of the lengths of CCBR dialogues in iNN(2)-G on breast cancer, lymphography, and SPECT heart, and in iNN(2)-L on contact lenses and primary tumor.

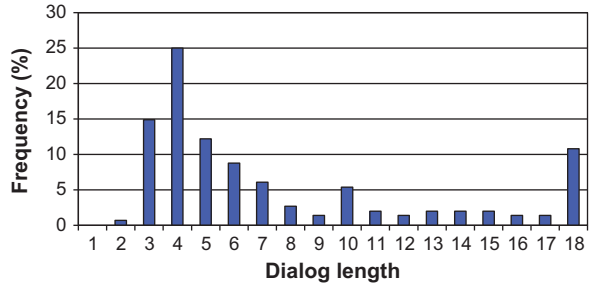

Fig. 3. Frequencies, in percentages, of the lengths of CCBR dialogues in iNN(2)-G on the lymphography dataset.

other hand, at most 8 questions are asked in 25% of primary tumor dialogues.

Fig. 3 shows the frequencies, in percentages, of the lengths of CCBR dialogues in iNN(2)-G on the lymphography dataset. Again the results clearly show the efficiency of iNN(2)-G on this dataset, with a modal dialogue length of 4 and dialogue lengths of 3, 4, and 5 accounting for more than 50% of CCBR dialogues.

5. Conclusions

In this paper, we presented and evaluated an approach to CCBR in medical classification and diagnosis that aims to increase transparency while also providing high levels of accuracy and efficiency. Feature selection in iNN(k), our CCBR algorithm, is driven by the goal of confirming a target class and informed by a measure of a feature's discriminating power in favor of the target class. As demonstrated in a CCBR system called CBR-*Confirm*, this enables a CCBR system to explain the relevance of any question it asks the user. We also presented the results of an empirical study in which iNN(k) was applied to a selection of datasets related to medicine and health care from the UCI machine learning repository [41].

The performance of iNN(k) on a given dataset was shown to depend on the value of k and on whether feature selection is performed locally or globally in the dataset. We investigated several combinations of these parameters and identified the version of iNN(k), among those studied, that was most effective in addressing the trade-off between accuracy and efficiency on each of the selected datasets. Our results show the ability of iNN(k) to provide high levels of accuracy on most of the selected datasets, while often requiring the user to provide only a small subset of the features in a complete problem description. For example, only 42% and 51% on average of the features in a complete problem description were needed for the maximum levels of accuracy achieved by iNN(k) on lymphography (86.5%) and SPECT heart (84.3%).

While our analysis of iNN(k) in this paper has focused on its potential role as an algorithm for CCBR, a non-interactive version of the algorithm could also be used to guide feature selection with the aim of increasing accuracy in situations where a problem description is provided in advance, as in traditional CBR approaches to medical classification and diagnosis. However, a limitation of our current approach to feature selection in iNN(k) is the requirement for all attributes in the dataset to be nominal or discrete. In future research, we aim to address this issue by investigating alternative approaches to feature selection in iNN(k) for datasets with continuous attributes.

Acknowledgments

Thanks to Matjaz Zwitter and Milan Soklic for providing the breast cancer, lymphography, and primary tumor datasets in the UCI machine learning repository. The author is also grateful to the reviewers for their insightful comments and suggestions.

