Article

Case-based reasoning for predicting the success of therapy

Rosanne Janssen,[1] Pieter Spronck[2] and Arnoud Arntz[1,3]

(1) Department of Clinical Psychological Science, Maastricht University, 6200 MD, Maastricht, The Netherlands
Email: rosanne.janssen@maastrichtuniversity.nl
(2) Tilburg center for Cognition and Communication (TiCC), Tilburg University, Tilburg, The Netherlands
(3) Department of Clinical Psychology, University of Amsterdam, Amsterdam, The Netherlands

Abstract: *For patients with mental health problems, various treatments exist. Before a treatment is assigned to a patient, a team of clinicians must decide which of the available treatments has the best chance of succeeding. This is a difficult decision to make, as the effectiveness of a treatment might depend on various factors, such as the patient's diagnosis, background and social environment. Which factors are the predictors for successful treatment is mostly unknown. In this article, we present a case-based reasoning approach for predicting the effect of treatments for patients with anxiety disorders. We investigated which techniques are suitable for implementing such a system to achieve a high level of accuracy. For our evaluation, we used data from a professional mental healthcare centre. Our application correctly predicted the success factor of 65% of the cases, which is significantly higher than the prediction of the baseline of 55%. Under the condition that the prediction was based on only cases with a similarity of at least 0.62, the success rate of 80% of the cases was predicted correctly. These results warrant further development of the system.*

Keywords: case-based reasoning, prediction, forecasting-by-analogy, weighted voting, information gain, nearest neighbour method, accuracy, mental health care, treatment outcome, anxiety disorders

1. Introduction

For mental health patients with anxiety disorders, a variety of possible treatments is available. The decision, of which treatment a patient is offered, is mainly the responsibility of healthcare professionals. For the patient's health, and for cost-effectiveness, that the treatment that is offered should be effective in dealing with the patient's problems. However, not every treatment is equally effective for every patient. Moreover, there is no consistent empirical evidence for patient-treatment matching rules (Spinhoven *et al.*, 2008). The effectiveness of most treatments for a specific patient cannot be predicted by a therapist. Despite such predictions being hard to make, they are the first step in successfully matching patient and treatment. In this study, we aim to use case-based reasoning (CBR) to make such predictions with a high level of accuracy.

This study was performed at the Community Mental Health Centre Maastricht[1] (CMHCM), the Netherlands. This centre treats patients with all sorts of mental disorders. As a starting point, the study focuses on patients with anxiety disorders and the effect of cognitive behavioural therapy. To make treatment predictions, a CBR system called *CBRth* was built, where *CBR* stands for *case-based reasoning* and *th* for *therapy*. The goal of this study is twofold: (1) to investigate which techniques are suitable for implementing CBRth and (2) to investigate the accuracy of CBRth.

This study is the first step in a research project in which the ultimate goal is the creation of an advisory system for a team of experts to assign the most effective therapy from the therapies offered by the centre to a patient. We chose CBR, because it is ideally suited for dealing with real-world-examples. Moreover, patient information often contains missing values, and as opposed to many competing methods, CBR offers various ways to deal with those, making it a suitable approach to our problem domain. Finally, because predictions made by a CBR can be made transparent to the users, they tend to have a high level of acceptability. We discussed the proposed system with our intended users, and found that they were positive about the concept.

This paper is organised as follows: Section 2 provides background information on the context of the clinical setting, and reviews CBR in mental health care and related techniques we used in this study. Section 3 specifies the CBRth casebase. Section 4 describes the CBRth process. Section 5 discusses the experimental setup. Section 6 gives the results of this study, which are discussed in Section 7. Section 8 presents our concluding remarks.

2. Background

This section provides background information on the context of the clinical setting where the research was carried out (Section 2.1). It also discusses background literature on CBR (Section 2.2) and CBR techniques (Section 2.3).

---

[1]Dutch name for Community Mental Health Centre Maastricht is Riagg Maastricht.

2.1. The context

The CMHCM is mainly concerned with providing treatments for mental health patients. It is also involved with research, for which it cooperates with Maastricht University. An important research domain is the effect of treatments. Over the years, the university gathered a large dataset on the effects of treatments given at CMHCM.

When new patients arrive at CMHCM, they are assigned to treatment after a screening procedure by an intake staff. This process involves the following people: (1) the patient, (2) the screener (the therapist who determines the patient's diagnoses), (3) the intake staff (team of screeners) and (4) the therapist who provides the treatment. The procedure is as follows:

1. The patient is screened in two or three sessions to determine the patient's diagnoses. During the screening, the screener does an anamnestic interview and formulates the targets of treatment together with the patient.
2. The screener presents the case in a meeting of the intake staff, after which the case is discussed and a decision is taken on which treatment, if any, is offered to the patient.
3. In a next session, the screener explains the treatment to the patient.
4. Before the treatment starts, there is a pre-assessment. The patient completes three questionnaires. The questionnaire scores represent the baseline state of the patient. Patients give informed consent for the use of these data for scientific research.
5. The patient receives treatment for several months (usually 4–6).
6. After the treatment, a post-assessment is carried out with the same questionnaires to determine which state the patient is in after treatment. By comparing the two assessments, the effect of the treatment can be determined.

To describe the patient's diagnoses, the screener uses the Diagnostic and Statistical Manual of mental disorders (DSM; (American Psychiatric Association, 1994)), a classification of mental disorders. The DSM consists of five axes (domains) on which disorders and other relevant issues can be assessed. Axis I contains all mental disorders except Personality Disorders and Mental Retardation, which form Axis II. The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I; First *et al.*, 1997) is a semi-structured clinical interview to determine the patient's Axis I diagnoses based on DSM. The screener uses the SCID-I during the screening (step 1 in the procedure). The present study focuses on patients with a primary diagnosis (i.e. the most serious disorder) on Axis I.

The diagnoses on Axis I are divided into clusters. Each diagnosis belongs to one cluster. An example of a diagnosis is *Social Phobia* (coded 300.23) in the cluster *Anxiety Disorders*. Another example is *Primary Insomnia* (coded 307.42) in the cluster *Sleep Disorders*. The result of the screening with the SCID-I is a list of zero or more diagnoses on Axis I. If there are no diagnoses, there is no need for the available kind of treatment. If there are one or more diagnoses, the screener indicates the order of seriousness. The primary diagnosis is the target of treatment.

