Artificial Intelligence in Medicine xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

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

Elicitation of neurological knowledge with argument-based machine learning

Vida Groznik[a,*], Matej Guid[a], Aleksander Sadikov[a], Martin Možina[a], Dejan Georgiev[b], Veronika Kragelj[c], Samo Ribarič[c], Zvezdan Pirtošek[b], Ivan Bratko[a]

[a] Artificial Intelligence Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
[b] Department of Neurology, University Medical Centre Ljubljana, Zaloška cesta 2, SI-1000 Ljubljana, Slovenia
[c] Faculty of Medicine, University of Ljubljana, Vrazov trg 2, SI-1104 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 11 July 2012
Received in revised form 7 August 2012
Accepted 19 August 2012

Keywords:
Argument-based machine learning
Knowledge elicitation
Decision support systems
Parkinsonian tremor
Essential tremor

ABSTRACT

Objective: The paper describes the use of expert's knowledge in practice and the efficiency of a recently developed technique called argument-based machine learning (ABML) in the knowledge elicitation process. We are developing a neurological decision support system to help the neurologists differentiate between three types of tremors: Parkinsonian, essential, and mixed tremor (comorbidity). The system is intended to act as a second opinion for the neurologists, and most importantly to help them reduce the number of patients in the "gray area" that require a very costly further examination (DaTSCAN). We strive to elicit comprehensible and medically meaningful knowledge in such a way that it does not come at the cost of diagnostic accuracy.
Materials and methods: To alleviate the difficult problem of knowledge elicitation from data and domain experts, we used ABML. ABML guides the expert to explain critical special cases which cannot be handled automatically by machine learning. This very efficiently reduces the expert's workload, and combines expert's knowledge with learning data. 122 patients were enrolled into the study.
Results: The classification accuracy of the final model was 91%. Equally important, the initial and the final models were also evaluated for their comprehensibility by the neurologists. All 13 rules of the final model were deemed as appropriate to be able to support its decisions with good explanations.
Conclusion: The paper demonstrates ABML's advantage in combining machine learning and expert knowledge. The accuracy of the system is very high with respect to the current state-of-the-art in clinical practice, and the system's knowledge base is assessed to be very consistent from a medical point of view. This opens up the possibility to use the system also as a teaching tool.

1. Introduction and motivation

Essential tremor (ET) is one of the most prevalent movement disorders [1]. It is characterized by postural and kinetic tremor with a frequency between 6 and 12 Hz. Although it is regarded as a symmetrical tremor, ET usually starts in one upper limb and then spreads to the other side affecting the contralateral upper limb, consequently spreading to the neck and vocal cords, giving rise to the characteristic clinical picture of the disorder. However, there are many deviations from this classical presentation of ET, e.g. bilateral tremor onset, limb tremor only, head tremor only, isolated voice tremor. Parkinsonian tremor (PT), on the other hand, is a resting tremor classically described as "pill rolling" tremor with a frequency between 4 and 6 Hz. It is one of the major signs of Parkinson's disease (PD), which also includes bradykinesia, rigidity and postural instability. PT is typically asymmetrical, being more pronounced on the side more affected from the disease onset. Although distinct clinical entities, ET is very often misdiagnosed as PT [2]. Results from clinical studies show that ET is correctly diagnosed in 50–63%, whereas PT in 76% of the cases. Co-existence of both disorders is also possible [3]. In addition, PT can be very often observed when the upper limbs are stretched (postural tremor) and even during limb movement (kinetic tremor), which further complicates the differential diagnosis of the tremors.

Digitalized spirography is a quantitative method of tremor assessment [4], based on spiral drawing by the patient on a digital tablet. In addition to precise measurement of tremor frequency, spirography describes tremors with additional parameters – these, together with physical neurological examination, offer new means

* Corresponding author. Present address: Faculty of Computer and Information Science, University of Ljubljana, Tržaška cesta 25, SI-1000 Ljubljana, Slovenia. Tel.: +386 1 4768987; fax: +386 1 4768386.
E-mail address: vida.groznik@fri.uni-lj.si (V. Groznik).
URL: http://www.ailab.si/vida (V. Groznik).

0933-3657/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2012.08.003

to differentiate between numerous types of tremors [4,5], including ET and PT. The use of spirography for diagnostic purposes is a relatively recent idea, and only a few medical centers in the world are currently using it, among them Columbia University Medical Center and University Clinical Centre Ljubljana.

The paper describes the process of building a decision support system (DSS) for diagnosing and differentiating between aforementioned three types of tremors, namely ET, PT, and mixed tremor (MT; both ET and PT at the same time). It mainly focuses on the task of knowledge acquisition as this is usually the most challenging part of the project. The motivation for the DSS is as follows. Although several sets of guidelines for diagnosing both ET and PT do exist [6,7], none of them enjoys general consensus in the community. Furthermore, none of these guidelines takes into account additional information from spirography. Our DSS combines all sources of knowledge, experts' background knowledge, machine-generated knowledge, and spirography data in an attempt to improve prediction accuracy. However, at the same time and even more importantly, our DSS uses a very comprehensible model, making it very suitable for explaining its decisions. Therefore it could be used as a teaching tool as well.

Apart from improved data acquisition and storage, the main expected benefits of the DSS are twofold. By acting as a second opinion, mostly for difficult cases, the combined diagnostic accuracy is expected to increase, reducing the need for patients to undergo an invasive, and very expensive further examination (DaTSCAN). This will also save both patients' and doctors' time.

Our knowledge acquisition process was based on argument-based machine learning (ABML) [8]. ABML seamlessly combines the domain expert's knowledge with machine-induced knowledge, and is very suitable for the task of knowledge elicitation as it involves the expert in a very natural dialogue-like way [9]. The expert is not required to give general knowledge of the domain (which can be hard), but is only asked to explain concrete examples which the machine cannot correctly classify on its own. The process usually results in improved accuracy and comprehensibility [8]. Such focused knowledge elicitation also saves a lot of expert's time.

The organization of the paper is as follows. Section 2 discusses related work and through it additional motivation for our work. Section 3 describes the domain, including the spirography and the DaTSCAN examinations. Section 4 relates the essential methodological ingredients of our approach and Section 5 sheds light on how the approach was applied in practice through detailed examples. The evaluation setup and the results are presented in Sections 6 and 7, respectively. The discussion of the results follows in Section 8, and at the end we give some concluding remarks and plans for the future.

2. Related work

The knowledge acquisition bottleneck [10] is one of the main issues in building a DSS, particularly in medical domains [11]. Several knowledge elicitation approaches have been proposed [12–16]. These approaches elicit knowledge by direct interaction between the expert and the knowledge engineer (e.g. questionnaires, interviews, observations, etc.). However, the problem of knowledge elicitation remains open [17,18].

There are several approaches to knowledge elicitation in medicine. For the use in earlier medical expert systems experts provided their knowledge in the form of general rules which were then encoded in the system. Expert system that uses this kind of approach is MYCIN.

Since it is hard for the expert to express the overall knowledge of the whole domain taking into account all of the specific cases, the approach of providing explanations to specific cases seems to work better in practice. This approach is used by ripple down rules (RDRs) which can be treated as a binary tree. Each node in a tree represents one *if-(and)-then* rule which has two branches. When a wrong classification occurs, a new rule is attached to the appropriate branch resulting in one new node. By correcting wrong classifications the tree grows over time. The weaknesses of the method are that knowledge can be repeated in the knowledge base and that it provides a single classification of the data [19]. To eliminate the last weakness the multiple classification ripple down rules have been developed to allow multiple independent classifications [20]. A well known medical expert system based on RDRs technique is Pathology Expert Interpretative Reporting System [21].

