Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data

Marco Zaffalon, [a,1] Keith Wesnes [b] and Orlando Petrini [c]

[a] *IDSIA, Galleria 2, 6928 Manno (Lugano), Switzerland*
[b] *CDR Ltd., CDR House, 24 Portman Road, Reading RG30 1EA, UK*
[c] *Pharmaton SA, Via ai Mulini, 6934 Bioggio, Switzerland*

Abstract

Dementia is a serious personal, medical and social problem. Recent research indicates early and accurate diagnoses as the key to effectively cope with it. No definitive cure is available but in some cases when the impairment is still mild the disease can be contained. This paper describes a diagnostic tool that jointly uses the naive credal classifier and the most widely used computerized system of cognitive tests in dementia research, the Cognitive Drug Research system. The naive credal classifier extends the discrete naive Bayes classifier to imprecise probabilities. The naive credal classifier models both prior ignorance and ignorance about the likelihood by sets of probability distributions. This is a new way to deal with small or incomplete data sets that departs significantly from most established classification methods. In the empirical study presented here the naive credal classifier provides reliability and unmatched predictive performance. It delivers up to 95% correct predictions while being very robust with respect to the partial ignorance due to the largely incomplete data. The diagnostic tool also proves to be very effective in discriminating between Alzheimer's disease and Dementia with Lewy Bodies.

Key words: Credal classification; Dementia; Cognitive tests; Naive credal classifier; Imprecise probabilities; Missing data; Data mining

[1] Corresponding author. Tel.: +41 91 610 8665; fax: +41 91 610 8661.
Email addresses: zaffalon@idsia.ch (M. Zaffalon), keithw@cdr.org.uk (K. Wesnes), petrini@lgn.boehringer-ingelheim.com (O. Petrini).

1 Introduction

Dementia is one of the leading causes for concern in the elderly. On the personal level it reduces the quality of life. Socially, it is one of the causes of major costs in health care for the elderly population, severely demented patients requiring constant supervision and medical care [2,23]. The most important and widely known type of dementia is Alzheimer's Disease (AD), which accounts for approximately 50% of all types of diagnosed dementia. Vascular Dementia (VD) has traditionally been considered to be the second most common cause of dementia (up to 20% of all dementias, either alone, or in combination with AD). However, in recent years Dementia with Lewy Bodies (DLB) has been identified as the second most common single form of dementia, accounting for over 20% of all dementias [23]. DLB has previously been mistakenly diagnosed as Alzheimer's disease and sometimes has been confused with schizophrenia. One of the major problems in the performance of clinical trials with DLB is the correct diagnosis of the disorder [23,36–38].

There is currently no cure for dementia, although galanthamine, rivastigmine and donepezil (all acetylcholinesterase inhibitors) have now been registered in several countries for the mild symptomatic relieve of AD. An extract of Ginkgo biloba has also shown some activity in the treatment of symptoms of dementia [14]. Current research is trying to provide an early diagnosis of AD, as there is some hope that these compounds may prove more effective in treating early stages of dementia, also often termed "mild cognitive impairment", and in preventing rather than reducing symptoms [41]. A first clinical trial has been completed and shows that rivastigmine can dramatically improve cognitive functions in DLB [22].

Several problems confront research in this field. First, not all of the available systems are sufficiently sensitive to detect the early stages of dementia. Secondly, it remains to be confirmed which tests may differentiate between different types of dementia.

The present paper addresses these two problems by coupling the power of emerging classification tools and the diagnostic capabilities of a well-targeted system of cognitive tests. We propose an automated diagnostic model that deals successfully with both the sensitivity of the methods and their differential properties.

The Cognitive Drug Research (CDR) computerized assessment system has been chosen for this study. This system has been designed to provide a valid, reliable and sensitive tool to assess cognitive functions in dementia [25,27,33,36–38,42]. The system is the most widely used automated system in dementia research [25] (see Sect. 2.1). We have used a database describing the actual health state and the past responses to the CDR system tests for about 3,400 patients (Sect. 2.4). Data were not collected with the specific purpose of statistical analysis, so they present a substantial amount of missing values. Missing data are a fundamental problem

for machine learning methods; treating them properly is essential to draw reliable conclusions.

To overcome these challenging issues we chose the classification model called *Naive Credal Classifier* [45,47] (NCC, Sect. 2.3). The NCC generalizes the well-known discrete *Naive Bayes Classifier* (or NBC [4]) to *imprecise probabilities* [39], a well-founded generalized framework for uncertain reasoning. The NCC models both prior ignorance and ignorance about the likelihood originated by missing data, by sets of probability distributions (see Sect. 2.2). This makes the NCC one of the most significant steps towards reliability and realism in classification. We can find a similar modelling of incomplete samples in Ramoni and Sebastiani's *Robust Bayes Classifier* (RoC) [29] (see Sect. 2.3 for a detailed comparison). When small or incomplete samples convey only scarce knowledge on a domain, the NCC maintains reliability by providing possible partially indeterminate classifications, i.e. more than one class for a given object. For example, in the present application, the NCC may map a patient to the set $\{\mathrm{AD},\mathrm{VD}\}$, which means that it is not possible to discriminate between these two diseases; all the others, however, can be discarded. In other words, the NCC transforms the imprecision detected in the data into output indeterminacy, without trying to reduce it any further by doing strong assumptions.

The characteristics of the new paradigm of credal classification, in fact, enable the NCC to make automatically reliable diagnoses of dementia. This is the first application of credal classification to dementia screening. We report a detailed empirical study, in Sect. 3, that analyzes the predictive behavior of the NCC on the data. (The reader interested in the results from the clinical viewpoint can find a summary in the concluding section of the paper.) The diagnostic accuracy described in past work on dementia [19] is thereby improved and we obtain up to 95% correct predictions. We also show that the system is very effective in discriminating among dementias, even between the two types that are currently only hardly distinguishable, AD and DLB. The study also compares the NCC with the NBC and the RoC. In comparison with the NCC, the predictions of the NBC appear to be unreliable for a large portion of the data. The NCC compares well with the RoC, and it avoids the overly caution exhibited by the RoC in some cases.

In summary, we deal successfully with the problem of obtaining reliable conclusions, which is fundamental for the application domain under consideration and is even more critical given the incompleteness of the database.

2 Methods

2.1 The CDR system

The *International Group on Dementia Drug Guidelines* has issued a position paper on assessing cognitive function in future clinical trials [9]. The working group concluded that existing testing procedures (e.g., the Alzheimer's disease assessment scale) do not properly identify all cognitive deficits characterizing AD patients, in particular attention deficits, and has recommended that automated procedures be used alongside more traditional ones to ultimately determine whether they should supersede traditional methods [9]. The CDR system has shown sensitivity in identifying mild cognitive impairment [27,41] and is able to differentiate various types of dementia (AD, DLB, VD, Huntington's Chorea [25,35–38]). The CDR system measures therapeutic response to a variety of medications in both AD [2,7,24,32,34] and DLB [22] and shows superior sensitivity in identifying AD and Huntington's disease as compared to all the most widely used non-automated procedures [25].