To determine the effect of treatment, there is a pre-assessment (step 4 in the procedure) and post-assessment (step 6 in the procedure) in which the patient completes three questionnaires. The questionnaires that are used in the assessments are as follows:

1. *Symptom Checklist-90* (SCL-90) (Arrindell & Ettema, 1986) – a series of 90 physical and psychological complaints, which the patient rates for the degree of distress associated with these complaints on a 5-point Likert scale. The sum score on all 90 items can be used as a global measure for the severity of Axis I psychopathology. The 90 items can be divided into nine groups of items, which belong to a specific complaint. There is no overlap between the items of a group. The sum score of these items forms the subscales. The subscales of the Dutch SCL-90 are agoraphobia, anxiety, depression, somatic complaints, inadequacy of thought and action, mistrust and interpersonal sensitivity, hostility, sleeping problems and 'other.'
2. *Fear Questionnaire* (Marks & Mathews, 1979; Zuuren, 1988) – a 32-item self-report measure. Each item is rated on a 9-point scale. The 32 items can by divided into eight groups of items, each group forming so-called the subscales. There is no overlap between the items of a subscale. The first four subscales refer to avoidance (Dutch abbreviation within the questionnaire: FQV) regarding the main phobia, social phobia, blood injury and agoraphobia. The last four subscales refer to anxiety (Dutch abbreviation within the questionnaire: FQA) regarding the main phobia, social phobia, blood injury and agoraphobia. The *main phobia* is formulated for each patient separately as the situation that the patient is most afraid of and requests treatment for.
3. *Biographical characteristics* – a list of questions concerning the patient's age, gender, religion, marital status, education level, use of medication, earlier treatment and duration of the complaint.

2.2. CBR in health care

Case-based reasoning is a technique for problem-solving that is based on the decision-making process of humans by learning from experiences (Kolodner, 1992; Aamodt & Plaza, 1994). The technique finds a solution for a problem by reusing the solutions of similar problems in the past. CBR can be used in domains where the knowledge is incomplete and complex. It has two major advantages: (1) *real-world examples* can be used as a knowledge base

(CBR does not require an explicit domain model), and (2) CBR can *explain* why it provides a solution by presenting similar cases of the past.

CBR is a popular AI-technique in the medical domain. CBR applications were built to enhance the work of health experts and to improve the efficiency and quality of health care (Holt *et al.*, 2006). Two of the earliest medical CBR systems for diagnosis and decision support were CASEY for heart failure (Koton, 1989) and MEDIC for dyspnoea (Holt *et al.*, 2006). More recent work in this area are systems for early detection of breast cancer (Hung & Chen, 2006), for diagnosing neuromuscular diseases (Pandey & Mishra, 2009a), diagnosing breast cytology (Ahn & Kim, 2009) and the system T-CARE, a temporal case retrieval system for medical scenarios used in an intensive care burn unit (Juarez *et al.*, 2011). CBR applications in health care are mostly used for determining diagnoses and as a medium for treatment (Pandey & Mishra, 2009b). In the mental health care, there are only a few CBR applications in use. The most famous system is SHRINK (Kolodner, 1983). It is one of the first CBR systems in the clinical field. SHRINK is also designed for determining the patient's diagnosis. More recent work in the area of mental health care is a CBR system for the diagnosis of attention-deficit hyperactivity disorder (ADHD) (Brien *et al.*, 2005).

Although CBR has not been used for predicting the effect of mental health treatment, CBR is successfully used for prediction in other areas. Some examples are predicting the success of an in vitro fertilisation treatment (Jurisica *et al.*, 1998), predicting the ecological risks of pesticides in freshwater ecosystems (Brink *et al.*, 2002) and predicting financial distress (Sun & Hui, 2006).

2.3. Techniques

Case-based reasoning is a methodology, which employs different techniques. The choice of which techniques are used in a CBR system depends on the contents and purpose of the system. In this section, CBR details and the techniques used for CBRth are discussed.

2.3.1. CBR cycle CBR seeks solutions to new problems by referencing a casebase with past experiences. The casebase forms the core of any CBR system. The CBR process is a cycle containing four steps (Aamodt & Plaza, 1994). The four steps are as follows:

1. RETRIEVE, which matches a new problem with the casebase and finds one or more similar cases;
2. REUSE, which uses the solutions for the similar cases found in the RETRIEVE step to suggest a solution for the new problem;
3. REVISE, which investigates whether the solution suggested by the REUSE step actually solves the problem after it is tried out; and

4. RETAIN, which stores the new problem and its solution in the casebase.

2.3.2. Forecasting-by-analogy If the solutions provided by a CBR system can be assigned to different classes, CBR may be used as a classifier system to predict to which class a target case belongs. This classification is called forecasting-by-analogy and contains three steps (Jo & Han, 1996; Li *et al.*, 2009):

1. identifying significant features to describe the target case;
2. searching for similar cases in the casebase; and
3. predicting the class of the target case based on the classes of the similar cases.

Compared to the CBR cycle, the first step concerns the case representation according to the structure of the casebase, the second step corresponds to RETRIEVE and the third step corresponds to REUSE. To complete the CBR cycle, the two steps REVISE and RETAIN can be added.

2.3.3. Nearest neighbour method For the search for similar cases (RETRIEVE), the nearest neighbour (NN) method is the most commonly used technique in CBR systems. NN is an exhaustive search method that evaluates the dissimilarity (or similarity) between all the past cases and the new case (Tsai *et al.*, 2005). There are several variants of the formula that is used in NN. The variant we used is shown as Formula 1 (Watson, 1999), where T is the target case, S is a case in the casebase and w_i is the weight of the ith feature. The expression $f_i(t_i, s_i)$ defines the similarity function for the similarity between T and S on the ith feature. Usually, the Euclidean distance is used for this equation, but different comparisons are possible.

$$Similarity(T, S) = \left(\sum_{i=1}^{n} f_i(t_i, s_i) \cdot w_i \right) / \left(\sum_{i=1}^{n} w_i \right) \quad (1)$$

The result of NN is one or more cases from the casebase, which are deemed 'most similar'. The two main variants of NN are K-nearest neighbour (KNN) and R-nearest neighbour (RNN). The result of KNN is a fixed number (K) of cases that have the minimum dissimilarity (or maximum similarity) with the new case. The result of RNN is a variable number of cases, namely, all cases that have a dissimilarity with the new case that is less than a threshold R (Tsai & Chiu, 2007).

2.3.4. Information gain In the NN method, weights are assigned to features. The information gain could be used as a value for feature weights. *Information gain* is term used in the information theory introduced by Shannon (1948). Quinlan (1986, 1989) used the gain in the ID3 and C4.5 methods to create decision trees. During the creation of

the decision tree, at each node of the tree, the information gain is calculated for every feature. The feature with the largest information gain is chosen as a node in the tree at that position in the tree. The information gain is a value that is easy to calculate, and the calculation is also easy to automate. Only little research was carried out in which the information gain is used as a feature weight. Daelemans *et al.* (1993), Ling *et al.* (1997) and Wettschereck and Dietterich (1995) used it in an instance based learning algorithm.