Forsyth and Rada proposed machine learning techniques as an alternative approach to solving knowledge elicitation problem [22]. Although successful in building knowledge bases [23], the automatically induced models are rarely compliant with the way experts want their knowledge to be expressed. It is likely that incomprehensible models will not be trusted by the experts and users.

Expert system that relies mainly on machine learning techniques for eliciting knowledge is Medical Knowledge Elicitation System (MediKES). In MediKES the expert's knowledge is elicited in two steps. The first step is the knowledge acquisition step which automatically elicits expert's knowledge from the electronic medical record. After this step the system has a decision logic for medical treatments from every expert at its disposal. Elicited knowledge is then "visualized by concept mapping technique, which presents the information graphically" [24] and makes it more understandable. However, the weak side of this system is that it can give us odd results if, let us say, there is a physician who only has specific cases that have to be treated differently than others or if a physician has a small number of cases or similar. The decision logic for the treatment could therefore be somewhat unusual if not incorrect.

There is a common belief that combining machine learning and expert's knowledge would give us the best results [25]. Inductive learning system LINUS [26,27] uses CN2 [28] attribute-value learner for learning diagnostic rules from the patients' data and combines them with the background knowledge of the domain expert. Knowledge was given "in the form of typical co-occurrences of symptoms" [29]. LINUS was used for learning rules for early diagnosis of rheumatic diseases.

3. Domain description

Our data set consists of 122 patients diagnosed and treated at the Department of Neurology, University Medical Centre Ljubljana. The patients were diagnosed by a physician with either ET, PT, or MT which represent possible class values for our classification task.

The class distribution is: 52 patients diagnosed with ET, 46 patients with PT and 24 patients with MT.

The patients were initially described using 69 attributes. These were reduced to 47 attributes during the preprocessing of these data. The excluded attributes contained mostly unknown values or comments and were as such irrelevant for building the model. One of the excluded attributes is the result of a DaTSCAN. About a half of the attributes were derived from the patient's history data and the neurological examination, the other half included data from spirography. All included attributes and their rate of missing values are detailed in Table 1.

3.1. Spirography

Digitalized spirography is a relatively new computer-assisted method for detection and evaluation of tremors [4]. It has

Please cite this article in press as: Groznik V, et al. Elicitation of neurological knowledge with argument-based machine learning. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.08.003

Table 1
The initial attributes used at the beginning of the knowledge elicitation process and their rate of missing values.

Attribute	Description	Number of missing values, *n* (%)
Age	Age of a patient	6 (5.26)
Alcohol.response	Response of tremor on alcohol – ET diminishes on alcohol in about 60% of cases; PT does not	88 (77.19)
Bare.left.freq.harmonics	Harmonic frequencies on spectral analysis when drawing without template – left hand	9 (7.89)
Bare.left.freq.maxamp	Maximal amplitude frequency of the tremor on the left side during bare hand drawing	33 (28.95)
Bare.left.freq.range	Frequency range of the tremor on the left side during bare hand drawing	24 (21.05)
Bare.left.radius.angle	Radius–angle transform on the left side during bare-hand drawing	20 (17.54)
Bare.left.speed.time	Speed–time transform on the left side during bare-hand drawing	19 (16.67)
Bare.right.freq.harmonics	Harmonic frequencies on spectral analysis when drawing without template on the right side	11 (9.65)
Bare.right.freq.maxamp	Maximal amplitude frequency of the tremor on the right side during bare hand drawing	38 (33.33)
Bare.right.freq.range	Frequency range of the tremor on the right side during bare hand drawing	16 (14.04)
Bare.right.radius.angle	Radius–angle transform on the right side during bare-hand drawing	11 (9.65)
Bare.right.speed.time	Speed–time transform on the right side during bare-hand drawing	8 (7.02)
Bradykinesia.left	Slowed movement on the left side	5 (4.39)
Bradykinesia.right	Slowed movement on the right side	5 (4.39)
Diagnosis	Diagnosis	0 (0.00)
Disease.duration	Duration of the disease	53 (46.49)
Education	Education	11 (9.65)
Gait	Clinical neurological examination of gait; gait has specificities in PB and other neurological disorders; gait is normal in ET	80 (70.18)
History	Family history (is there somebody in your family who has the same condition (tremor/disease?))	29 (25.44)
Hypokinesia.left	Paucity of movement at the left side	51 (44.74)
Hypokinesia.right	Paucity of movement at the right side	51 (44.74)
Kinetic.tremor.up.left	Tremor when the limb moves (isotonic contraction) on the left upper extremity	12 (10.53)
Kinetic.tremor.up.right	Tremor when the limb moves (isotonic contraction) on the right upper extremity	12 (10.53)
Postural.tremor.up.left	Tremor when the limb is activated, but does not move (isometric contraction) on the left upper extremity	10 (8.77)
Postural.tremor.up.right	Tremor when the limb is activated, but does not move (isometric contraction) on the right upper extremity	10 (8.77)
Qualitative.spiral	Qualitative evaluation of the spiral drawing	10 (8.77)
Resting.tremor.up.left	Tremor when the limb rests on the left side	6 (5.26)
Resting.tremor.up.right	Tremor when the limb rests on the right upper extremity	6 (5.26)
Rigidity.low.left	Rigidity hypertonia of extrapyramidal (parkinsonian) type on the left lower extremity	92 (80.70)
Rigidity.low.right	Rigidity hypertonia of extrapyramidal (parkinsonian) type on the right lower extremity	92 (80.70)
Rigidity.neck	Rigidity hypertonia of extrapyramidal (parkinsonian) type on the neck	97 (85.09)
Rigidity.up.left	Rigidity hypertonia of extrapyramidal (parkinsonian) type on the left upper extremity	7 (6.14)
Rigidity.up.right	Rigidity hypertonia of extrapyramidal (parkinsonian) type on the right upper extremity	7 (6.14)
Sex	Gender	0 (0.00)
Template.left.freq.harmonics	Harmonic frequencies on spectral analysis during template drawing on the left side	9 (7.89)
Template.left.freq.maxamp	Maximal amplitude frequency on the left side during template drawing	55 (48.25)
Template.left.freq.range	Frequency range on the left side during template drawing	49 (42.98)
Template.left.radius.angle	Radius–angle transform on the left side during template drawing	44 (38.60)
Template.left.speed.time	Speed–time transform on the left side during template drawing	43 (37.72)
Template.right.freq.harmonics	Harmonic frequencies on spectral analysis during template drawing on the right side	10 (8.77)
Template.right.freq.maxamp	Maximal amplitude frequency of the tremor on the right side during template drawing	53 (46.49)
Template.right.freq.range	Frequency range of the tremor on the right side during template drawing	53 (46.49)
Template.right.radius.angle	Radius–angle transform on the right side during template drawing	46 (40.35)
Template.right.speed.time	Speed–time transform on the right side during template drawing	54 (47.37)
Tremor.duration	Duration of the tremor	14 (12.28)
Tremor.neck	Neck tremor	10 (8.77)
Tremor.start	On which side did the tremor start?	44 (38.60)

been used to evaluate different types of tremor. The system for acquisition and analysis of spiral drawings is composed of a computer, a tablet for digital acquisition of the signal and a special pencil. The task of the patient is to draw an Archimedes spiral on the tablet. Different quantitative parameters are provided by spirography. Besides the spectral analysis of the acquired signal, which provides information about the tremor frequency, commonly used are also radius–angle transform and speed–time transform. Radius–angle transform depicts changes of the radius as a function of changes of the angle during spiral drawing. Speed–time transform represents acceleration during spiral drawing. Linear and angular acceleration transform is being calculated by the