A selection of tasks from the CDR computerized cognitive assessment system for demented patients was used in the present study. All tasks were computer-controlled, the information being presented on high resolution monitors, and the responses recorded via a response module containing two buttons, one marked "no" and the other "yes". The language versions of the tests were appropriate for the country in which they were administered. The tests took between 25 and 40 minutes to administer, depending on the level of dementia shown by the patients, and were carried out in the following order.

- Word presentation: twelve words were presented on the monitor at the rate of 1 every 3 seconds for the patient to remember. The patient was then given 1 minute to recall as many of the words as possible.
- Word recognition: during the "word presentation test" described above, 12 distracting words were presented one at a time in a randomized order. For each word the patient was required to indicate whether or not he recognized it as being from the original list of words.
- Picture presentation: a series of 14 pictures was presented on the monitor at the rate of 1 every 3 seconds for the patient to remember.
- Picture recognition: during the "picture presentation test" described above, 14 distracting pictures were presented one at a time in a randomized order. For each picture the patient had to indicate whether or not he recognized it as being from the original series.
- Simple reaction time: the patient was instructed to press the "yes" response button as quickly as possible every time the word "yes" was presented on the monitor. Thirty stimuli were presented with a varying inter-stimulus interval.

- Digit vigilance task: a target digit was randomly selected and then constantly displayed to the right of the monitor screen. A series of digits were then presented in the center of the screen at the rate of 80 per minute and the patient was required to press the "yes" button as quickly as possible every time the digit in the series matched the target digit. Fifteen targets were used.
- Choice reaction time: either the word "no" or the word "yes" was presented on the monitor and the patient was instructed to press the corresponding button as quickly as possible. Thirty trials were used, for which each stimulus word was chosen randomly with equal probability and with varying inter-stimulus intervals.
- Spatial working memory: a picture of a house was presented on the screen with four of its nine windows lit. The patient had to memorize the position of the lit windows. For each of the 24 subsequent presentations of the house, the patient was required to decide whether or not the one window that was lit was also lit in the original presentation.
- Numeric working memory: a series of three digits was presented to the patient to memorize. This was followed by a series of 18 probe digits for each of which the patient had to decide whether or not it was in the original series and press buttons as appropriate.

2.2 Imprecise probabilities and credal classification

A classifier is an algorithm that allocates new objects to one out of a finite set of previously defined groups (or *classes*) on the basis of observations on several characteristics of the objects, called *attributes* or *features* [5]. We call *instance* a possibly incomplete assignment of values to the attributes. The records of the data set are also called *units*. A unit always contains the class value, besides the related instance. In the present application, each test of the CDR system is regarded as an attribute, the values of which are the possible outcomes of the test. The class is the variable the values of which are the possible states of dementia (including the state "normal"). The purpose of a classifier is to perform a diagnosis by assigning a patient to the correct health state.

Classification is one of the most important techniques for knowledge discovery in databases. Classifiers are inferred from data sets, making explicit the knowledge underlying data. Classification aims at achieving this goal with minimal or no user intervention. To do this, a classifier must be able to model ignorance in the most appropriate way. Basically, ignorance arises for two reasons. When data are the only source of information, we start learning from them in a state of *prior ignorance*. Modelling prior ignorance is a long-standing and difficult problem. The Bayesian literature [3] proposes models based on so-called noninformative prior distributions (or *priors*). This method, however, is very controversial [39, pp. 226–235]. Many people use noninformative priors because the effects of a given prior are severe only

for small samples and will disappear in the limit of an infinite sample, no matter which prior is used. However, data sets are often small and, moreover, "small" and "large" depend on the sample space. Also a data set with, say, 10^6 records is small if the sample space is sufficiently complex.

In order to understand the second type of ignorance, we must consider the problem of incomplete samples. These are data sets in which some values have been turned into missing data. In such cases the mechanism responsible for the missing data should not be ignored if one aims to obtain reliable conclusions, unless data are subject to a condition known as "missing at random" [17]. Unfortunately, such condition cannot be tested statistically [20, pp. 73–74] and hence it does not seem to be suited for data mining applications. More generally speaking, missing data prevent us from having complete knowledge of the likelihood, i.e. we have partial ignorance about the likelihood (or *likelihood ignorance*, for short). Of course, likelihood ignorance can produce effects for any size of the data set. Likelihood ignorance is a serious problem for knowledge discovery applications, among others, for which established methods do not provide a widely accepted solution.

Recently, there has been a great development of innovative proposals to model ignorance. Walley presents strong arguments that support the use of sets of prior densities to model prior ignorance [40]. A number of contributions show that also likelihood ignorance can be modelled satisfactorily by sets of measures [10,11,21,30,46]. The underlying idea of these modern approaches to prior and likelihood ignorance is the same: the body of all the possible states of knowledge *is* the model of ignorance. For example, all possible mechanisms responsible for the missing data, taken as a whole, are a model for likelihood ignorance. Overall, sets of probability distributions appear as a well-founded framework suited to model ignorance. Sets of probability distributions belong to the wider theory of imprecise probabilities [39] (see http://www.sipta.org for up-to-date information).

Two major consequences stem from adopting imprecise probabilities in classification problems. On the one hand, there is a much greater modelling flexibility and realism. On the other, classifications can be partially indeterminate: an object is generally mapped to a set of classes. In general, the output classes should be all interpreted as candidates for the actual class, as we have no way to rank them. In other words, the different types of ignorance may limit the strength of the conclusions. Precisely determinate classifications, i.e. strong conclusions, are a special case achieved only when the conditions justify precision.

This more general way to address classification problems is called *credal classification* (*credal set* is another name for sets of distributions, after Levi [15]). Credal classification was introduced in [44] and discussed more widely in [47]. A *credal classifier* is defined as a function that maps an object to a set of classes. A credal classifier is not only a new classifier, it implements a new way to perform classification. The concepts developed in the classification literature do not always apply to

credal classification in a straightforward way. For example, the empirical evaluation of credal classifiers is a challenging issue (see Sect. 3).

Credal classification can be explained more clearly by focusing on the special case of sequential learning tasks (here we assume that data are complete). The classifier starts in condition of prior ignorance. Every new instance is first classified and only then stored in the knowledge base together with the actual class, which is unknown at the classification stage. The classifier's knowledge grows incrementally, so that its predictions become more reliable as more units are collected. A credal classifier naturally shows this behavior. Initially it will produce all the classes (i.e., complete indeterminacy); with more units, the average output set size will decrease approaching one in the limit. If one compares this behavior with that of common classifiers that always produce a single class, even when very few units have been read, these will appear to be overconfident.