The information gain (IG) can be defined as

$$IG = (E \text{ before partitioning})\text{-}(E \text{ after partitioning}) \quad (2)$$

In this formula, E stands for Shannon Entropy. Entropy within the information theory is a measure to represent the uncertainty of a message as an information source (Munakata, 1998). Shannon Entropy E of variable $X(\omega)$ with a probability distribution $P(X(\omega) = x)$ is defined by (Munakata, 1998) as

$$E = -\sum_{all\ x} P(X(\omega) = x) \times \log_2 P(X(\omega) = x) \quad (3)$$

2.3.5. Weighted voting Sun and Hui (2006) described how the method *weighted voting* can be used by CBR in combination with KNN as a classifier. The result of KNN is a list of k cases that are most similar to the target case. The similarity of the ith of these k cases and the target cases is denoted as sim_i. The outcome class of an ith case is denoted as d_i. Every ith case belongs to an outcome class C_l (with $l = 1, 2, …, q$ in which q is the amount of possible outcome classes) when $C_l = d_i$. For every possible outcome, we can calculate the similarity weighted voting probability $prob(C_l)$, which indicates the probability that the target case belongs to class C_l, with Formula 2 based on Sun and Hui (2006):

$$prob(C_l) = \frac{\sum_{i=1}^{k} sim_i^*}{\sum_{i=1}^{k} sim_i} \quad (4)$$

$$\text{with } sim_i^* = \begin{cases} sim_i \text{ if } d_i = C_l \\ 0 \text{ otherwise} \end{cases}$$

We can use this weighted voting technique to make a prediction (classification) for the new case by choosing the class that has the highest similarity voting probability.

3. CBRth casebase

The CMHCM database allows to assess the effect of treatment by comparing pre and post-assessments. A large dataset has been collected, which is the basis for the casebase of CBRth. In this section, the structure of the casebase is discussed. We describe the case features (Section 3.1), the

outcome measure (Section 3.2), the selection of cases (Section 3.3) and the case format (Section 3.4).

3.1. Features

The basic principle of this study is to use all the features that are shared by all the cases. The data we used were collected during the pre-test. From these data, we extracted 30 different features. The features are named in the first column of Table 1. Section 2.1 discusses the meaning of each of these features. Because not every feature is equally important, we assigned a weight to each feature according to its importance. The different weights have a value in the range from 0 to 1. Determining the values of these weights is part of the study and is discussed in Section 4.3.

In this study, we used nominal (numerical values with no order in rank) and ordinal (a finite number of numerical values with an order in rank). In the second column of Table 1, the type of each feature is specified. The ordinal features have ranges of values. In the third column of Table 1, the range of each feature is specified. We refer to the values of these nominal and ordinal features as *scores*, which refer to the questionnaire scores.

Because some ranges of scores are quite large, we discretised the values into a smaller number of *categories*. For each score values, we assigned a category (also for the nominal types and ordinal types). In this way, every feature has a score value and a category value. We assigned categories as follows:

1. For the nominal types and the ordinal types with a small range, the category value is the same value as the score value.
2. For the features of the SCL-90, we used the classification of the norm table that is provided for the Dutch version (Arrindell & Ettema, 1986). They divided the scores of the features into a 7-point scale according to the following categories: very low, low, below average, average, above average, high and very high.
3. For the remaining features, we used the equal-frequency binning method (Witten & Frank, 2000; Liu *et al.*, 2002), an unsupervised method that divides a range into intervals with a predetermined number of cases per interval. To increase the liability of the equal-frequency binning, we used not only the cases that were included in the casebase (Section 3.3) but also cases of patients that received different treatments (1091 cases instead of 219).

In the fourth column of Table 1, the method for discretisation of the corresponding feature is indicated, with the arity (the number of categories after discretisation) listed in the fifth column.

3.2. Outcome measure

With CBRth, we want to predict whether a treatment will be successful or not. Therefore, there are two classes for the outcome measure: *successful* or *not successful*. To determine

Table 1: *Features used in case-based reasoning therapy*

Feature	Type	Range	Discretisation	Arity	Gain	a_i
SCL_agoraphobia	Ordinal	7–35	Norm SCL-90	7	0.027	0.241
SCL_anxiety	Ordinal	10–50	Norm SCL-90	7	0.041	0.171
SCL_depression	Ordinal	16–80	Norm SCL-90	7	0.084	0.108
SCL_somatic_complaints	Ordinal	12–60	Norm SCL-90	7	0.089	0.143
SCL_inadequacy	Ordinal	9–45	Norm SCL-90	7	0.071	0.189
SCL_sensitivity	Ordinal	18–90	Norm SCL-90	7	0.024	0.096
SCL_hostility	Ordinal	6–30	Norm SCL-90	7	0.042	0.280
SCL_sleeping_problems	Ordinal	3–15	Norm SCL-90	7	0.069	0.538
FQA_main	Ordinal	0–8	None		0.010	1.000
FQA_agoraphobia	Ordinal	0–40	Equal freq.	5	0.015	0.122
FQA_blood_injury	Ordinal	0–40	Equal freq.	5	0.043	0.122
FQA_social_fear	Ordinal	0–40	Equal freq.	5	0.027	0.122
FQV_main	Ordinal	0–8	None		0.015	1.000
FQV_agoraphobia	Ordinal	0–40	Equal freq.	5	0.040	0.122
FQV_blood_injury	Ordinal	0–40	Equal freq.	5	0.010	0.122
FQV_social_fear	Ordinal	0–40	Equal freq.	5	0.023	0.122
Gender	Nominal	1–2	None		0.003	1.000
Marital_status	Nominal	1–8	None		0.002	1.000
Religion	Nominal	1–5	None		0.017	1.000
Education_level	Ordinal	1–11	None		0.013	1.000
Age	Ordinal	16–72	Equal freq.	5	0.012	0.088
Medication	Nominal	0–1	None		0.000 *	1.000
Earlier_treatment	Nominal	0–1	None		0.008	1.000
Duration_complaint	Ordinal	0–600	Equal freq.	5	0.011	0.008
Cluster	Nominal	1–15	None		0.022	1.000
Diagnose	Nominal	1–36	None		0.115	1.000
Medication_category1	Nominal	0–1	None		0.001	1.000
Medication_category2	Nominal	0–1	None		0.001	1.000
Medication_category3	Nominal	0–1	None		0.002	1.000
Medication_category4	Nominal	0–1	None		0.007	1.000

*<0.000 but not equal to 0.

which of these classes a case belongs to, we used the c-criterion (Jacobson & Truax, 1991). The c-criterion is a cut-off point for clinically significant change. At the time of the post-assessment, a value is measured for the severity of the patient's disorder. In order for the patient to be classified as *free from the disorder*, this value should lie beneath the c-criterion.

Formula 5 (Jacobson & Truax, 1991) for the c-criterion is

$$c = \frac{s_0 M_1 + s_1 M_0}{s_0 + s_1} \qquad (5)$$

with M_0 and s_0 the mean and standard deviation of a well-functioning normal population, and M_1 and s_1 the mean and standard deviation of a clinical population.