Please cite this article in press as: Groznik V, et al. Elicitation of neurological knowledge with argument-based machine learning. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.08.003

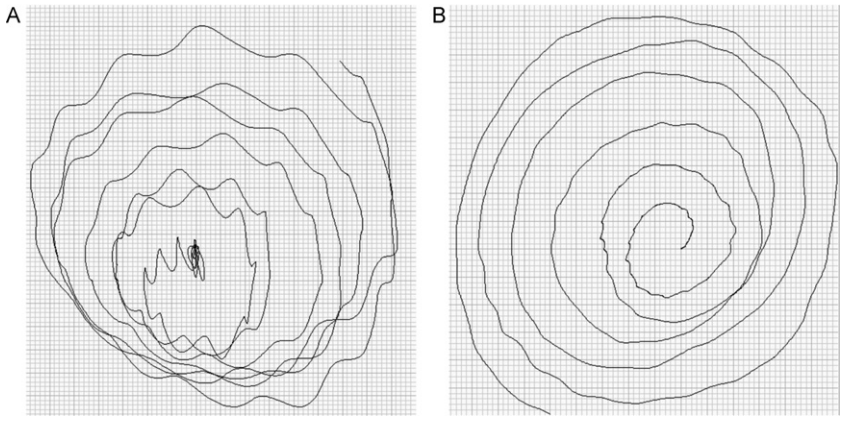

Fig. 1. Spiral drawing of a patient with an essential tremor: (a) left hand, (b) right hand.

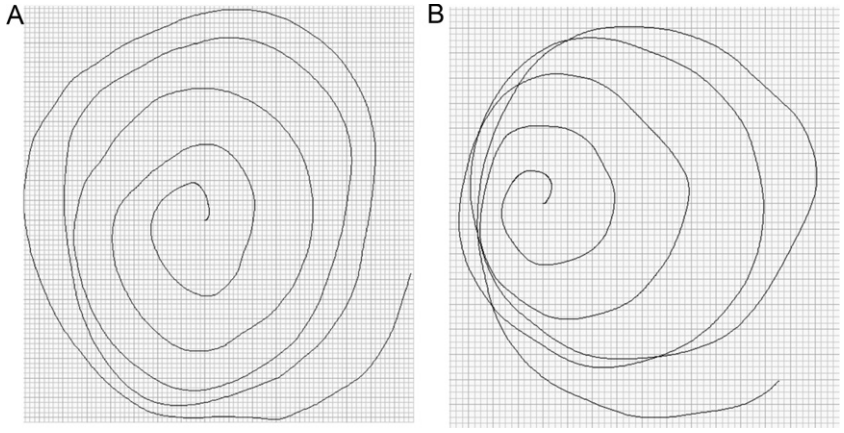

Fig. 2. Spiral drawing of a patient with Parkinsonian tremor: (a) left hand, (b) right hand.

system. Different types of tremor typically have different values for various parameters. For example, Parkinsonian tremor has a frequency between 4 and 6 Hz, and a characteristic radius–angle and speed–time transform [30].

Spiral drawing of a patient with ET can be seen in Fig. 1, of a patient with PT in Fig. 2 and of a patient with MT in Fig. 3. Spiral on the left side is made with a left hand, and on the right side with a right hand.

3.2. DaTSCAN

DaTSCAN is a single photon emission computed tomography of the dopamine transporter (DAT) in the striatum. During the procedure, a radioactive agent (ioflupane (123)I-FP-CIT) is injected in the blood. Ioflupane (123)I-FP-CIT specifically binds to the dopamine transporter on the presynaptic membrane in the striatum. DAT is a transmembrane protein that re-uptakes dopamine from the synaptic cleft into the presynaptic neuron. In PD, because of degeneration of substantia nigra, which projects to the basal ganglia, there is a remarkable loss of DAT activity (as labeled by ioflupane (123)I-FP-CIT) in nucleus caudatus and putamen. On the contrary ioflupane (123)I-FP-CIT DAT activity in ET is normal. Although it has been proven as a useful tool for the differentiation between ET and PD with high sensitivity (93.7%) and specificity (97.3%), the main disadvantages of the method are its high cost and limited access to the method, as it is usually available in bigger hospitals only [31].

4. Methodology

In this section we describe the essential ingredients of our approach: ABML learning, the handling of comorbidities, and the interaction between the expert and the learning program (ABML refinement loop).

4.1. Argument-based machine learning

ABML [8] is machine learning extended with concepts from argumentation. In ABML, arguments are used as means for experts to elicit some of their knowledge through explanations of the learning examples. The experts need to focus on one specific case at a time only and provide knowledge that seems relevant for this case. We will use the ABCN2 [8] method, an argument based extension of the well-known CN2 method [28], that learns a set of unordered probabilistic rules from examples with attached arguments, also called *argumented examples*.[1]

4.2. Handling comorbidities

The problem domain described in this paper contains a class variable with three values: ET, PT, and MT. Since the MT class

[1] Reader can find more about ABML and ABCN2 in [8] and at its website www.ailab.si/martin/abml.

Please cite this article in press as: Groznik V, et al. Elicitation of neurological knowledge with argument-based machine learning. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.08.003

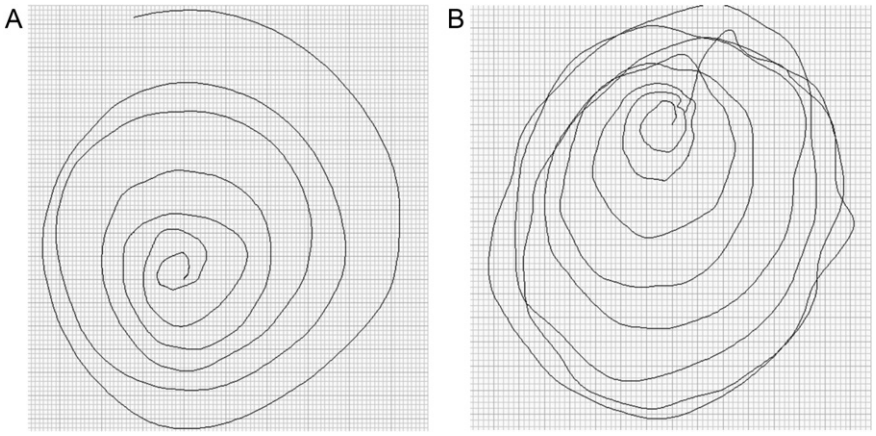

Fig. 3. Spiral drawing of a patient with mixed tremor: (a) left hand, (b) right hand.

implies the presence of essential tremor and Parkinsonian tremor (comorbidity), a rule learning method will have difficulties distinguishing between ET and MT (and likewise between PT and MT). To avert this difficulty, we decided to translate our three-class problem into two two-class problems. In the first, ET and MT are combined in EMT (*essential mixed*) class. All patients in the EMT class contain some signs of essential tremor. The rules for EMT would therefore contain in their conditions features that are indicating essential tremor and are not indicating Parkinsonian tremor. While it is true that EMT contains patients with Parkinsonian tremor (within the MT class), the features indicating Parkinsonian tremor would not be included, as they are not relevant to distinguish between the EMT and PT classes.

In the learning problem with EMT and PT, we learn a set of rules for EMT class only. We skip learning rules for PT, as EMT contains also patients with Parkinsonian tremor, therefore it is unlikely that learning rules for PT would produce good and understandable rules for Parkinsonian tremor. The second two-class problem is analogous to the first one, where PT and MT are combined into PMT (*Parkinsonian mixed*) class. The results of learning in the second problem is a set of rules for PMT class.

To diagnose new cases with the induced rules, we need a mechanism to enable reasoning about new cases. We developed a technique that can infer a classification in one of the three possible classes from induced rules for EMT and PMT classes.