References

[1] Aamodt A, Plaza E. Case-based reasoning: foundational issues, methodological variations, and system approaches. AI Commun 1994;7:39–59.
[2] Leake DB. CBR in context: the present and future. In: Leake DB, editor. Case-based reasoning: experiences, lessons & future directions. Menlo Park, CA: AAAI Press/MIT Press; 1996. p. 3–30.
[3] López de Mántaras R, McSherry D, Bridge D, Leake D, Smyth B, Craw S, et al. Retrieval, reuse, revision and retention in case-based reasoning. Knowl Eng Rev 2005;20:215–40.
[4] Bichindaritz I. Guest editorial: Case-based reasoning in the health sciences. Artif Intell Med 2006;36:121–5.
[5] Holt A, Bichindaritz I, Schmidt R, Perner P. Medical applications in case-based reasoning. Knowl Eng Rev 2005;20:289–92.
[6] Leake D, McSherry D. Introduction to the special issue on explanation in case-based reasoning. Artif Intell Rev 2005;24:103–8.
[7] McSherry D. Interactive case-based reasoning in sequential diagnosis. Appl Intell 2001;14:65–76.
[8] Sørmo F, Cassens J, Aamodt A. Explanation in case-based reasoning: perspectives and goals. Artif Intell Rev 2005;24:109–43.
[9] Rissland EL. The fun begins with retrieval: explanation and CBR. In: Roth-Berghofer TR, Göker MH, Güvenir HA, editors. Proceedings of the 8th European conference on case-based reasoning. Heidelberg: Springer; 2006. p. 1–8.
[10] Cunningham P, Doyle D, Loughrey J. An evaluation of the usefulness of case-based explanation. In: Ashley KD, Bridge DG, editors. Proceedings of the 5th international conference on case-based reasoning. Heidelberg: Springer; 2003. p. 122–30.
[11] Doyle D, Cunningham P, Bridge D, Rahman Y. Explanation oriented retrieval. In: Funk P, González-Calero PA, editors. Proceedings of the 7th European conference on case-based reasoning. Heidelberg: Springer; 2004. p. 157–68.
[12] Evans-Romaine K, Marling C. Prescribing exercise regimens for cardiac and pulmonary disease patients with CBR. In: McGinty L, editor. ICCBR 2003 workshop proceedings. Trondheim: NTNU, Department of Computer and Information Science; 2003. p. 45–52.
[13] Massie S, Craw S, Wiratunga N. A visualisation tool to explain case-base reasoning solutions for tablet formulation. In: Macintosh A, Ellis R, Allen T, editors. Proceedings of the 24th SGAI international conference on innovative techniques and applications of artificial intelligence. London: Springer; 2004. p. 222–34.
[14] Maximini R, Freßmann A, Schaaf M. Explanation service for complex CBR applications. In: Funk P, González-Calero PA, editors. Proceedings of the 7th European conference on case-based reasoning. Heidelberg: Springer; 2004. p. 302–16.
[15] McSherry D. Explaining the pros and cons of conclusions in CBR. In: Funk P, González-Calero PA, editors. Proceedings of the 7th European conference on case-based reasoning. Heidelberg: Springer; 2004. p. 317–30.
[16] Roth-Berghofer TR. Explanations and case-based reasoning: foundational issues. In: Funk P, González-Calero PA, editors. Proceedings of the 7th European conference on case-based reasoning. Heidelberg: Springer; 2004. p. 389–403.
[17] Aha DW, Breslow LA, Muñoz-Avila H. Conversational case-based reasoning. Appl Intell 2001;14:9–32.
[18] Aha DW, McSherry D, Yang Q. Advances in conversational case-based reasoning. Knowl Eng Rev 2005;20:247–54.
[19] Aha DW. The Omnipresence of case-based reasoning in science and application. Knowl-Based Syst 1998;11:261–73.
[20] Branting K, Lester J, Mott B. Dialogue management for conversational case-based reasoning. In: Funk P, González-Calero PA, editors. Proceedings of the 7th