The total sum score (*SCL-total*) on the SCL-90 is a global measure for the severity of Axis I psychopathology. We use the SCL-total as outcome attribute and calculated the c-criterion for this attribute. Arrindell and Ettema (1986) report the results of the SCL-90 for a large group of a well-functioning normal population (referred to as 0 in Formula 5) and the result of the SCL-90 for a large group of patients (referred to as 1 in Formula 5). These results were officially used for compiling the norms for the SCL-90. We used these results to calculate the value of c. Because Arrindell and Ettema (1986) make a distinction between the

male-population and the female-population, we have different values for *c* for men (141.4) and women (159.1).

We used the value of *c* as follows. If for a patient the *SCL-total* of the post-assessment was lower than the *c*, the treatment was deemed *successful*, otherwise *not successful*. For example, if the *SCL-total* of the post-assessment of a woman was 150, then treatment was deemed *successful* because $150 < 159.1$.

In this study, we used the c-criterion (as described previously) for the definition of success. This method focuses on the end state functioning of the patient (i.e. the *recovery*) after treatment. Other definitions of success are possible, for example, the Reliable Change Index (Jacobson & Truax, 1991), which focuses on the *change* of the patient between the state before treatment and after treatment.

3.3. Cases

The cases we used in the casebase are the *real-world examples* of the large dataset of pre and post-assessments. The selection of cases were based on the following principles:

- Only cases in which the treatment was cognitive behavioural therapy were included.
- Only cases that completed the treatment were included.

- Only cases of which all the features that we selected for this study were available were included. We left out the concept of missing values (these are the focus of follow-up research).
- Only cases of which the *SCL-total* in the post-assessment was completed were included. We need this value to determine the outcome of the treatment.
- Only cases of which the *SCL-total* in the pre-assessment is higher than the *c* were included. Is already lower than *c* than we cannot define the effect of the treatment in terms of the c-criterion.

This resulted in a dataset of 219 cases. According to our criteria, for 98 cases, the treatment was successful, and for 121 cases, the treatment was not successful.

3.4. Case format

A typical case in a CBR system consists of three parts (Watson & Marir, 1994):

- the *problem* that describes the state of the world when the case occurred;
- the *solution* that states the derived solution to that problem; and
- the *outcome* that describes the state of the world after the case occurred.

If CBR is used for prediction, the *solution* is not a part of the case format. The *outcome* of the new case is based on the outcome of the similar cases that were retrieved based on the *problem*.

In each case in our casebase, the *problem* describes the state of the patient before treatment starts. The problem description consists of values (the score- and the category values) of all features for that case. A full description of the case features is given in Table 1. For the outcome, there are only two possible values: 0 and 1. The value 0 means that the cognitive behavioural therapy was *not successful*, and the value 1 means that it was *successful*.

4. CBRth process

For CBRth, we use CBR to predict in which class a new target case falls. We classify according to forecasting-by-analogy (Section 2.3.2). In this section, we describe the process steps that a target case follows within CBRth. These are the tree steps of forecasting-by-analogy. The first step is already described in Section 3.3. Sections 4.1 and 4.2 describe the other two steps. Forecasting-by-analogy only covers the first two steps of the CBR cycle (RETRIEVE and REUSE). In Section 4.3, we describe the remaining two steps (REVISE and RETAIN) of the CBR cycle.

4.1. Search for similar cases (RETRIEVE)

For the search for similar cases, we use Formula 1 for the NN method (Section 2.3.3). In this formula, two

components can be customised, namely, the weight w_i of every ith feature and the equation for the similarity function between T and S $f_i(t_i, s_i)$ for every ith feature. How we defined w_i is discussed in Section 4.3. For the similarity function, we used two different methods, the *category method* and the 'score method'. For each method, we built a variant of CBRth to investigate which of the two methods provides better results.

The *category method* is the simplest of both methods. It uses the category values of the features. The possible output of the similarity function only includes 0 and 1. The output 1 is given when the corresponding categories of the feature that are compared are equal for both cases, and the output 0 is given when they are not equal for both cases.

$$f_i(t_i, s_i) = \begin{cases} 1 \text{ if } t_i = s_i \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

The *score method* uses the score values.

$$f_i(t_i, s_i) = \begin{cases} 1 - a_i \cdot |t_i - s_i| \text{ if } |t_i - s_i| < 1 \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

In Formula 7, the parameter a_i can have a different value for each feature i (for $a_i = 1$, Formulas 6 and 7 are equal). The effect of a_i is that *near* values obtain a high degree of similarity, and *distant* values obtain a low degree of similarity. Because in nominal features, there is no order in the values, $a_i = 1$ for nominal values. For the ordinal features with a small range, we also set $a_i = 1$. For the ordinal features with a larger range, a_i is a value between 0 and 1 calculated by Formula 8. The used values for a_i are given in the seventh column of Table 1. In the third column (range) of Table 1, the minimum and maximum values are given.

$$a_i = (\textit{number of classes feature i})/ \\ (\textit{maximum value feature i} - \textit{minimum value feature i} + 1) \tag{8}$$

4.2. Predicting the outcome of the target case (REUSE)

For predicting the outcome of the target case, we used weighted voting (as explained in Section 2.3.5). There are two possible values for the outcome class C_l, namely, *successful* and *not successful*. For the target case, $prob(C_l)$ is calculated according to Formula 2 for both the possible values of the C_l. The prediction for the target case is the value of C_l for with $prob(C_l) > 0.5$.

Sun and Hui (2006) base weighted voting on KNN. We applied a more fine-grained approach by using by using a combination of RNN and KNN, in which we limit the maximum number of selected cases (K), and on top exclude the cases that are too dissimilar (R). This means k in Formula 2 is not a fixed value, but k is a value between 0 and K.

4.3. Remaining steps of the CBR cycle (REVISE and RETAIN)

After a patient received and completed cognitive behavioural therapy, we can determine the actual outcome, that is, whether the treatment was successful or not, based on the results of the post-assessment. In the REVISE step, we compare the predicted outcome with the actual outcome. We enter the value of the actual outcome as *outcome* in the case. The knowledge whether a prediction was correct or not is useful for an indication of CBRth's accuracy.

In the RETAIN step, the target case is added to the casebase. Changing the composition of the casebase can also provide insight into which features are most important in determining the best match. Particular values of the weights may evolve over time as the casebase expands. This means that tuning the features weights (w_i) can also positively influence the future performance of a CBR system. If the feature weights must be tuned after each time a case is added, it is strongly preferable that they are easy to calculate.

In the present study, we researched whether the information gain can be used as a feature weight. For every feature, we use the information gain that is calculated to decide which feature should be in the top note of the decision tree. We built a variant of CBRth in which the feature weights are all equal ($w_i = 1$ for every i) and a variant with the information gain for every feature. We compared both variants to investigate which of the two provides better results. Note that we only want to investigate whether gain works as feature weight, not whether it is the most suitable feature weight; at present, our dataset is simply not large enough to draw conclusions on which kind of feature weight works best.