2.3 The naive credal classifier

This section presents an overview of the NCC. For details about the model, please refer to [45].

Definition of the model. Let us denote the classification variable by C, taking values in the nonempty and finite set $\mathcal{C}$, where the possible classes are denoted by lower-case letters. We measure k features $(A_1, \ldots, A_k)$ taking generic values $(a_1, \ldots, a_k) = \mathbf{a}$ from the sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$, which are assumed to be nonempty and finite.

We assume the N units of the data set, each with known values of the attributes and the class (for the moment we consider the case of complete data), are generated independently from an unknown multinomial process. Let the unknown chances of the multinomial distribution be denoted by $\theta_{c,\mathbf{a}}$ $((c, \mathbf{a}) \in \mathcal{C} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_k)$. Denote by $\theta_{a_i|c}$ the chance that $A_i = a_i$ conditional on c; similarly, let $\theta_{\mathbf{a}|c}$ be the chance that $(A_1, \ldots, A_k) = (a_1, \ldots, a_k)$ conditional on c. Let $n(c)$ and $n(a_i, c)$ be the observed frequencies of class c and of the joint state (a_i, c) in the N observations, respectively.

Both the naive Bayes classifier and the naive credal classifier are based on the assumption of probabilistic independence of the attributes conditional on the class:

$$\theta_{\mathbf{a}|c} = \prod_{i=1}^{k} \theta_{a_i|c} \quad \forall\, (c, \mathbf{a}) \in \mathcal{C} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_k. \tag{1}$$

Based on this assumption and imposing a Dirichlet prior over the chances, it is

possible to obtain a Dirichlet posterior distribution:

$$P(\theta|\mathbf{n}) \propto \prod_{c\in\mathcal{C}} \left[\theta_c^{st(c)+n(c)-1} \prod_{i=1}^{k} \prod_{a_i\in\mathcal{A}_i} \theta_{a_i|c}^{st(a_i,c)+n(a_i,c)-1} \right], \tag{2}$$

where $\mathbf{n}$ is the sample, $t(c)$ and $t(a_i, c)$ are hyperparameters corresponding to $n(c)$ and $n(a_i, c)$, respectively, and $s > 0$ is a constant representing the prior weight (also known as the number of virtual units). So far we have presented a Bayesian learning approach for the NBC. The extension to imprecise probabilities and the NCC is achieved by modelling prior ignorance by a set of Dirichlet prior densities. We consider the set of all Dirichlet priors, and, consequently, posteriors of the form (2), that are obtained by letting the t-hyperparameters vary in the following region:

$$\sum_c t(c) = 1 \tag{3}$$

$$\sum_{a_i\in\mathcal{A}_i} t(a_i, c) = t(c) \quad \forall(i, c) \tag{4}$$

$$t(a_i, c) > 0 \qquad \forall(i, a_i, c). \tag{5}$$

These constraints resemble the structural constraints to which the counts $n(c)$ and $n(a_i, c)$ naturally obey.

The model obtained in this way is a special version of the *imprecise Dirichlet model* [40] and is coherent in the strong sense of Walley [39, Section 7.8]. In this framework s is interpreted as a degree of caution that Walley suggests choosing in the interval $[1, 2]$.

Classification procedure. Let $E[U(c)|\mathbf{a}, \mathbf{n}, \mathbf{t}]$ denote the expected utility with respect to (2) from choosing class c, given $\mathbf{a}$, the previous data $\mathbf{n}$ and a vector $\mathbf{t}$ of hyperparameters. Since $\mathbf{t}$ belongs to a region, there are many such utilities for every class c, so that we cannot always compare two classes: generally, we have a partial order on the classes that only allows us to discard the dominated ones. The partial order depends on the chosen dominance criterion. We use credal dominance, defined below.

We say that class c' *credal-dominates* class c'' if and only if $E[U(c')|\mathbf{a}, \mathbf{n}, \mathbf{t}] > E[U(c'')|\mathbf{a}, \mathbf{n}, \mathbf{t}]$ for all values of $\mathbf{t}$ in the imprecise model.

Credal dominance is a special case of *strict preference* [39, Sect. 3.7.7] justified by Walley on the basis of rationality (behavioral) arguments. It was previously proposed by Seidenfeld in the commentary of a paper by Kyburg [13, p. 260, P-III′].

In the following we consider 0-1 valued utility functions, i.e., we receive utility 1 if we choose the correct class c and 0 if we do not, so $E[U(c)|\mathbf{a}, \mathbf{n}, \mathbf{t}] = P(c|\mathbf{a}, \mathbf{n}, \mathbf{t})$.

With the NCC, credal dominance reduces itself to the following nonlinear optimization problem:

$$\inf \left\{ \left[\frac{n(c'') + st(c'')}{n(c') + st(c')} \right]^{k-1} \prod_i \frac{n(a_i, c')}{n(a_i, c'') + st(c'')} \right\} \tag{6}$$

$$t(c') + t(c'') = 1 \tag{7}$$

$$t(c'), t(c'') > 0. \tag{8}$$

The optimum value of the problem is greater than 1 if and only if c' credal-dominates c''.

The classification procedure simply takes each pair of classes, tests credal dominance by the above optimization and discards the dominated class, if any. The output of the classifier is the set of classes that are not dominated. The overall computational complexity to credal-classify an instance is $O(k\,|\mathcal{C}|^2)$, given that the above optimization can be reduced to a convex optimization problem. Let us stress that the computation is exact, no approximations are involved.

The extension to incomplete samples is straightforward. We limit ourselves to the common case when the class is never missing. Everything is unchanged in the classification procedure except that the counts $n(a_i, c')$ and $n(a_i, c'')$ in (6) are replaced by $\underline{n}(a_i, c')$ and $\overline{n}(a_i, c'')$, respectively. These are the minimum value of $n(a_i, c')$ and the maximum value of $n(a_i, c'')$ achieved by replacing the missing values of feature i with the values in $\mathcal{A}_i$ in all the possible ways. The frequencies $\underline{n}(a_i, c')$ and $\overline{n}(a_i, c'')$ can be computed in linear time in the size of the data set, so that the extension to incomplete samples does not increase the computational complexity to learn the model with respect to the NBC.

Comparison with other models. The NCC allows us to model prior and likelihood ignorance under very weak assumptions, so that the classifications are inherently robust to small sample sizes and missing data. These are benefits of the innovative ideas brought by credal classification. Credal classification is a promising field and new credal classifiers have already been proposed [1,6,28], although the NCC is still the only one that deals with both prior and likelihood ignorance.

It is particularly important to compare the NCC with Ramoni and Sebastiani's robust Bayes classifier [29]. Although developed independently, the NCC and the RoC share many characteristics. Both are generalizations of the NBC to sets of distributions. Their modelling of prior ignorance is substantially different, but the RoC treatment of likelihood ignorance is much in the same spirit of the NCC. Let us analyze the differences in detail.