Let e be the example to classify. Let R_{EMT} be a set of rules that cover example e and predict the EMT class. Similarly, rules in R_{PMT} predict PMT class and cover e. Let probability $P(e = EMT)$, where $e = EMT$ is an abbreviation for $class(e) = EMT$, be the predicted class probability of the "best" rule (with highest predicted class probability) from R_{EMT} and $P(e = PMT)$ the predicted class probability of the "best" rule from R_{PMT}. If R_{EMT} (or R_{PMT}) is empty, then $P(e = EMT) = 0$ (or $P(e = PMT) = 0$). Our method will use the following formulae to compute the probabilities for the three classes:

$$P(e = ET) = 1 - P(e = PMT),$$
$$P(e = PT) = 1 - P(e = EMT),$$
$$P(e = MT) = P(e = EMT) + P(e = PMT) - 1,$$

where $P(e = ET)$, $P(e = PT)$, $P(e = MT)$ correspond to predicted probabilities of ET, PT, and MT classes, respectively. In special cases, where the sum $P(e = EMT) + P(e = PMT)$ is less than one, the probability $P(e = MT)$ would be negative. This happens when there are no rules covering example e (R_{EMT} or R_{PMT} is empty). In such cases, we first set the probability $P(e = MT)$ to zero and afterwards normalize the probabilities $P(e = ET)$, $P(e = PT)$, $P(e = MT)$ to sum to one. Example e is classified as the class with the highest predicted probability.

The described mechanism was used during the knowledge elicitation loop presented in the following section. An improved strategy for classification, that can adjust for the change in class distribution between the old and the new data, is described within Section 6.

4.3. Interactions between expert and ABCN2

As asking experts to give arguments to the whole learning set is not feasible, we use the following loop to pick out the *critical examples* that should be explained by the expert.

Step 1: **Learn a hypothesis** with ABCN2 using given data.
Step 2: **Find the "most critical" example** and present it to the expert. If a critical example cannot be found, stop the procedure.
Step 3: **The expert explains the example;** the explanation is encoded in arguments and attached to the learning example.
Step 4: **Return to step 1**.

To finalize the procedure, we need to answer the following two questions: (a) How do we select critical examples? and (b) How can we ensure to get all necessary information for the chosen example?

4.3.1. Identifying critical examples

A critical example is an example the current hypothesis cannot explain well. As our method gives probabilistic class predictions, we will first identify the "most problematic" example, with the highest probabilistic error. To estimate the probabilistic error we used a k-fold cross-validation repeated n times (e.g. $n = 4$, $k = 5$), so that each example is tested n times. The critical example is thus selected according to the following two rules.

1 If the problematic example is from MT, it becomes the critical example.
2 If the problematic example is from the ET (or PT) class, the method will seek out which of the rules predicting PMT (or EMT) is the culprit for the example's misclassification. As the problematic rule is likely to be bad since it covers our problematic example, the critical example will become an example from PT or MT class (or ET or MT) covered by the problematic rule. Then, the expert will be asked to explain what are the reasons that this patient was diagnosed with Parkinsonian tremor (or essential

tremor). Explaining this example should lead to the replacement by the ABCN2 algorithm of the problematic rule with a better one for the PMT (or EMT) class, which hopefully will not cover the problematic example.

4.3.2. Are expert's arguments good or should they be improved?

Here we describe in details step 3 of the above algorithm:

Step 3a: **Explaining critical example.** If the example is from the MT class, the expert can be asked to explain its Parkinsonian and essential signs (which happens when the problematic example is from MT) or to explain only one of the diseases. In other two cases (ET or PT), the expert always explains only signs relevant to the example's class. The expert then articulates a set of reasons suggesting the example's class value. The provided argument should contain a minimal number of reasons to avoid overspecified arguments.

Step 3b: **Adding arguments to example.** An argument is given in natural language and needs to be translated into domain description language (attributes). If the argument mentions concepts currently not present in the domain, these concepts need to be included in the domain (as new attributes) before the argument can be added to the example.

Step 3c: **Discovering counter examples.** Counter examples are used to spot if an argument is sufficient to successfully explain the critical example or not. If not, ABCN2 will select a counter example. A counter example has the opposite class of the critical example, however it is covered by the rule induced from the given arguments.

Step 3d: **Improving arguments with counter examples.** The expert has to revise his initial argument with respect to the counter example.

Step 3e: **Return to step 3c if counter example found.**

5. Knowledge elicitation with ABML

The knowledge elicitation process consisted of 19 iterations. During the process, 17 new attributes were included into the domain. All new attributes were derived from the original attributes and are based on the explanations given by the expert. They are described in Table 2.

5.1. Argumentation of examples from class ET/PT

At the start of the process, only original attributes were used and no arguments have been given yet. Example E.2 (classified as ET in the data set) was the first critical example selected by our algorithm. The expert was asked to describe which features of E.2 are in favor of ET. He selected the following features: resting tremor, rigidity, and bradykinesia, and chose bradykinesia (represented with two attributes in the data set, one for the left side and one for the right side) to be the most influential one of the three features. The expert used his domain knowledge to come up with the following answer: "E.2 is ET because there is no bradykinesia, either on the left nor on the right side." Based on his general knowledge about the domain he also explained that the side (left or right) does not play any particular role in differentiating between ET and PT.

The expert's explanation led the knowledge engineer to introduce new attribute BRADYKINESIA with possible values *true* (bradykinesia is present on the left side *or* on the right side) and *false* (bradykinesia was not indicated on either side). At the same time the original two attributes were excluded from the domain – it is their combination (reflected in the expert's argument) that provides relevant information according to the expert.

Based on the expert's explanation, argument "BRADYKINESIA is *false*" was added as the argument for ET to the critical example E.2. No counter examples were found by the method and thus the first iteration was concluded. New rules were induced before entering the next iteration. One of the notable changes was that the following rule appeared:

IF BRADYKINESIA = *false* THEN *class* = EMT;

The rule covers 20 learning examples, and all of them are from class ET.

5.2. Argumentation of examples from class MT

In the previously described iteration the critical example E.2 was classified as purely ET by the neurologist. In one of the following iterations, however, the critical example E.61 was classified as both PT and ET. In such a case, the expert is asked to describe which features are in favor of ET *and* which features are in favor of PT. The expert explained that the presence of postural tremor speaks in favor of ET, while the presence of rigidity speaks in favor of PT. Again he relied on his general knowledge to advocate that

Table 2
The new attributes derived from the original ones during the knowledge elicitation process and their rate of missing values. These attributes are based on the explanations given by the expert.

Attribute	Description	Number of missing values, *n* (%)
Sim.tremor.start	Bilateral tremor start	0(0.00)
Diff.age.tremor.duration	Difference between the age of the patient and tremor duration	14(12.28)
Diff.disease.tremor.duration	Difference between the duration of the disease and duration of the tremor	59(51.75)
Sim.resting.tremor.up	Bilaterally equal resting tremor on the upper limbs	6(5.26)
Sim.postural.tremor.up	Bilaterally equal postural tremor on the upper limbs	10(8.77)
Sim.rigidity.up	Bilaterally equal rigidity of the upper limbs	7(6.14)
Sim.bare.speed.time	Bilaterally equal speed–time transform-bare hand drawing	21(18.42)
Sim.bare.radius.angle	Bilaterally equal radius–angle transform-bare hand drawing	23(20.18)
Sim.template.speed.time	Bilaterally equal speed–time transform-template drawing	45(39.47)
Sim.template.radius.angle	Bilaterally equal radius–angle transform-template drawing	49(42.98)
Bradykinesia	At least one sided bradykinesia	5(4.40)
Resting.tremor.up	At least one sided resting tremor	6(5.26)
Postural.tremor.up	At least one sided postural tremor	10(8.77)
Rigidity.up	At least one sided rigidity	7(6.14)
Harmonics	At least one harmonic at any condition	10(8.77)
Spiro.Parkinsonian.only	All spirography data are parkinsonian	0(0.00)
Spiro.Essential.only	All spirography data are essential	0(0.00)

Please cite this article in press as: Groznik V, et al. Elicitation of neurological knowledge with argument-based machine learning. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.08.003

Table 3
Class distributions of learning and testing data.