5. Experimental setup

The introduction describes the goal of this study as twofold: (1) to investigate which techniques are suitable for implementing CBRth and (2) to investigate the accuracy of CBRth. Section 4 describes the system design of CBRth and the steps that are successively followed to use CBRth for making predictions on the success of treatments. To test CBRth, we used 219 cases (Section 3.3) to investigate both parts of our goal. In our approach, we used combinations of a training set of 218 cases and a test set of 1 case. All possible 219 combinations of training set and test set were used.

The procedure that we followed consists of the following eight steps:

1. Place the 219 cases in the casebase of CBRth.
2. Select one case and remove this case from the casebase.
3. Recalculate the weights corresponding to the casebase with the remaining 218 cases (RETAIN).
4. Determine the best match of the resulting casebase for the selected case (RETRIEVE).
5. Predict the outcome for the selected case (REUSE).
6. Check whether the predicted outcome is correct (REVISE) and log the results.
7. Repeat steps 2–6 until all cases for all 219 combinations of training set and test set.
8. Evaluate the accuracy of CBRth.

Our approach is based on 219-fold cross validation, to be exact *leave-one-out cross validation* (Witten & Frank, 2000). We used this approach for testing both goals of the study. The results of this experimental setup are presented in the next section.

6. Results

This section describes our results in two sections corresponding to the two parts of our goal. The results are shown in Figures 1–5. For more detailed results, see Janssen (2009).

6.1. Best combination of techniques

The first part of the goal is to investigate which techniques are suitable for implementing CBRth. We investigated whether the information gain is a useful value for feature weights, and which of the methods described in Section 4.1 is best suited for feature matching. To determine the best combination of techniques, the following parts were investigated in succession:
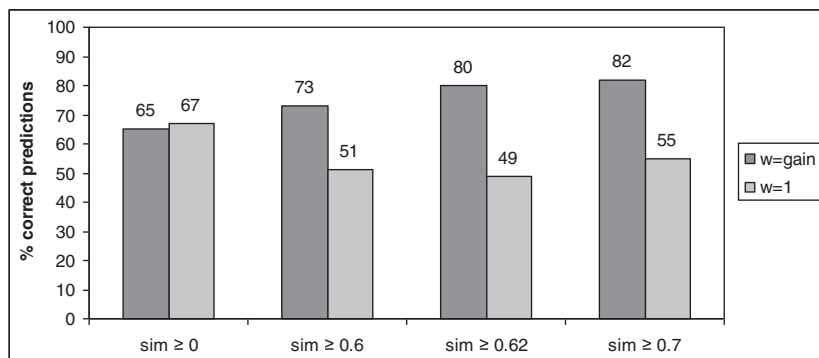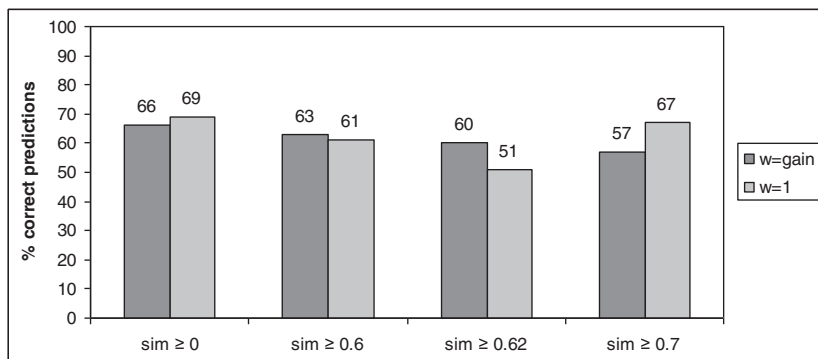

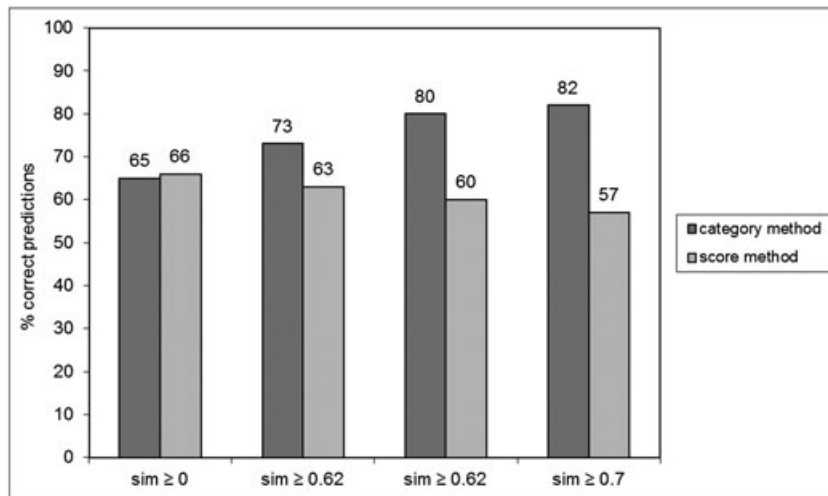**Figure 1:** *Category method.*

Figure 2: *Score method.*

Figure 3: *Category method and score method comparison.*
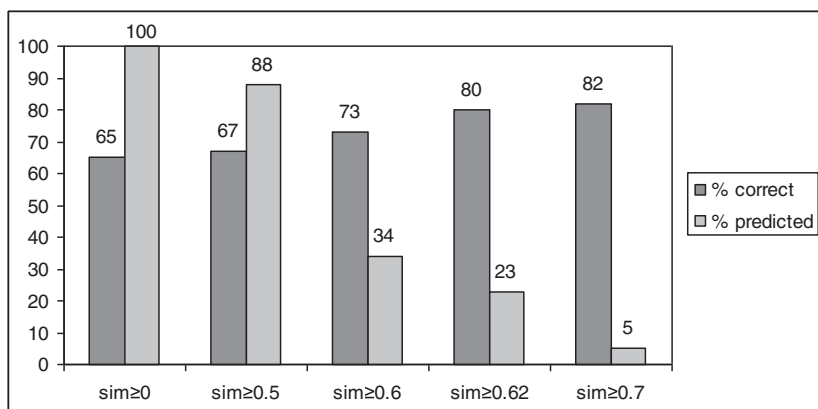

Figure 4: *Results case-based reasoning therapy with K = max 1.*

1. feature weights and
2. feature matching.

6.1.1. Feature weights We started by investigating whether to use the gain as feature weight or not. For this decision, we compared the results with the gain as a feature weight to the results without a feature weight.

We did that for both the category method and the score method.