- The most important difference has to do with the fundamental problem of modelling prior ignorance, which is addressed reliably by the NCC for the first time,

as explained above. On the other hand, the RoC gives the option to choose a Bayesian prior. In this framework, one models prior ignorance using noninformative priors. Whatever prior is chosen, the RoC behaves exactly like an NBC if the learning sample is complete. The unreasonably precise classifications often produced when small learning samples are available, are a consequence of the overconfident modelling of prior ignorance provided by noninformative priors.

- A second difference is related to the dominance criterion and the classification procedure. Ramoni and Sebastiani propose two scores to rank classes: the *strong dominance score* and the *complete-admissible score*. The latter requires the assumption that all missing data mechanisms are equally possible [29, p. 218]. The problems we focus on do not allow one to make strong assumptions about the mechanism responsible for the missing data, so we base our comparison only on the strong dominance score.
  - The strong dominance score is based on *interval dominance* ("stochastic dominance" in [29]; and originally called "strong dominance" in [18]). In our notation, c' *interval-dominates* c'' iff $\underline{E}[U(c')|\mathbf{a}, \mathbf{n}] > \overline{E}[U(c'')|\mathbf{a}, \mathbf{n}]$, where $\underline{E}$ and $\overline{E}$ are the minimum and the maximum, respectively, of the expected utility obtained over all possible $\mathbf{t}$ vectors. Interval dominance is unnecessary cautious (see [13, p. 260, P-III vs. P-III′] and [47, Section 4]): interval dominance implies credal dominance but the converse does not hold. In practice, interval dominance makes the RoC overcautious for some instances: i.e. the RoC does not always recognize dominated classes as such.
  - There is a further problem with the classification procedure based on the strong dominance score. An instance is either classified precisely or it is unclassified. This is equivalent to have a credal classifier that outputs all classes unless one of them dominates all the others. Such a procedure can produce a very large excess of cautious. For example, an experiment in Sect. 3.3 shows the NCC produces about 2 (out of 4) classes on average, for 102 instances (about 20% of the tested cases), with about 98% chance that the actual class is in the two predicted ones. These cases are unclassified according to the strong dominance score, thus losing quite some useful information.
- Finally, the RoC deals with missing classes. This case does not seem to be frequent in practice (e.g., see the "University of California at Irvine" repository of machine learning data sets [26]); and the RoC deals with this case by an approximate approach, which is responsible for an excess of caution.

Overall, apart from the questionable model of prior ignorance, the RoC is a valuable but overly cautious tool. If we forced the NCC to use the same prior-ignorance model of the RoC, the output set of classes of the latter would always be unnecessarily larger than or equal to the output set of the former. In practice this inclusion will not necessarily be observed due to the different models of prior ignorance.

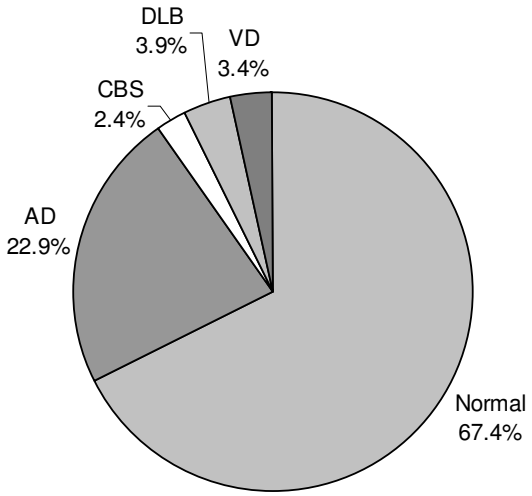

Fig. 1. Per cent distribution of the classes in the database.

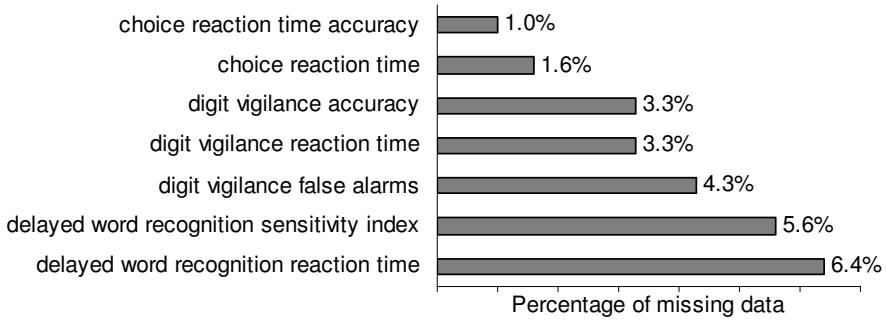

Fig. 2. Attributes used in the first analysis. The length of the bars represents the percentage of missing values.

2.4 The database

The CDR database contains 3,385 records. Each record stores the responses to the CDR system tests for a patient. The results are expressed by either continuous or integer numbers. Each record also reports the actual health state of the patient, which is classified into 5 categories: normal, AD, to undergo Coronary Bypass Surgery (CBS), DLB and VD. Figure 1 shows the distribution of the classes.

The database records more features for the dementia patients than for the normal controls, so the number of attributes depends on whether or not normal people are involved in the study. The first analysis (Sect. 3.2), taking normal people into account, uses 7 features. These are shown in Fig. 2.

The second analysis (Sect. 3.3) is based on the 1,103 records of the database re-