	Learning data (47)			Testing data (67)		
	Essential	Mixed	Parkins.	Essential	Mixed	Parkins.
n	22	13	12	28	10	29
Proportions	46.81%	27.66%	25.53%	41.79%	14.93%	43.28%

keeping separate attributes for both the left and the right side does not have any impact on deciding between ET and PT, and suggested one attribute for each feature instead.

Attributes POSTURAL.TREMOR.UP and RIGIDITY.UP were introduced into the domain instead of the original ones that describe features postural tremor and rigidity. The former was used as an argument for ET and the latter was used as an argument for PT – both of these arguments were added to the critical example E.61. While no counter examples were found for the expert's argument in favor of ET, the method selected E.45 (ET) as a counter example for his argument in favor of PT.

The expert was now asked to compare the critical example E.61 with counter example E.45, and to explain what is the most important feature in favor of PT that applies for E.61 and does *not* apply for E.45. According to the expert's judgement, it was the presence of harmonics in E.45 (or their absence in E.61), which are typical of ET. The attribute HARMONICS that was added into the domain earlier with possible values of *true* and *false* was added to the previous argument. However, the method then found another counter example, E.30 (ET). The expert explained that the tremor in E.30 did not have symmetrical onset, as opposed to the one in the critical example. The argument was further extended using the attribute SIM.TREMOR.START and added to the critical example E.61. No new counter examples were found and this particular iteration was therefore concluded.

5.3. Improving on the arguments

There are three possible ways for the expert's arguments to be improved: (1) by the expert, with the help of counter examples selected by the method, (2) by the method alone, and (3) by the expert, upon the observation of induced rules. The first option was covered in the previous subsection. In the sequel, the latter two options are described.

In the first iteration (as described in Section 5.1), the expert's arguments proved to be sufficient for the method to induce rules with clear distributions. Sometimes, however, the method automatically finds additional restrictions to improve the expert's argument. Such was the case in one of the following iterations, where the following argument occurred: "RESTING.TREMOR.UP IS *true* and HARMONICS is *false* and SIM.TREMOR.START is *false*." The following rule that also occurs in the final model was induced with the help of this argument:

IF RESTING.TREMOR.UP = *true* AND HARMONICS = *false*
AND SIM.TREMOR.START = *false* **AND** SPIRO.ESSENTIAL.ONLY **= *false***
THEN class = PMT;

The method automatically improved on the expert's argument by adding an additional restriction in the above rule. The attribute SPIRO.ESSENTIAL.ONLY was introduced by the expert in one of the previous iterations. Its meaning is the following: if qualitative assessment of the spiral in any of the eight observations (attributes) in the original data is essential, and none of them is Parkinsonian (or any other), then the value of SPIRO.ESSENTIAL.ONLY is *true*, otherwise it is *false*. The above rule covers 14 examples (all of them from class ET) and was particularly praised by the expert – one of the reasons for this being that it effectively combines clinical data with the results of spirography.

The third possibility occurred only once in the ABML knowledge elicitation process presented in this paper. Upon the final examination of the rules the expert approved all the obtained rules but the following one:

IF POSTURAL.TREMOR.UP = *true* AND SIM.BRADYKINESIA = *true*
THEN class = EMT;

Although the rule covers 23 examples (out of 47) and has a clear distribution, the expert found the attribute SIM.BRADYKINESIA meaningless. This was the automatically induced part of the rule from the expert's argument to the example E.61, as described in Section 5.2. Based on the expert's explanation this argument was now extended to "POSTURAL.TREMOR.UP is *true* and BRADYKINESIA = *false*." Such changes should not by any circumstances be made *after* examining results on the testing data, and it is particularly important that the expert relies on his common knowledge of the domain when doing this. The following rule was induced from the expert's argument:

IF POSTURAL.TREMOR.UP = *true*
AND BRADYKINESIA = *false*
THEN class = EMT;

Although the rule has notably worse coverage, the expert found it consistent with his domain knowledge. At this point, the expert approved all the rules in the final model and thus the iterative process was concluded.

6. Evaluation setup

6.1. Data

The data was gathered in two batches; the first part contained 67 patients and the second, which we received a few months later, contained 55 patients. When the first part arrived, we made a stratified split of the 67 patients into 47 patients used for learning the model and the remaining 20 patients to evaluate it. The results of that experiment can be found in [32].

When we received the second part of the data, we decided to add these examples to the initial 20 testing examples to enhance the evaluation (and make the results more significant). Of the new 55 cases only 47 could be used, as 8 of them were invalid due to missing values of all clinical attributes. These 47 were thus added to the initial 20, resulting in 67 testing examples in total.[2]

After examination of the new data, we noticed a significant difference in class distributions. The figures in Table 3 show a large increase of patients with Parkinsonian disease among the newly obtained data. The subsequent analysis revealed that the initial data had a skewed distribution, because we used the cases readily available to our expert. These cases were mostly more difficult to diagnose correctly. The second set consists of unselected

[2] These 67 patients used for testing should not be confused with the initially obtained 67 patients. It is by pure chance that the two numbers are the same.

Table 4
The list of parameters used for tuning each method in the internal optimization process and their possible values.

Method	Parameter	Values
NB	*m*-Estimate	2, 5, 10, 25, 50, 100
DT	Measure	Information gain, Gini Index, Gain ratio
	m-Estimate (post-pruning)	2, 5, 10, 25, 50, 100
	Min. instances in leaves (pre-pruning)	2, 3, 5, 10, 20
RF	Number of trees	10, 25, 50, 100
	Min. instances in leaves	2, 3, 5, 10
SVM	gamma	$1/114, 2^x$ where $x \in [-5, -3, -1, 1, 3, 5]$
	nu	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Stepwise LR	Imputation	Average
	Add criteria (probability of F)	0.01, 0.05, 0.1
	Remove criteria (probability of F)	0.1, 0.2, 0.5, 1.0

consecutive patients, and their distribution is as usually observed with respect to the incidence of the involved tremors; the rate of mixed tremors is at about 15%, and the rest are equally spread between the essential and Parkinsonian tremors.

6.2. Evaluation setup

The difference in distributions did not allow for the standard evaluation technique, where a model is learned on the learning data and tested on the testing data. Furthermore, we were unable to run standard cross-validation, since the 47 learning examples were already used within the ABML refinement loop and hence could not be used as testing cases. To carry out the standard k-fold cross-validation, the complete ABML refinement loop would have to be repeated for each fold, with new expert's arguments each time.

We used a variant of cross-validation where 67 testing examples were split into 10 folds using stratified sampling. In each iteration the original 47 learning examples coupled with 9 folds of the testing examples thus constituted the complete learning data, and the learned model was tested on the remaining fold. Consequently, the 47 learning examples were never used during testing.

With the described technique we evaluated ABCN2, naive Bayes classifier (NB), decision trees (DT), random forests (RF), SVM and stepwise logistic regression (LR). The measures used were classification accuracy (CA), area under the curve (AUC), sensitivity (sens.) and specificity (spec.).