The results for the category method are shown in Figure 1. In this figure, the results with the gain as feature weight are shown as $w = gain$, and the results without a feature weight are shown as $w = 1$. We show the results for four different values of R in the RNN method, namely, sim ≥ 0 ($R = 1$), sim ≥ 0.6 ($R = 0.4$), sim ≥ 0.62 ($R = 0.38$) and sim ≥ 0.7
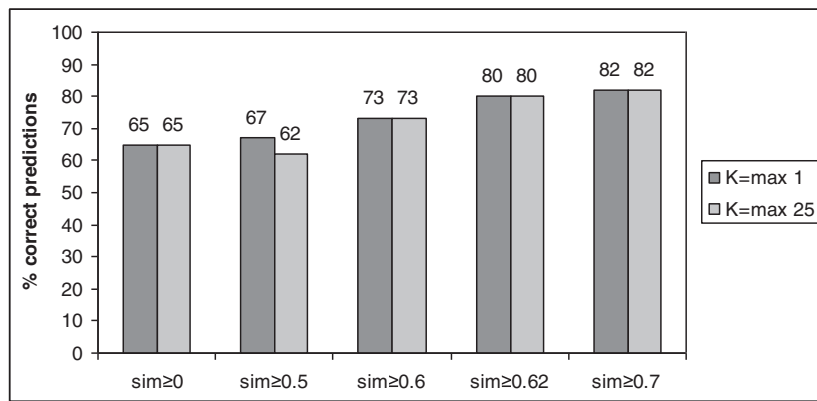
Figure 5: *case-based reasoning therapy with K = max 1 and K = max 25.*

($R = 0.3$). We combined RNN with KNN and the choice for $K = $ max 25.

A prediction can be made when there is at least one match in the casebase that achieves the minimum similarity R. The *% correct predictions* is the percentage predictions that were correct for all the predictions that could be made. When $sim \geq 0$, all the cases that enter the CBR cycle in the casebase meet the restriction. The prediction is thus based on the 25 best matches. When $sim \geq 0.62$ for only 23% of the cases that entered the CBR cycle, a prediction could be made. For 80% of these, the prediction was correct.

For the category method, when $sim \geq 0$, the *% correct predictions* is higher for $w = 1$ (67%) than for $w = gain$ (65%), but the difference between the two rates is only 2%. Once the restriction of the minimum similarity is set higher, the differences between $w = gain$ and $w = 1$ are greater. Here, CBRth with the gain as feature weight gives better results.

In Figure 2, the results for the score method are shown. The results are also shown for $w = gain$ and $w = 1$, and for the same values of R and K.

When $sim \geq 0$, the *% correct predictions* is marginally higher for $w = 1$ (69%) than for $w = gain$ (66%). When $sim \geq 0.6$ and $sim \geq 0.62$, the results for $w = gain$ are higher than the results of $w = 1$ (63% and 61% vs 60% and 51%). For $sim \geq 0.6$, the difference is minimal (2%), but for $sim \geq 0.62$, the difference is 9%. The difference is also relatively large when $sim \geq 0.7$ (57% vs 67%); however, when $sim \geq 0.7$ for only a very small number of cases that enter the cycle, a prediction could be made (3% of $w = gain$ and 5% for $w = 1$).

Based on the results in this section, we decided to use the gain as feature weight.

6.1.2. Feature matching After deciding on the feature weight estimation, we investigated whether to use the category method or the score method for feature matching. The results for the category method and the score method with the gain as feature weight are shown in Figure 3. We show the results for four different values of R in the RNN method, namely, $sim \geq 0$ ($R = 1$), $sim \geq 0.6$ ($R = 0.4$), $sim \geq 0.62$ ($R = 0.38$) and $sim \geq 0.7$ ($R = 0.3$). We combined RNN with KNN and the choice for $K = $ max 25.

When $sim \geq 0$, the *% correct predictions* is marginally higher for the *score method* (66%) than for the *category method* (65%), the difference is negligible. With $sim \geq 0.6$, $sim \geq 0.62$ and $sim \geq 0.7$, the results of the *category method* are higher than the results of the *score method* (73%, 80% and 82% vs 63%, 60% and 57%, respectively). We conclude that the category method gives better results than the score method.

One of the characteristics of the CBR method is that the percentage of correct predictions improves as the restriction on similarity increases, because the prediction is only made if there are matches found that have a certain degree of similarity with the new case. If we examine the results in Figure 3, we can conclude that this is indeed the case for the category method. The score method shows the opposite results, namely, that there are fewer correct predictions when the degree of similarity increases. That is surprising, but the differences are small.

Based on the results in this section and the previous section, we decided for CBRth to use the category method with the gain as a feature weight.

6.2. CBRth accuracy with available data

We now present CBRth's accuracy with the techniques we chose in the previous section: the category method with the gain as a feature weight. This study was performed with the 219 cases detailed in Section 3.3. Of these 219 cases, the treatment was *successful* for 98 cases (45%) and *not successful* for 121 cases (55%). As the prediction *never successful* is correct in 55% of the cases, 55% is the frequency baseline. To perform well, CBRth must make predictions with an accuracy significantly higher than 55%.

Figure 4 shows the results for the application of CBRth to the available data. These are the results for five different values of R in the RNN method, namely, $sim \geq 0$ ($R = 1$), $sim \geq 0.5$ ($R = 0.5$), $sim \geq 0.6$ ($R = 0.4$), $sim \geq 0.62$ ($R = 0.38$) and $sim \geq 0.7$ ($R = 0.3$). In this figure, the results are shown for which we combined RNN with KNN with $K = $ max 1 (results for $K = $ max 25 are shown in Figure 5, which is shown later). The figure shows two bars. The first bar shows the percentage correct predictions of the predictions that

could be made (i.e. at least one case was selected as a *best match* from the casebase), and the second bar shows for how many cases a prediction could be made.

Obviously, with sim ≥ 0 for all the cases, a prediction could be made. The percentage correct predictions is 65%, that is, significantly exceeding the frequency baseline. When the restriction on similarity is higher, we may expect the number of cases for which a match can be found to decrease. However, we may also expect an increase in the percentage of correct predictions. As can be seen in Figure 4, the results show this pattern. For example, for sim ≥ 0.5 for 88% of the cases, a prediction is made, which is correct for 67% of them. For sim ≥ 0.62, a prediction is made for only 23% of the cases, but it is correct for 80% of them.

Figure 5 shows that the best results were achieved with K = max 1. However, the results for K = max 25 are almost equal. As the percentage of cases for which a prediction is made only depends on the value of R, it is the same for both values of K. The percentage of correct predictions for both values of K is compared in Figure 5. We only see a small drop in the percentage for sim ≥ 0.5. A straightforward explanation for the fact that for sim ≥ 0.6 the results are the same for both values of K is that for the higher restrictions on R for the majority of cases, only one match is found. Naturally, when the number of cases in the casebase increases, the number of matches may increase too. The effect of that on the accuracy of the predictions has not been tested yet, but it may be expected to improve the accuracy as better and more matches can be found.