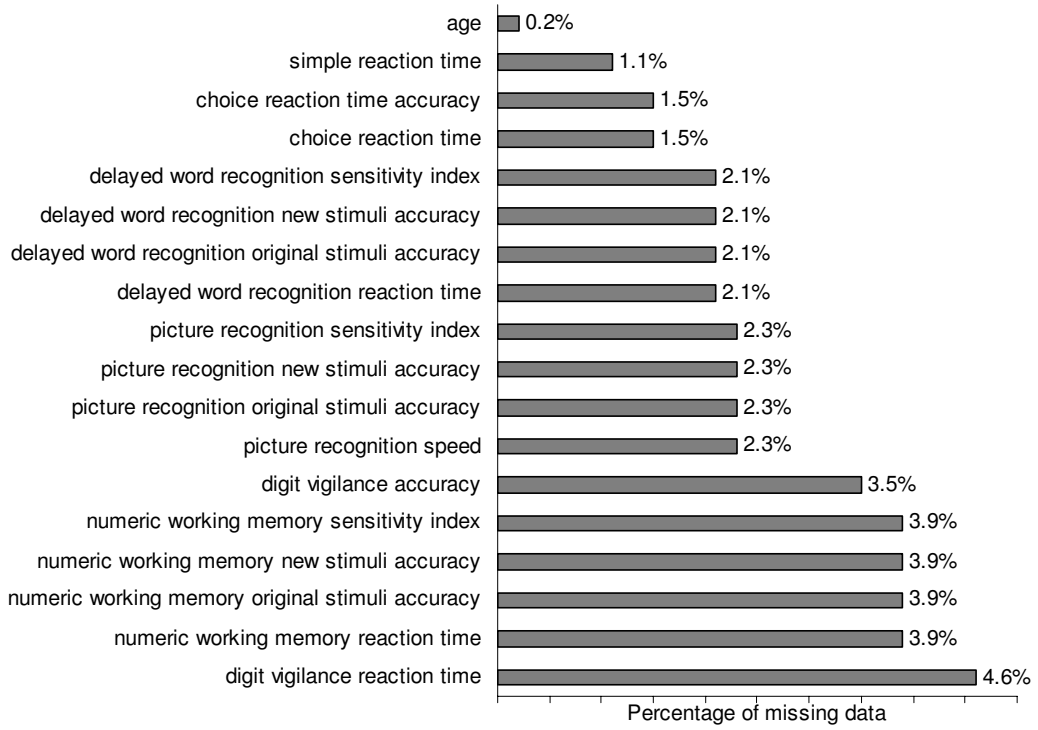

Fig. 3. Attributes used in the second analysis. The length of the bars represents the percentage of missing values.

stricted to the dementia group. In the second analysis we have included 18 attributes (see Fig. 3).

3 Experiments

The experiments aim at evaluating empirically the NCC. This stage involves comparing the NCC with the NBC and the RoC [2] .

In the following, when either the NCC or the RoC produce more than one class for an instance, we say that the classification is (partially or totally) indeterminate and the classifier is imprecise or suspends the judgment. In the opposite case we speak of determinate classifications and precise classifiers.

[2] In the version publicly available from http://kmi.open.ac.uk/projects/bkd/.

3.1 Experimental methodology

For each of the two experiments, the following steps were undertaken. The database was randomly split once into equal-size learning and test sets. The initial preprocessing stage involved (i) discretizing the data set, since the classifiers assume that the attributes are categorical; and (ii) selecting a subset of features. The database was discretized by MLC++ [12] with default options, i.e. by the common entropy-based discretization [8]. Step (ii) used Weka's [3] feature selection filters [43] (e.g., the common filter based on mutual information [16]). These two steps were carried out on the basis of the learning set only. (The numbers of attributes reported in Sect. 2.4 refer to the post-processed data set.) On the remaining test set the true classes were hidden to the classifiers to study the prediction accuracy, i.e. the relative number of correct predictions on a set of unseen instances.

As far as the classifiers are concerned, we have the choice of the priors weight. This study uses $s = 1$ for the NCC (see Sect. 2.3). The NBC [4] and the RoC were both inferred by using the noninformative uniform prior distribution with weight 1 (the weight and s play the same role). This is equivalent to initialize each count, necessary to estimate the model probabilities, to 1. As the NBC cannot deal with likelihood ignorance, the NBC was inferred by discarding the missing values, separately for each attribute. This corresponds to assume that the data are missing at random.

Finally, the notation $a \pm b$ used in the following sections represents, each time, a 0.95 confidence interval (the possible upper bounds greater than 100 should be regarded as 100%). All comparisons were tested for statistical significance using the two-samples one-tail t-test at level $p = 0.05$.

3.2 Detecting dementia

The first experiment aimed at differentiating normal from demented people. Dementias are clustered into one class so that the class variable is binary with values in: "normal" (67.4%) and "demented" (32.6%). Seven attributes describe a patient, as reported in Sect. 2.4.

3.2.1 NCC vs. NBC

First, we compare the NCC with the NBC. The results are shown in Tab. 1.

[3] Publicly available from http://www.cs.waikato.ac.nz/~ml/weka/.

[4] We also used other noninformative priors with the NBC: Haldane, Perks and Jeffreys (see [45] for the definition). The results were very similar to those presented here.

Table 1
Results of the comparison of the NCC and the NBC for the discrimination between normal people and people in the dementia group. The cells contain confidence intervals on percentages.

C_1%	N%	Ns%	S%
94.77±1.14	92.41±1.30	70.37±7.26	9.68±1.45

The columns are described below. The prediction accuracy corresponds to the relative number of correct predictions.

- C_1% is the accuracy of the NCC on the subset of instances where it is precise.
- N% is the accuracy of the NBC on the entire test set.
- Ns% is the accuracy of the NBC on the subset of instances for which the NCC is imprecise.
- S% is the percentage of instances for which the NCC is imprecise.

When the credal classifier isolates a single class, the accuracy of prediction is very high (C_1%). In about 10% of cases (S%), it suggests that there is not enough knowledge to isolate a single class and therefore it outputs both, not giving any judgement. The NBC does a significantly $(p = 2 \cdot 10^{-3})$ worse prediction (Ns%) than the one achieved on the entire set (N%). Ns% is also significantly $(p = 1.2 \cdot 10^{-9})$ worse than C_1%.

The NCC is thus able to isolate a subset of instances where robust and determinate (C_1%) predictions are possible: i.e., these predictions are independent both of whatever values might replace the missing data and the unknown state of prior knowledge. The NBC realizes non-random predictions on the subset of instances where the NCC does not provide any judgement: the NBC achieves about 70% accuracy, larger than the 50% accuracy that would be obtained by randomly guessing. This may suggest that the data violate the assumption of independence among attributes conditional on the class and that it might be worth trying more structured credal classifiers [45].

3.2.2 NCC vs. RoC

As a general remark, when both the NCC and the RoC are precise, they will tend to produce the same class. In fact, when the knowledge provided by data increases, both converge to an NBC.

Tab. 2 partitions the 1673 units of the test set according to whether the NBC and the RoC produce a determinate classification or not (i.e., two classes out of two). In a subset of 1511 instances both classifiers produce a single class and, notably, the two classes coincide. When the RoC suspends the judgement (144 instances), also the NCC does so. Conversely, in 18 cases the NCC does not provide any judgement, whereas the RoC is precise. In this subset the confidence interval for the accuracy of

Table 2
Partition of the test set according to the different behaviors of the NCC and the RoC. The notation "d" represents precisely determinate classifications, "¬d" indeterminate ones.