In the experiments, we used the implementations of the above methods as provided by the data mining package *Orange*.[3] For SVM we used Gaussian radial basis function.

All of the methods, except for ABCN2, were tuned using internal cross-validation as follows. In each iteration of the aforementioned cross-validation the complete learning data (47 learning examples coupled with 9 folds) were in turn split into 10 folds for internal cross-validation, using stratified sampling. In each iteration we learned the optimal set of parameter settings for each method on 9 "internal" folds, and tested the method with this setting on the remaining fold. In this way we obtained 10 best sets of parameter values. The setting giving the highest CA was used to learn a model on all the learning examples, i.e. 47 learning examples plus 9 "external" folds. The learned model was tested on the remaining "external" fold. Parameters used in the optimization process can be found in Table 4. For tuning the parameters we used the *Wrap* method from *Orange* package.

[3] Reader can find more about *Orange* at its website http://orange.biolab.si/.

6.3. Tuning the decision model of ABCN2

Using new data, our goal was to evaluate the final rules obtained with ABCN2 at the end of the elicitation process. As explained above, it would be inappropriate to use all the examples (47 plus 9 folds) for ABCN2 learning. Therefore, we used only the original 47 examples for ABCN2 learning, while the remaining data from 9 folds were used to (a) update the classification accuracy estimates of the learned rules to account for new class distribution and (b) improve the decision model from Section 4.2.

The classification accuracies of the rules were estimated using the *m*-estimate [33] formula where parameter *m* was set to 2 (that is the default setting for *m* in ABCN2). The *m*-estimate of CA of a rule is:

$$CA(rule) = \frac{s + m \times p_a}{n + m}, \quad (1)$$

where s is the number of positive examples covered by the rule, p_a is the prior probability of the predicted class, and n is the number of all examples covered by the rule. During learning these accuracies were estimated on the original learning data (47 examples). To account for the change in class distributions, the values s, p_a, and n were revised using only the remaining part of the learning data (9 folds) and the accuracy estimates of the rules were reevaluated.

With respect to improving the decision model, we used the PILAR [34] algorithm to estimate probabilities $P(e = EMT)$ and $P(e = PMT)$. This algorithm is known to produce better probabilistic predictions with ABCN2 than other methods, e.g. using the most accurate rule only. The idea of PILAR is to assign weights to each rule while taking correlations between rules into account. The probability of a class for a certain example is therefore obtained by summing the weights of all the rules covering this example and computing the probability through the logit transformation. After the probabilities $P(e = EMT)$ and $P(e = PMT)$ are returned by the algorithm, the formulae from Section 4.2 are used to compute probabilities $P(e = ET)$, $P(e = MT)$, and $P(e = PT)$, and to make a decision as to the correct diagnosis.

7. Results

7.1. Quantitative comparison

Table 5 presents the results of applying different machine learning techniques: ABCN2, NB, DT, SVM, RF and stepwise LR. The numbers represent classification accuracy, area under curve, specificity and sensitivity before and after knowledge elicitation process. We have also included the confidence intervals for CA and AUC.

It is worth noting again that the non-standard 10-fold cross-validation (described earlier) was used for standard machine

Table 5
The results of applying different machine learning techniques before and after knowledge elicitation process.

			ABCN2	NB	DT	SVM	RF	LR
	CA	Before	0.82 ± 0.09	0.82 ± 0.13	0.60 ± 0.15	0.71 ± 0.09	0.76 ± 0.07	0.76 ± 0.09
		After	0.91 ± 0.09	0.88 ± 0.09	0.64 ± 0.16	0.78 ± 0.09	0.78 ± 0.11	0.78 ± 0.09
	AUC	Before	0.95 ± 0.06	0.98 ± 0.02	0.77 ± 0.15	0.92 ± 0.06	0.93 ± 0.04	0.89 ± 0.09
		After	0.96 ± 0.08	0.98 ± 0.03	0.79 ± 0.16	0.94 ± 0.06	0.93 ± 0.06	0.91 ± 0.07
ET	Spec.	Before	0.90	0.85	0.46	0.77	0.59	0.90
		After	0.97	0.87	0.46	0.77	0.62	0.82
	Sens.	Before	0.86	0.93	0.89	0.75	1.00	0.82
		After	0.93	1.00	0.89	0.86	1.00	0.75
PT	Spec.	Before	0.87	0.95	0.97	0.82	1.00	0.82
		After	0.89	0.97	0.97	0.95	1.00	0.92
	Sens.	Before	0.86	0.83	0.41	0.79	0.72	0.83
		After	0.97	0.86	0.48	0.83	0.72	0.86
MT	Spec.	Before	0.95	0.93	0.91	0.93	1.00	0.91
		After	0.98	0.96	0.95	0.93	1.00	0.91
	Sens.	Before	0.60	0.50	0.30	0.30	0.20	0.40
		After	0.70	0.60	0.30	0.40	0.30	0.60

learning techniques: learning with fixed 47 examples and nine folds of the remaining 67 cases, and testing on the remaining fold.

The initial (before elicitation) classification accuracy for ABCN2 is 0.82, which is comparable to CA of NB (0.82). This result is further supported with high AUC score of 0.95. Specificity and sensitivity for each class (ET, PT and MT) are above 0.80, except for MT, which has sensitivity 0.60.

The CA of the final ABCN2 model is 0.91. The AUC has slightly increased to 0.96 and similar applies to specificity and sensitivity.

To compare ABCN2 with the other methods we tested for statistical significance (at the 0.05 level) of the difference in CA (independent sample *t*-test) and the difference in AUC (Mann–Whitney–Wilcoxon rank-sum test) for the results after the knowledge elicitation process. There was no significant difference in AUC, except for decision trees. There was, however, significant difference in CA between ABCN2 and all the methods, except NB. On the basis of this, we believe that ABNC2 performs at least on par with the best competing method in diagnostic accuracy.

We also measured the net time investment of the domain expert. It was slightly more than 20 h. The knowledge engineers spent approximately 150 h total.

7.2. Qualitative comparison of the initial and the final model

There were 13 rules at the end of the knowledge elicitation process. The final model is given in Table 6. Each of the rules was evaluated independently by two neurologists (other than our expert in the knowledge elicitation process). They found all the rules to correctly indicate the predicted class. In the sequel, we will present the experts' explanations of three of these rules.

Nine rules contain attributes dealing with spirography tests, and three of those are based exclusively on spirography. The following rule is one of them:

IF HARMONICS = *true* THEN *class* = EMT;

The rule states that if there are harmonic frequencies in the tremor frequency spectra, then the tremor is essential. It is known that the appearance of harmonic frequencies is very specific for ET.

The following rule in the final model is based solely on the attributes of clinical examination:

IF BRADYKINESIA = *true*
  AND RIGIDITY.UP = *true*
  AND RESTING.TREMOR.UP = *true*
THEN *class* = PMT;

This says that if a patient has a resting tremor, bradykinesia, and rigidity, the tremor is Parkinsonian. This rule, namely the

Table 6
Rules in the final model.

Rule number	Rule
1	IF Bradykinesia = false THEN class = EMT
2	IF Qualitative.spiral = essential THEN class = EMT
3	IF Harmonics = true THEN class = EMT
4	IF Spiro.Parkinsonian.only = false AND Postural.tremor.up.left > 0 THEN class = EMT
5	IF History = positive AND Bare.right.freq.range > 5 THEN class = EMT
6	IF Qualitative.spiral = Parkinsonian THEN class = PMT
7	IF Bradykinesia = true AND Rigidity.up = true THEN class = PMT
8	IF Bare.right.speed.time = Parkinsonian AND Tremor.neck ≤ 0 THEN class = PMT
9	IF Rigidity.up = true AND Harmonics = false AND Tremor.start = right side THEN class = PMT
10	IF Bradykinesia = true AND Rigidity.up = true AND Resting.tremor.up = true THEN class = PMT
11	IF Resting.tremor.up = true AND Harmonics = false AND Spiro.Essential.only = false AND Sim.tremor.start = false THEN class = PMT
12	IF Bradykinesia = true AND Diff.age.tremor.duration ≤ 60 years AND Harmonics = false AND Sim.tremor.start = false THEN class = PMT
13	IF Postural.tremor.up = true AND Bradykinesia = false THEN class = EMT

Table 7
Rules in the initial model.