Tables 2 and 3 show the results for CBRth with K = max 1 for sim ≥ 0 and sim ≥ 0.62. In these tables, all the cases for which a prediction could be made are given in four quadrants depending on the outcome of the prediction and the actual outcome. Again, it shows that fewer predictions can be made for a higher restriction on similarity. With the figures in the tables, we can calculate the sensitivity (Formula 9), the specificity (Formula 10) and the likelihood ratio (Formula 11) (based on Woodward (2005))

$$sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives} \quad (9)$$

$$specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives} \quad (10)$$

Table 3: *Results for CBRth with sim ≥ 0.62 and K = max 1*

	Actual outcome		
Prediction	Success	No success	Total
Success	16	6	22
No success	4	24	28
Total	20	30	50

$$likelihood\ ratio = \frac{sensitivity}{1 - specificity} \quad (11)$$

in which positive refers to *successful* and negative refers to *not successful*. The sensitivity, specificity and likelihood ratio for CBRth with K = max 1 is shown in Table 4 for five measures of similarity. The table shows that for a higher restriction on similarity, both sensitivity and specificity increase. This naturally increases the likelihood ratio.

7. Discussion

The first goal of this study was to investigate which techniques are suitable for implementing CBRth. With respect to this goal, we found that information gain is suitable as a feature weight. The gain is easy to calculate. Therefore, the gain can be recalculated every time a case is added to the casebase in the RETAIN step. This can be carried out without the intervention of a developer, so a part of the maintenance of the system is automated. When in time it is discovered that some features increase in importance for the success of the treatment, the system will adjust itself automatically.

The use of the gain as a feature weight allows CBRth to use all the features that are available. It is not necessary for an expert to make a selection first. However, we included only those features for which all the values for all cases were available. CBRth does not deal with missing values yet. It is therefore possible that we have left out important features. In future research, we will incorporate the handling of missing values.

We developed two methods to compare features. The category method handles strict boundaries between the values of the features. The score method uses a smooth

Table 2: *Results for CBRth with sim ≥ 0 and K = max 1*

	Actual outcome		
Prediction	Success	No success	Total
Success	64	43	107
No success	34	78	112
Total	98	121	219

Table 4: *Sensitivity, specificity and likelihood ratio for CBRth with K = max 1*

	Sensitivity	Specificity	Likelihood ratio
sim ≥ 0	0.65	0.64	1.84
sim ≥ 0.5	0.69	0.65	1.96
sim ≥ 0.6	0.72	0.74	2.74
sim ≥ 0.62	0.80	0.80	4.00
sim ≥ 0.7	0.83	0.80	4.17

change. We expected that the score method would perform better, because of its smooth character. To our surprise, the category method performed better. One possible explanation is that the score method makes similarity values drop rapidly when the feature values differ slightly. For the category method, the similarity is maximal when the features are part of the same category even if they differ slightly in feature values.

With a constraint on the similarity, CBRth can only predict an outcome when there is at least one match that meets the condition. With $sim \geq 0.62$ for only 23% of the cases, a prediction can be made. The reliability of this prediction is higher than without a restriction, namely, a correct prediction is made for 80% of the cases. When used in practice, the casebase will grow automatically. This means that the number of cases for which a prediction can be made with $sim \geq 0.62$ will increase. As the 80% correct predictions is mainly tied to the value of R, we may expect that for this higher percentage of predictions, the accuracy will remain at least 80%, and may even increase when the predictions are based on more similar cases.

The CBRth can give a prediction very fast. The computational cost for making a prediction depends on the number of cases and the number of features. Because the number of features is constant, the computational costs only depend on the number of cases; this means that computation costs are O(n), n being the number of cases. When adding new cases, the casebase, the information gain of all the features has to be recalculated. The computational costs of that are, again, O(n).

The CBRth makes binary predictions: the only two possible outcome values are successful and not successful. We have not looked at the degree of improvement. In practice, it is possible that a patient is not cured after the treatment, but has significantly improved and thus benefits from the treatment. In future research, we will focus on adding a second outcome measure, namely, the degree of improvement.

This study focuses on cognitive behavioural therapy for patients with an anxiety disorder. In practice, there are more treatment options for this group of patients. For each treatment option, we can use CBRth with a different casebase specific for that treatment. When the features of a new patient are entered, CBRth can make a prediction for each treatment option, and a final decision on treatment offered can be based on the predictions for each form of treatment.

Because of the transparency and the interpretability of CBR, we expect that therapists will accept CBRth for decision support. Offering treatments, however, remains a human activity, and it will be a long time before CBRth grows beyond a purely advisory system.

8. Conclusions

We were able to build a CBR system (CBRth) for prediction the success of cognitive behavioural therapy for a patient before start of the treatment. The best techniques for CBRth

turned out to be (1) the use of the information gain for feature weighting as described in Section 2.3.4 and 4.3 and (2) the use of the category method as described in Section 4.1. For the effectiveness of CBRth, we showed that without restriction on the similarity, 65% of the predictions of CBRth were correct, which is considerably higher than the frequency baseline of 55%. With restrictions on the similarity, CBRth becomes even more reliable: with $sim \geq 0.62$, 80% of the predictions were correct, but for only 23% of the cases a prediction could be made. Future studies are necessary to assess the effectiveness of CBR compared to the classical method in the mental healthcare research (like regression) and the classical methods in the artificial intelligence (like neural networks or support vector machines).

The CBRth was demonstrated to provide useful advice in relation to the treatment outcome of cognitive behavioural therapy. Future studies should incorporate different diagnostic groups and different treatment modalities. Of special interest is the use of CBR in deciding which treatment modality is indicated for a specific case, for example, pharmacotherapy versus cognitive behavioural therapy. In future research, we will also compare the predictions of CBRth with predictions an intake staff does before the start of treatment. If the CBRth predictions have a higher accuracy than those of the intake staff, it is definitely suitable as an advisory system. As CBRth even with $sim \geq 0$ has an accuracy of 65%, the results of this study provide sufficient confidence in CBR as a prediction model in the mental health area. With restrictions on the similarity, CBRth reaches an accuracy of 80%. Although then only for a small part of the patients this prediction can be carried out, these are very good results. The casebase grows during use, and the larger the casebase becomes, the more chance on finding a good match. The percentage of reliable (with at least $sim \geq 0.62$) predictions will only increase. It has therefore been decided that CBRth will see further development, with the ultimate goal to incorporate it in the daily practice of treatment assignments.

References

AAMODT, A. and E. PLAZA (1994) Case-based reasoning: foundational issues, methodological variations, and system approaches, *AI Communications*, **7**, 39–59.

AHN, H. and K.-J. KIM (2009) Global optimization of case-based reasoning for breast cytology diagnosis, *Expert Systems with Applications*, **36**, 724–734.

ARRINDELL, W.A. and J.H.M. ETTEMA (1986) SCL-90. Handleiding bij een multidimensionele psychopathologie-indicator, Lisse: Swets & Zeitlinger.

American Psychiatric Association (1994) Diagnostic and Statistical Manual of Mental Disorders, Washington, DC: American Psychiatric Association.