		NCC	
		d	¬d
RoC	d	1511	18
	¬d	0	144

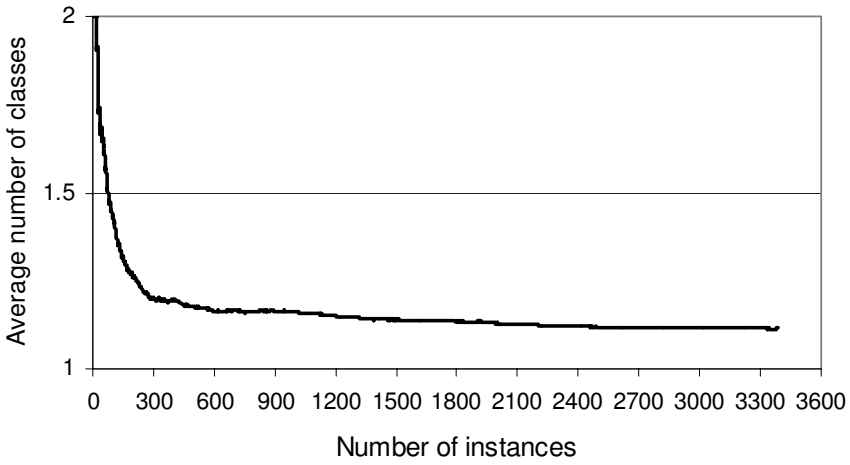

Fig. 4. The average number of classes produced by the NCC as a function of the number of instances used to infer the classifier.

the RoC is $66.7\% \pm 22.2$. The hypothesis that the RoC prediction accuracy is larger than 50% is rejected $(p = 0.09)$. For these 18 cases, the NCC correctly withholds the judgement as consequence of its imprecise model for prior ignorance.

3.2.3 A further analysis

An experiment with the entire database provides results that allow one to obtain a clearer picture of how credal classification works. Figure 4 shows the number of classes produced by the NCC as a function of the number of units used to learn the classifier. Initially, when the NCC is fed with very few past examples, the output indeterminacy is very high: the NCC tends to produce totally indeterminate classifications (2 classes). Fairly quickly, as more data accumulate, the average number of classes gets closer to 1. Note that if the data were complete, the decrease would be much steeper: the missing data create an substantial area of ignorance also after using 3,000 units.

3.3 Discriminating among dementias

The goal of the second analysis is assigning a diseased patient to the correct disease type. The class takes values in the set of 4 dementias reported in Sect. 2.4, which also reports the 18 attributes used.

3.3.1 NCC vs. NBC

The results are presented in Tab. 3.

Table 3
Results of the discrimination among dementias. The cells contain confidence intervals on percentages.

C_1%	Cs%	N%	Ns%	Rs%	S%
94.05±2.31	98.42±2.21	89.76±2.59	75.59±7.62	45.6±8.84	23.22±3.61

There are two more columns as compared to Tab. 1:

- Cs% is a set-accuracy, i.e. it represents the probability that the actual class belongs to the set of classes proposed by the NCC. This measure is computed in the subset of instances where the NCC is imprecise.
- Rs% is the accuracy of a uniformly random predictor that randomly chooses one of the classes proposed by the NCC. This measure is computed in the subset of instances for which the NCC is imprecise and the actual class is contained in the NCC output set.

The performance of the credal classifier is very good when the classification is determinate (C_1%). C_1% is significantly greater than N% ($p = 6.9 \cdot 10^{-3}$). The number of determinate classifications are about 3/4 of the test set. (S% is larger than that shown in Tab. 1 because now the learning set size is about 1/3 of the learning set of the first experiment and more attributes are used; larger collections of data, with respect to the number of missing values, would quickly decrease this type of indeterminacy.)

In the area where the NCC is not precise, the NBC improves on predicting at random (Ns% vs. Rs%) but its behavior is not very satisfactory (Ns%); both C_1% ($p = 4.15 \cdot 10^{-6}$) and N% ($p = 2.9 \cdot 10^{-4}$) are significantly much greater than Ns%.

In our view, the NCC provides a more reliable approach to instances that are hard to classify, given the available knowledge. The NCC produces about 2.3 classes on average in the area of imprecision (S%). Although such partial indeterminacy does not allow one to draw strong conclusions, the output set of classes almost certainly (Cs%) contains the actual one. The ability to discard almost 2 useless classes out of 4, on average, is a very significant and useful predictive characteristic of the NCC,

which by doing so maintains reliability even under hard conditions.

High accuracy can be obtained trivially when set-predictions are allowed (e.g., by producing all the classes each time, a 100% prediction accuracy would be reached). It is therefore useful to compare the NCC with a dumb predictor in order to better appreciate the results yielded by the NCC. When the NCC is imprecise, the dumb predictor we use as a comparison outputs a set of classes drawn from the sets the NCC produces, with the same (set) frequencies. For example, the dumb predictor will output the set {DLB,VD} in 40% of instances if the NCC outputs such set with the same frequency. The predictor is defined as dumb because it does not take advantage of the values of the attributes, but it is useful to show how much the NCC improves over trivial predictions. The accuracy of the dumb predictor is 70.5±8.09. By comparing this with Cs%, we see that the NCC exploits the knowledge in the data to achieve a significantly ($p = 1.92 \cdot 10^{-20}$) much superior accuracy.

3.3.2 NCC vs. RoC

Tab. 4 partitions the 547 units of the test set depending upon the production of a determinate or a partially indeterminate classification by the NBC and the RoC.

Table 4
Partition of the test set according to the different behaviors of the NCC and the RoC. The notation "d" represents precisely determinate classifications, "¬d" indeterminate ones.

		NCC	
		d	¬d
RoC	d	406	25
	¬d	14	102

Both classifiers are precise in a subset of 406 instances, and the two classes coincide.

The RoC classifications are totally indeterminate for 102 instances, i.e., in an equivalent way, the RoC outputs all the 4 classes each time. Here the average set size produced by the NCC is 2.3 and the chance that the actual class is contained there is 98%±2.74. In this case the credal dominance criterion enables the NCC to exploit the data much better than the RoC.

The NCC is precise in 14 instances where the RoC classifications are totally indeterminate. The NCC accuracy is 50%±26.7 here, and is significantly ($p = 0.046$) larger than the accuracy of the uniformly random predictor: 25%.

In the remaining subset of 25 instances, the RoC is precise and predicts all instances correctly. The NCC outputs sets of average size 2.1, which always contain

the actual class. With the NCC, these 25 cases are on the border of determinate classifications; the partial indeterminacy depends on the NCC's more cautious model of prior ignorance which needs slightly more knowledge from the data for the NCC to be precise. This can be checked by setting the parameter s slightly smaller than 1, by which these cases will be classified precisely.

In summary, as far as the comparison between NCC and RoC is concerned, the experiments shows on the one hand the drawbacks of the RoC's strong dominance score, as well as the negligible impact on this application of the remaining differences between the classifiers.

3.3.3 Confusion matrix

Table 5
Confusion matrix. The boldface values are the numbers of instances correctly classified by the NCC.