European conference on case-based reasoning. Heidelberg: Springer; 2004. p. 77–90.
[21] Bogaerts S, Leake D. What evaluation criteria are right for CCBR? Considering rank quality. In: Roth-Berghofer TR, Göker MH, Güvenir HA, editors. Proceedings of the 8th European conference on case-based reasoning. Heidelberg: Springer; 2006. p. 385–99.
[22] Carrick C, Yang Q, Abi-Zeid I, Lamontagne L, Activating CBR. systems through autonomous information gathering. In: Althoff K-D, Bergmann R, Branting K, editors. Proceedings of the 3rd international conference on case-based reasoning. Heidelberg: Springer; 1999. p. 74–88.
[23] Doyle M, Cunningham P. A dynamic approach to reducing dialog in on-line decision guides. In: Blanzieri E, Portinale L, editors. Proceedings of the 5th European workshop on case-based reasoning. Heidelberg: Springer; 2000. p. 49–60.
[24] Göker MH, Thompson CA. Personalized conversational case-based recommendation. In: Blanzieri E, Portinale L, editors. Proceedings of the 5th European workshop on case-based reasoning. Heidelberg: Springer; 2000. p. 99–111.
[25] Gómez-Gauchía H, Díaz-Agudo B, González-Calero P. Ontology-driven development of conversational CBR systems. In: Roth-Berghofer TR, Göker MH, Güvenir HA, editors. Proceedings of the 8th European conference on case-based reasoning. Heidelberg: Springer; 2006. p. 309–24.
[26] Gu M, Aamodt A. A knowledge-intensive method for conversational CBR. In: Muñoz-Avila H, Ricci F, editors. Proceedings of the 6th international conference on case-based reasoning. Heidelberg: Springer; 2005. p. 296–311.
[27] Gu M, Aamodt A, Evaluating CBR. systems using different data sources: a case study. In: Roth-Berghofer TR, Göker MH, Güvenir HA, editors. Proceedings of the 8th European conference on case-based reasoning. Heidelberg: Springer; 2006. p. 121–35.
[28] Gupta KM. Taxonomic conversational case-based reasoning. In: Aha DW, Watson I, editors. Proceedings of the 4th international conference on case-based reasoning. Heidelberg: Springer; 2001. p. 219–33.
[29] Kohlmaier A, Schmitt S, Bergmann R. A similarity-based approach to attribute selection in user-adaptive sales dialogues. In: Aha DW, Watson I, editors. Proceedings of the 4th international conference on case-based reasoning. Heidelberg: Springer; 2001. p. 306–20.
[30] McSherry D, Hassan S, Bustard D. Conversational case-based reasoning in self-healing and recovery. In: Althoff K-D, Bergmann R, Minor M, Hanft A, editors. Proceedings of the 9th European conference on case-based reasoning. Heidelberg: Springer; 2008. p. 340–54.
[31] McSherry D. Increasing dialogue efficiency in case-based reasoning without loss of solution quality. In: Gottlob G, Walsh T, editors. Proceedings of the 18th international joint conference on artificial intelligence. San Francisco, CA: Morgan Kaufmann; 2003. p. 121–6.
[32] McSherry D. Minimizing dialog length in interactive case-based reasoning. In: Nebel B, editor. Proceedings of the 17th international joint conference on artificial intelligence. San Francisco, CA: Morgan Kaufmann; 2001. p. 993–8.
[33] Shimazu H, ExpertClerk:. A conversational case-based reasoning tool for developing salesclerk agents in e-commerce webshops. Artif Intell Rev 2002;18:223–44.
[34] Shimazu H, Shibata A, Nihei K, ExpertGuide:. A conversational case-based reasoning tool for developing mentors in knowledge spaces. Appl Intell 2001;14:33–48.
[35] McSherry D. Conversational case-based reasoning in medical classification and diagnosis. In: Combi C, Shahar Y, Abu-Hanna A, editors. Proceedings of the 12th conference on artificial intelligence in medicine. Heidelberg: Springer; 2009. p. 116–25.
[36] Elstein A, Schwarz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. Brit Med J 2002;324:729–32.
[37] Elstein A, Schulman L, Sprafka S. Medical problem solving: an analysis of clinical reasoning. Cambridge, MA: Harvard University Press; 1978.
[38] Kassirer A, Kuipers BJ, Gorry GA. Toward a theory of clinical expertise. Am J Med 1982;73:251–9.
[39] Patel V, Arocha J, Zhang J. Thinking and reasoning in medicine. In: Holyoak K, Morrison RG, editors. The Cambridge handbook of thinking and reasoning. New York: Cambridge University Press; 2005. p. 727–50.
[40] Cendrowska J. PRISM: an algorithm for inducing modular rules. Int J Man Mach Stud 1987;27:349–70.
[41] Frank A, Asuncion A. UCI Machine learning repository. Irvine, CA: University of California, School of Information and Computer Science; 2010.
[42] Cover TM, Hart PE. Nearest neighbor pattern classification. IEEE Trans Inform Theory 1967;1:21–7.
[43] R Development Core Team. A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2009.
[44] Ripley BD. Pattern classification and neural networks. Cambridge, UK: Cambridge University Press; 1996.
[45] Zhua M, Chena W, Hirdes J, Stolee P. The k-nearest neighbor algorithm predicted rehabilitation potential better than current clinical assessment protocol. J Clin Epidemiol 2007;60:1015–21.
[46] Kurgan LA, Cios KJ, Tadeusiewicz R, Ogiela M, Goodenday LS. Knowledge discovery approach to automated cardiac SPECT diagnosis. Artif Intell Med 2001;23:149–69.
[47] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Mellish CS, editor. Proceedings of the 14th international joint conference on artificial intelligence. San Mateo, CA: Morgan Kaufmann; 1995. p. 1137–43.