BRIEN, D., J. GLASGOW, D. MUNOZ, H. MUÑOZ-ÁVILA and F. RICCI (2005) The Application of a Case-Based Reasoning System to Attention-Deficit Hyperactivity Disorder Case-Based Reasoning Research and Development, Berlin/Heidelberg: Springer.

BRINK, P.J.V.D., J. ROELSMA, E.H.V. NES, M. SCHEFFER and T.C.M. BROCK (2002) PERPEST model, a case-based reasoning approach to predict ecological risks of pesticides, *Environmental Toxicology and Chemistry*, **21**, 2500–2506.

DAELEMANS, W., S. GILLIS, G. DURIEUX and A.V.D. BOSCH (1993) Learnability and markedness in data-driven acquisition of stress. ITK Research Report. Institute for Language Technology and Artificial Intelligence.

FIRST, M.B., R.L. SPITZER, M. GIBBON and J.B.W. WILLIAMS (1997) Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I/P), New York: New York State Psychiatric Institute, Biometrics Research Department.

HOLT, A., I. BICHIDARITZ, R. SCHMIDT and P. PENER (2006) Medical applications in case-based reasoning, *The Knowledge Engineering Review*, **20**, 289–292.

HUNG, S.-Y. and C.-Y. CHEN (2006) Mammographic case base applied for supporting image diagnosis of breast lesion, *Expert Systems with Applications*, **2006**, 93–108.

JACOBSON, N.S. and P. TRUAX (1991) Clinical significance: a statistical approach to defining meaningful change in psychotherapy research, *Journal of Consulting and Clinical Psychology*, **59**, 12–19.

JANSSEN, R. (2009) Case-based reasoning voor het voorspellen van succes van therapie. faculteit Informatica, Heerlen: Open Universiteit Nederland.

JO, H. and I. HAN (1996) Integration of case-based forcasting, neural network, and discriminant analysis for bankruptcy prediction, *Expert Systems with Applications*, **11**, 415–422.

JUAREZ, J.M., M. CAMPOS, J. PALMA and R. MARIN (2011) T-CARE: temporal case retrieval system, *Expert Systems*, **28**, 324–338.

JURISICA, I., J. MYLOPOULOS, J. GLASGOW, H. SHAPIRO and R.F. CASPER (1998) Case-based reasoning in IVF: prediction and knowledge mining, *Artificial Intelligence in Medicine*, **12**, 1–24.

KOLODNER, J.L. (1983) Towards an understanding of the role of experience in the evolution from novice to expert. *International Journal of Man-machine Studies*, **19**, 497–518.

KOLODNER, J.L. (1992) An introduction to case-based reasoning, *Artificial Intelligence Review*, **6**, 3–34.

KOTON, P. (1989) A medical reasoning program that improves with experience, *Computer Methods and Programs in Biomedicine*, **30**, 177–184.

LI, H., J. SUN and B.-L. SUN (2009) Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors, *Expert Systems with Applications*, **36**, 643–659.

LING, C.X., J.J. PARRY and H. WANG (1997) Deciding Weights for Nearest Neighbour Algorithms using Decision Trees, Ontario: The University of Western Ontario.

LIU, H., F. HUSSAIN, C.L. TAN and M. DASH (2002) Discretization: an enabling technique, *Data Mining and Knowlege Discovery*, **6**, 393–423.

MARKS, I.M. and A.M. MATHEWS (1979) Case histories and shorter communications, *Behavior Research and Therapy*, **17**, 263–267.

MUNAKATA, T. (1998) Fundamentals of the New Artificial Intelligence: Beyond Traditional Paradigms, New York: Springer-Verlag.

PANDEY, B. and R.B. MISHRA (2009a) An integrated intelligent computing model for the interpretation of EMG based neuromuscular diseases, *Expert Systems with Applications*, **36**, 9201–9213.

PANDEY, B. and R.B. MISHRA (2009b) Knowledge and intelligent computing system in medicine, *Computers in Biology and Medicine*, **39**, 215–230.

QUINLAN, J.R. (1986) Induction of decision trees, *Machine Learning*, **1**, 81–106.

QUINLAN, J.R. (1989) Unknown attribute values in induction, in *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, NY: Morgan Kaufmann.

SHANNON, C.E. (1948) A mathematical theory of communication, *The Bell System Technical Journal*, **27**, 379–423.

SPINHOVEN, P., J. GIESEN-BLOO, R.V. DYCK and A. ARNTZ (2008) Can assessors and therapists predict the outcoume of long-term psychotherapy in Borderline Personlity disorder? *Journal of Clinical Psychology*, **64**, 667–686.

SUN, J. and X.-F. HUI (2006) Financial Distress Prediction Based on Similarity Weighted Voting CBR. Advanced Data Mining and Applications, Berlin/Heidelberg: Springer.

TSAI, C.-Y. and C.-C. CHIU (2007) A case-based reasoning system for PCB principal process parameter identification, *Expert Systems with Applications*, **32**, 1183–1193.

TSAI, C.-Y., C.-C. CHIU and J.-S. CHEN (2005) A case-based reasoning system for PCB defect prediction, *Expert Systems with Applications*, **28**, 813–822.

WATSON, I. (1999) Case-based reasoning is a methodology not a technology, *Knowledge-Based Systems*, **12**, 303–308.

WATSON, I. and F. MARIR (1994) Case-based reasoning: a review, *The Knowledge Engineering Review*, **9**, 355–381.

WETTSCHERECK, D. and T.G. DIETTERICH (1995) An experimental comparison of the nearest neighbor and nearest-hyperrectangle algoritms, *Machine Learning*, **19**, 5–27.

WITTEN, I. and E. FRANK (2000) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation, San Francisco: Margan Kaufmann Publishers.

WOODWARD, M. (2005) Epidemiology: Study Design and Data Analysis, Boca Raton: Chapman & Hall/CRC.

ZUUREN, F.J.V. (1988) The fear questionnaire: some data on validity, reliability and layout, *British Journal of Psychiatry*, **153**, 659–662.

The authors

Rosanne Janssen

Rosanne Janssen is a PhD candidate at the Faculty of Psychology & Neuroscience at the University of Maastricht, the Netherlands. She received the BSc degree and MSc degree with honours in computer science at Open University, the Netherlands. Her main research interests include health informatics and artificial intelligence techniques such as case-based reasoning.

Pieter Spronck

Pieter Spronck is an Associate Professor in Artificial Intelligence at the University of Tilburg, the Netherlands. He authored and co-authored about 100 papers on artificial intelligence and machine learning in international journals and conference proceedings. At present, his research focuses on player modelling, data mining and adaptive intelligence.

Arnoud Arntz

Arnoud Arntz is Professor of Clinical Psychology at the University of Amsterdam, the Netherlands. His main research interests lie in the fields of anxiety and personality disorders, both applied and fundamental. He is director of various multisite studies investigating new psychological treatments of personality disorders and post-traumatic stress disorder. He also practises as a psychotherapist at the Viersprong Institute in Amsterdam, where he mainly treats patients with personality disorders.