		Predicted class			
		AD	CBS	DLB	VD
Actual class	AD	**318**	3	1	0
	CBS	0	**34**	0	0
	DLB	0	2	**25**	8
	VD	4	0	7	**18**

To better analyze the ability of the credal classifier to assign a patient to the actual class, we represent the *confusion matrix* in Tab. 5. We restrict the attention to the instances for which the NCC is precise. (With the other instances, we already know that the 2.3 average output classes contain the actual class almost surely; in addition, the indeterminate cases need a more general representation than the one offered by the confusion matrix). A cell of the matrix reports the number of instances related to given predicted and actual classes, as computed from the test set: e.g., the number of Alzheimer-diseased people that are correctly diagnosed is 318.

The confusion matrix shows that the NCC performance to assign patients to actual classes is excellent: for instance, the misdiagnosed subjects suffering from Alzheimer's disease are only 4 out of 322, 3 as CBS and 1 as DLB. The pair DLB and VD is an exception, as there is substantial confusion between the two types of dementia: 8 DLB patients are misdiagnosed as VD and, vice versa, 7 VD patients are classified as DLB. Another evidence arising from the confusion matrix is the capability of the NCC to discriminate between AD and DLB. This is a very important result for research on dementia: DLB has frequently been misdiagnosed as AD. Here we show in an objective way that the presented system can differentiate them very accurately.

3.3.4 A final remark

Clearly, the evaluation of credal classifiers is challenging, as is the comparison with other classifiers. Drawing stable conclusions requires an analytical effort superior to that needed for common classifiers. Some of these challenging issues are also faced by RoC users. Ramoni et al. [31] propose an approach to evaluate empirically the RoC based on cost analysis. Costs are placed not only on wrong decisions but also on the impossibility to decide, due to the total indeterminacy of some RoC outputs. This approach is attractive and might be extended to the partial indeterminacies of the NCC. However, the costs, especially of the second type, can be very difficult to quantify in practice. This is also the case of our application. For this reason we have chosen to analyze the results of the experiments from different viewpoints, thus allowing the domain experts to be presented with a broad picture of the NCC behavior.

4 Conclusions

Cognitive tests for dementias are becoming more and more important, as early diagnosis seems to be the basis for coping successfully with the diseases. This paper shows that coupling targeted cognitive tests such as the CDR computerized system with a reliable classifier such as the NCC, enables very accurate automated diagnoses. The results are briefly summarized below.

1673 records of patients' data entered the first experiment, set up to detect dementia. 9.68% of these were not classified, due to limited knowledge. On the remaining 90.32%, 94.77% were correctly classified either as "demented" or "normal".

547 records of patients' data entered the second experiment, set up to differentiate the four types of dementias. 23.22% of these could not be given a unique classification, but 98.42% patients were correctly classified in one of the 2.3 diseases the system proposed on average. On the remaining 76.78% of patients, 94.05% were correctly classified.

It is often useful for clinicians to know how accurate is a system in differentiating diseases i and j. We define such measure here as the probability that the actual and predicted classes coincide given that both belong to $\{i, j\}$. Using Tab. 5, we report the discrimination accuracies in Tab. 6, with reference to the mentioned 76.78% of patients.

Overall, the system shows an excellent discrimination ability. The discrimination between Alzheimer's disease and dementia with Lewy bodies is particularly important, as non-automated and non-computerized diagnoses often fail to detect the subtle differences in symptoms linked to the two disease types.

Table 6
Per cent differentiation accuracy between dementia pairs.

	CBS	DLB	VD
AD	99.15	99.71	98.82
CBS		96.72	100.00
DLB			74.14

More generally speaking, diagnoses have also been shown to be robust to prior ignorance and a substantial number of missing values in the learning database. This result is due to the powerful characteristics of credal classification. However, the imprecision resulting from the missing data did result in less precise inferences. Data sets with significantly less missing values could reduce the indeterminacies in the predicted classes, thus enabling a full predictive capability of the method.

Acknowledgements

Marco Zaffalon would like to thank Peter Walley for his encouragement to develop credal classification and for many important suggestions and enlightening discussions. Thanks also to L. M. Gambardella and C. Lepori for their kind attention and support. The authors are grateful to two anonymous referees for useful and detailed comments. This research was partially supported by the Swiss NSF grant 2100-067961.02/1, and by the Swiss CTI grant 4217.1.

References

[1] J. Abellán and S. Moral. Building classification trees using the total uncertainty criterion. In G. de Cooman, T. Fine, and T. Seidenfeld, editors, *ISIPTA'01*, pages 1–8, The Netherlands, 2001. Shaker Publishing.

[2] H. Allain, E. Neuman, M. Malbezin, V. Salzman, Guez D., K Wesnes, and J. M. Gandon. Bridging study of S12024 in 53 in-patients with Alzheimer's disease. *J. Am. Geriatr. Soc.*, 45:125–126, 1997.

[3] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, New York, 1996.

[4] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley, New York, 1973.

[5] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, 2001. 2nd edition.

[6] E. Fagiuoli and M. Zaffalon. Tree-augmented naive credal classifiers. In *IPMU 2000: Proceedings of the 8th Information Processing and Management of Uncertainty in*

Knowledge-Based Systems Conference, pages 1320–1327, Spain, 2000. Universidad Politcnica de Madrid.

[7] T. D. Fakouhi, Jhee S. S., J. J. Sramek, C. Benes, P. Schwartz, G. Hantsburger, R. Herting, E. A. Swabb, and N. R. Cutler. Evaluation of cycloserine in the treatment of Alzheimer's disease. *J. Geriatr. Psychiatry Neurol.*, 8:226–230, 1995.

[8] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th international joint conference on artificial intelligence*, pages 1022–1027, San Francisco, CA, 1993. Morgan Kaufmann.

[9] S. Ferris, U. Lucca, R. Mohs, B. Dubois, K. Wesnes, H. Erzigkeit, D. Geldmacher, and N. Bodick. Objective psychometric tests in clinical trials of dementia drugs. *Alzheimer Disease and Associated Disorders*, 11(3):34–38, 1997. Position paper from the International Working Group on Harmonisation of Dementia Drug Guidelines.

[10] J. L. Horowitz and C. F. Manski. Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. *Journal of Econometrics*, 84:37–58, 1998.

[11] J. L. Horowitz and C. F. Manski. Imprecise identification from incomplete data. In G. de Cooman, T. Fine, and T. Seidenfeld, editors, *ISIPTA'01*, pages 213–218, The Netherlands, 2001. Shaker Publishing.

[12] R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger. MLC++: a machine learning library in C++. In *Tools with Artificial Intelligence*, pages 740–743. IEEE Computer Society Press, 1994.

[13] H. E. Jr. Kyburg. Rational belief. *The behavioral and brain sciences*, 6:231–273, 1983.

[14] P. L. Le Bars, M. M. Katz, N. Berman, T. M. Itil, A. M. Freedman, and A. F. Schatzberg. A placebo-controlled, double-blind, randomized trial of an extract of Ginkgo biloba for dementia. *JAMA*, 278(16):1327–1332, 1997.