Rule number	Rule
1	IF Qualitative.spiral = essential THEN class = EMT
2	IF Bare.right.speed.time = essential THEN class = EMT
3	IF Bradykinesia.right ≤ 0 THEN class = EMT
4	IF Rigidity.up.right ≤ 0 AND Bradykinesia.left ≤ 1 THEN class = EMT
5	IF Rigidity.up.right ≤ 1 AND Resting.tremor.up.left ≤ 2 AND Postural.tremor.up.left > 0 AND Disease.duration ≤ 9 years THEN class = EMT
6	IF Bare.right.speed.time = Parkinsonian AND Disease.duration ≤ 12 years THEN class = PMT
7	IF Rigidity.up.right > 0 AND Age ≤ 83 years THEN class = PMT
8	IF Rigidity.up.right > 0 AND Disease.duration ≤ 12 years THEN class = PMT
9	IF Bradykinesia.right > 0 AND Age ≤ 74 years THEN class = PMT

combination of a resting tremor, bradykinesia and rigidity actually clinically defines PT.

Finally, we present a rule that successfully combines the knowledge from spirography and clinical examination:

IF RESTING.TREMOR.UP = *true*
AND HARMONICS = *false*
AND SPIRO.ESSENTIAL.ONLY = *false*
AND SIM.TREMOR.START = *false*
THEN *class* = PMT;

The rule defines an indication of PT by the presence of resting tremor, and the lack of harmonic frequencies in the tremor frequency spectra, while not all spirography data are essential, and a non-bilateral tremor start occurred. All this was found to be in accordance with the knowledge of the experts.

In the initial model, before the beginning of the knowledge elicitation process, there were 9 rules (see Table 7). Five of these rules used the duration of the disease or age at disease onset. However, both of these attributes are not very informative according to the expert. Both disorders can start at any age. While ET typically occurs earlier, several patients with ET tend to visit a neurologist only many years after the occurrence of the disease, and therefore their age at onset and duration of the disease are sometimes not recorded properly.

In contrast to the final model, several rules in the initial model were found to be senseless from the medical point of view. The next two examples illustrate this. Let us take a closer look at the following rule in the initial model:

IF RIGIDITY.UP.RIGHT >0
AND AGE ≤83 years
THEN *class* = PMT;

The first part states correctly that if the rigidity is greater in the right upper extremity, this is a sign of PT. However, according to the expert the side (left or right) does not play any particular role for differentiating between ET and PT. The second condition is incorrect. There is no upper limit for the age at disease onset for either PT or ET. Here is another rule in the initial model:

IF RIGIDITY.UP.RIGHT ≤1
AND RESTING.TREMOR.UP.LEFT ≤2
AND POSTURAL.TREMOR.UP.LEFT >0
AND DISEASE.DURATION ≤9 years
THEN *class* = EMT;

According to the expert, a positive value of RIGIDITY.UP.RIGHT in general speaks in favor of PT. Similarly, it is commonly accepted that a positive value of RESTING.TREMOR.UP.LEFT also speaks in favor of PT. However, the above rule obviously does not distinguish between the cases in which the values of these two attributes are positive or not. Moreover, as we explained above, ET typically occurs earlier than PT, therefore a symbol ≥ instead of ≤ would be more logical, if any.

Table 8
The confusion matrix for the final model.

True class	Predicted class		
	ET	MT	PT
ET	26	1	1
MT	0	7	3
PT	1	0	28

7.3. Misclassification analysis

Using the rules of our final model, 6 out of 67 cases in the test data set were misclassified (see Table 8). We asked our domain expert to examine the misclassified cases and the rules responsible for their classification.

After precise evaluation, the expert actually agreed with two of the computer's classifications as the neurologist overlooked some of the details at the time of diagnosis. Therefore the class of these two cases should actually be modified.

Moreover, the expert changed the class of another case. In this case he did not agree with the computer's evaluation, but the rules nevertheless helped the expert to spot the earlier mistake.

The domain expert was also asked to examine 12 misclassified cases produced by the rules of the initial model (see Table 9). Although the expert agreed with two of the computer's classifications and the class was changed on two more occasions after careful examination, three misclassified cases of the worst type remained: PT was wrongly classified as ET twice and ET was wrongly classified as PT once.

8. Discussion

The evaluation suggests that the ABML knowledge elicitation process resulted in improved diagnostic accuracy. The classification is better after the elicitation process both in ABML and in the other machine learning methods used in the comparison. The performance of ABML in terms of CA and AUC is at least comparable to the performance of the other state-of-the-art methods involved in the comparison. However, in our view the main result

Table 9
The confusion matrix for the initial model.

True class	Predicted class		
	ET	MT	PT
ET	24	2	2
MT	1	6	3
PT	3	1	25

Please cite this article in press as: Groznik V, et al. Elicitation of neurological knowledge with argument-based machine learning. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.08.003

of ABML knowledge elicitation is the comprehensibility and the medical meaningfulness of the obtained knowledge which do not come at the cost of diagnostic accuracy. The comparison of the comprehensibility between the initial and the final set of rules clearly demonstrates the effect of knowledge elicitation.

We believe comprehensibility is the crucial aspect that makes the final set of rules appropriate for use in a decision support system. It is important for the explanation supporting the suggested decisions (diagnoses), especially so in the medical domains where treatment is based on diagnosis. Apart from boosting confidence in the suggested diagnosis, well-formulated explanations can help the doctors spot their own potential mistakes, or prompt them to rethink the diagnosis. As such, the DSS can really act as a second opinion, especially for the harder cases.

Another benefit of comprehensibility is that a DSS can easily be turned into a valuable teaching tool. For this it is equally if not even more important that the explanations are sensible and correct. The evaluation of our models by the neurologists clearly confirms the benefit of ABML in this respect; while initial knowledge was hardly comprehensible and it was sometimes even illogical from the medical point of view (usually an artifact of chance), the final set of rules was much more textbook like.

The analysis of the confusion matrices for the initial and final models does not suggest a significant decrease in the severity of errors made. As all the misdiagnoses are not equally harmful, cost-sensitive learning to tune the DSS is one obvious improvement to think of in the future.

A further look at the final set of rules reveals another interesting result. Of the 13 rules, nine contain attributes dealing with spirography tests, and of those nine, three are based exclusively on spirography. It is important to note that during the knowledge elicitation there was no special incentive to use specifically spirographic data. This suggests that spirography is very useful for the task at hand. The three rules also work very well on eight patients that had no clinical data.

This warrants further work on spirography, perhaps to use as an early screening method, even remotely on gadgets like the iPad or smart phones.

9. Conclusions and further work

The paper detailed some aspects of building a decision support system for diagnosing and differentiating between three types of tremors. Our DSS also takes into account the information from spirography which was not used in previous work on this problem. According to our results, spirography provides valuable diagnostic information. After carrying out the ABML refinement loop, the accuracy of the system improved over the initial model on our test set as well as the specificity and sensitivity for each class. There has also been a slight decrease in severity of misdiagnoses. As new patients will constantly be enrolled into the study we will be able to precisely quantify the accuracy of the system in the long run.