[15] I. Levi. *The Enterprise of Knowledge*. MIT Press, London, 1980.

[16] D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proc. of Speech and Natural Language Workshop*, pages 212–217, San Francisco, 1992. Morgan Kaufmann.

[17] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.

[18] R. D. Luce and H. Raiffa. *Games and Decisions*. Wiley, New York, 1957.

[19] S. Mani, M. B. Dick, M. J. Pazzani, E. L. Teng, D. Kempler, and I. M. Taussig. Refinement of neuro-psychological tests for dementia screening in a cross cultural population using machine learning. In W. Horn, Y. Shahar, G. Lindberg, S. Andreassen, and J. Wyatt, editors, *Lecture Notes in Computer Science*, volume 1620, pages 326–335. Springer, 1999. Proc. of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, AIMDM'99, Aalborg, Denmark.

[20] C. Manski. The selection problem in Econometrics and Statistics. In C. R. Rao, G. S. Maddala, and H. Vinod, editors, *Handbook of Statistics, Vol. 11: Econometrics*, pages 73–84. North-Holland, Amsterdam, 1993.

[21] C. Manski. *Partial Identification of Probability Distributions*. Department of Economics, Northwestern University, USA, 2002. Draft book.

[22] I. McKeith, T. Del Ser, F. Spano, K. Wesnes, R. Anand, A. Cicin-Sain, R. Ferrera, and R. Spiegel. Efficacy of rivastigmine in dementia with Lewy bodies: results of a randomised placebo-controlled international study. *Lancet*, 356:2031–2036, 2000.

[23] I. G. McKeith and G. A. Ayre. Consensus criteria for the clinical diagnosis of dementia with Lewy bodies. In K. Iqbal, B. Winblad, T. Nishimura, M. Takeda, and H. M. Wisnieswski, editors, *Alzheimer's Disease: Biology, Diagnosis and Therapeutics*, pages 167–178. Wiley, 1997.

[24] E. Mohr, V. Knott, M. Sampson, K. Wesnes, R. Herting, and T. Mendis. Cognitive and quantified electroencephalographic correlates of cycloserine treatment in Alzheimer's disease. *Clinical Neuropsychopharmacology*, 18:23–38, 1995.

[25] E. Mohr, D. Walker, C. Randolph, M. Sampson, and T. Mendis. The utility of clinical trial batteries in the measurement of Alzheimer's and Huntington's dementia. *International Psychogeriatrics*, 3:397–411, 1996.

[26] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases, 1995. http://www.sgi.com/Technology/mlc/db/.

[27] C. G. Nicholl, S. Lynch, C. A. Kelly, L. White, L. Simpson, P. M. Simpson, K. Wesnes, and B. M. N. Pitt. The cognitive drug research computerised assessment system in the evaluation of early dementia—is speed of the essence? *International Journal of Geriatric Psychiatry*, 10:199–206, 1995.

[28] P. Nivlet, F. Fournier, and J.-J. Royer. Interval discriminant analysis: an efficient method to integrate errors in supervised pattern recognition. In G. de Cooman, T. Fine, and T. Seidenfeld, editors, *ISIPTA'01*, pages 284–292, The Netherlands, 2001. Shaker Publishing.

[29] M. Ramoni and P. Sebastiani. Robust Bayes classifiers. *Artificial Intelligence*, 125(1–2):209–226, 2001.

[30] M. Ramoni and P. Sebastiani. Robust learning with missing data. *Machine Learning*, 45(2):147–170, 2001.

[31] M. Ramoni, P. Sebastiani, and R. Dybowski. Robust outcome prediction for intensive-care patients. *Methods of Information in Medicine*, 40:39–45, 2001.

[32] K. R. Siegfried. Pharmacodynamic and early clinical studies with velnacrine. *Acta Neurol. Scand.*, 149(10):26–28, 1993.

[33] P. M. Simpson, D. J. Surmon, K. A. Wesnes, and G. R. Wilcock. The cognitive drug research computerised assessment system for demented patients: a validation study. *International Journal of Geriatric Psychiatry*, 6:95–102, 1991.

[34] L. Templeton, A. Barker, K. Wesnes, and D. Wilkinson. A double-blind, placebo-controlled trial of intravenous flumazenil in Alzheimer's disease. *Human Psychopharmacology*, 14:239–245, 1999.

[35] M. P. Walker, G. A. Ayre, C. H. Ashton, V. R. Marsh, K. Wesnes, E. K. Perry, J. T. O'Brien, I. G. McKeith, and C. G. Ballard. A psychophysiological investigation of fluctuating consciousness in neurodegenerative dementias. *Human Psychopharmacology*, 14:483–489, 1999.

[36] M. P. Walker, G. A. Ayre, J. L. Cummings, K. Wesnes, I. G. McKeith, J. T. O'Brien, and C. G. Ballard. Quantifying fluctuation in dementia with Lewy bodies, Alzheimer's disease and vascular dementia. *Neurology*, 54:1616–1625, 2000.

[37] M. P. Walker, G. A. Ayre, J. L. Cummings, K. Wesnes, I. G. McKeith, J. T. O'Brien, and C. G. Ballard. The clinician assessment of fluctuation and the one day fluctuation assessment scale. *British Journal of Psychiatry*, 177:252–256, 2000.

[38] M. P. Walker, G. A. Ayre, E. K. Perry, K. Wesnes, I. G. McKeith, M. Tovee, J. A. Edwardson, and C. G. Ballard. Quantification and characterisation of fluctuating cognition in dementia with Lewy bodies and Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 11:327–335, 2000.

[39] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

[40] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J. R. Statist. Soc. B*, 58(1):3–57, 1996.

[41] K. Wesnes. Predicting, assessing, differentiating and treating the dementias: experience in MCI and various dementias using the CDR computerised cognitive assessment system. In B. Vellas and L. J. Fitten, editors, *Research and practice in Alzheimer's disease*, volume 3, pages 59–65. Serdi, Paris, 2000.

[42] K. Wesnes, K. Hildebrand, and E. Mohr. Computerised cognitive assessment. In G. W. Wilcock, R. S. Bucks, and K. Rocked, editors, *Diagnosis and management of dementia: a manual for memory disorders teams*, pages 124–136. Oxford Univ. Press, Oxford, 1999.

[43] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[44] M. Zaffalon. A credal approach to naive classification. In G. de Cooman, F. Cozman, S. Moral, and Walley P., editors, *ISIPTA'99*, pages 405–414, Univ. of Gent, Belgium, 1999. The Imprecise Probabilities Project.

[45] M. Zaffalon. Statistical inference of the naive credal classifier. In G. de Cooman, T. Fine, and T. Seidenfeld, editors, *ISIPTA'01*, pages 384–393, The Netherlands, 2001. Shaker Publishing.

[46] M. Zaffalon. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105(1):105–122, 2002.

[47] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002.