We have also measured the net time involvement of the expert in building a knowledge base for the system. We believe ABML saves a significant amount of expert's time, and the expert agreed that the process itself felt very natural and stimulating. However, it is very difficult to make a fair comparison with other methods, and we resolved to just stating the net times measured.

As already mentioned, our long-term goal is to build a DSS able to act as a second opinion, and a valuable teaching tool. To this end, the obvious future work is to redo the whole learning process on a much larger scale, taking into account all we learned from this pilot project. The plan is also to extend the system to other types of tremors.

The other major topic for further work is to extensively evaluate and validate the system. We plan to enroll a large number of patients into the study, where the patients will routinely undergo the DatSCAN examination (also for other purposes than our study).

Spirography was revealed to have real potential for tremor diagnosis, in conjunction with clinical data, but also as a stand-alone early screening method. This possibility should also be investigated.

Acknowledgment

The work was partly funded by the Slovenian Research Agency (ARRS).

References

[1] Deuschl G, Wenzelburger R, Loeffler K, Raethjan J, Stolze H. Essential tremor and cerebellar dysfunction: clinical and kinematic analysis of intention tremor. Brain 2000;123(8):1568–80.
[2] Thanvi B, Lo N, Robinson T. Essential tremor – the most common movement disorder in older people. Age and Ageing 2006;35(4):344–9.
[3] Quinn N, Schneider S, Schwingenschuh P, Bhatia K. Tremor-some controversial aspects. Movement Disorders 2011;26(1):18–23.
[4] Miotto GAA, Andrade AO, Soares AB. Measurement of tremor using digitizing tablets. In: V Conferencia de Estudos em Engenharia Eletrica (CEEL). 2007.
[5] Kraus P, Hoffmann A. Spiralometry: computerized assessment of tremor amplitude on the basis of spiral drawing. Movement Disorders 2010;25(13):2164–70.
[6] Hughes A, Daniel S, Kilford L, Lees A. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. Journal of Neurology, Neurosurgery and Psychiatry 1992;55(3):181–4.
[7] Pahwa R, Lyons KE. Essential tremor: differential diagnosis and current therapy. American Journal of Medicine 2003;115:134–42.
[8] Mož ina M, Žabkar J, Bratko I. Argument based machine learning. Artificial Intelligence 2007;171(10/15):922–37.
[9] Mož ina M, Guid M, Krivec J, Sadikov A, Bratko I. Fighting knowledge acquisition bottleneck with argument based machine learning. In: Ghallab M, Spyropoulos CD, Fakotakis N, Avouris NM, editors. Proceedings of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence, vol. 178 of Frontiers in Artificial Intelligence and Applications. Amsterdam, The Netherlands: IOS Press; 2008. p. 234–8.
[10] Feigenbaum EA. The art of artificial intelligence: I. Themes and case studies of knowledge engineering. Tech. Rep., Stanford, CA, USA; 1977.
[11] Giuse DA, Giuse NB, Miller RA. Towards computer-assisted maintenance of medical knowledge bases. Artificial Intelligence in Medicine 1990;2(1):21–33.
[12] Bainbridge L. Asking questions and accessing knowledge. Future Computing Systems 1986;1(2):143–9.
[13] Neale IM. First generation expert systems: a review of knowledge acquisition methodologies. The Knowledge Engineering Review 1988;3(2):105–45.
[14] Boose JH. A survey of knowledge acquisition techniques and tools. Knowledge Acquisition 1989;1(1):3–37.
[15] Cooke NJ. Varieties of knowledge elicitation techniques. International Journal of Human–Computer Studies 1994;41(6):801–49.
[16] Yang HL. Information/knowledge acquisition methods for decision support systems and expert systems. Information Processing & Management 1995;31(1):47–58.
[17] Feigenbaum EA. Some challenges and grand challenges for computational intelligence. Journal of the ACM 2003;50:32–40.
[18] Okafor EC, Osuagwu CC. The underlying issues in knowledge elicitation. Interdisciplinary Journal of Information, Knowledge, and Management 2006;1:95–107.
[19] Compton P, Edwards G, Kang B, Lazarus L, Malor R, Preston P, et al. Ripple down rules: turning knowledge acquisition into knowledge maintenance. Artificial Intelligence in Medicine 1992;4(6):463–75.
[20] Kang BH, Compton P, Preston P. Multiple classification ripple down rules: evaluation and possibilities. In: Gaines B, Musen M, editors. Proceedings 9th Banff knowledge acquisition for knowledge based systems workshop, Banff. 1995. p. 17.1–20.
[21] Edwards G, Compton P, Malor R, Srinivasan A, Lazarus L. Peirs: a pathologist-maintained expert system for the interpretation of chemical pathology reports. Pathology 1993;25(1):27–34.
[22] Forsyth R, Rada R. Machine learning – applications in expert systems and information retrieval. Ellis Horwood series in artificial intelligence Ellis Horwood; 1986.
[23] Langley P, Simon HA. Applications of machine learning and rule induction. Communications of the ACM 1995;38(11).
[24] Ting S, Wang W, Tse Y, Ip W. Knowledge elicitation approach in enhancing tacit knowledge sharing. Industrial Management & Data Systems 2011;111(7):1039–64.

[25] Webb GI, Wells J, Zheng Z. An experimental evaluation of integrating machine learning with knowledge acquisition. Machine Learning 1999;35(1):5–23.
[26] Lavrač N, Džeroski S, Grobelnik M. Learning nonrecursive definitions of relations with LINUS. In: Kodratoff Y, editor. Machine learning – EWSL-91, vol. 482 of Lecture notes in computer science. Berlin/Heidelberg: Springer; 1991. p. 265–81.
[27] Lavrač N, Džeroski S. Weakening the language bias in LINUS. Journal of Experimental & Theoretical Artificial Intelligence 1994;6(1):95–119.
[28] Clark P, Boswell R. Rule induction with CN2: some recent improvements. In: Kodratoff Y, editor. EWSL, vol. 482 of Lecture notes in computer science. Berlin/Heidelberg: Springer; 1991. p. 151–63.
[29] Lavrač N, Džeroski S, Pirnat V, Križman V. The utility of background knowledge in learning medical diagnostic rules. Applied Artificial Intelligence 1993;7:273–93.
[30] Grimaldi G, Manto M. Tremor from pathogenesis to treatment; chap. Characterization of tremor. In: Synthesis lectures on biomedical engineering. Morgan & Claypool Publishers; 2008. p. 39–51.
[31] Towey DJ, Bain PG, Nijran KS. Automatic classification of 123i-fp-cit (DaTSCAN) spect images. Nuclear Medicine Communications 2011;32(8):699–707.
[32] Groznik V, Guid M, Sadikov A, Mož ina M, Georgiev D, Kragelj V, et al. Elicitation of neurological knowledge with ABML. In: Peleg M, Lavrač N, Combi C, editors. Artificial Intelligence in Medicine, vol. 6747 of Lecture notes in computer science. Berlin/Heidelberg: Springer; 2011. p. 14–23.
[33] Cestnik B. Estimating probabilities: a crucial task in machine learning. In: Aiello LC, editor. Proceedings of the ninth European Conference on Artificial Intelligence. London: Pitman; 1990. p. 147–9.
[34] Mož ina M. Argument based machine learning. Ph.D. thesis. University of Ljubljana, Faculty of Computer and Information Sciences, Ljubljana, Slovenia; 2009.

Please cite this article in press as: Groznik V, et al. Elicitation of neurological knowledge with argument-based machine learning. Artif Intell Med (2012), http://dx.doi.org/10.1016/j.artmed.2012.08.003