Book Title
Book Editors
IOS Press, 2003

Argument based machine learning in medical domain

Jure Žabkar [a,1], Martin Možina [a], Jerneja Videčnik [b] and Ivan Bratko [a]
[a] *Faculty of Computer and Information Science, University of Ljubljana, Slovenia*
[b] *Clinic for Infectious Diseases, Ljubljana, Slovenia*

Abstract. Argument Based Machine Learning (ABML) is a new approach to machine learning in which the learning examples could be accompanied by arguments. The arguments for specific examples are a special form of expert's background knowledge which the expert uses to substantiate the class value for the chosen example. Možina et al. developed ABCN2 algorithm - an extension of a well known rule learning algorithm CN2 - that can use argumented examples in the learning process. In this work we present an application of ABCN2 in the medical domain which deals with severe bacterial infections in geriatric population. The elderly population, people over 65 years of age, is rapidly growing and it is estimated that it will double in the next 30 years. In Slovenia, it was accounted for over 15% in 2004 which is nearly the same as in USA and other developed countries. The costs of treating the patients aged over 65 are growing rapidly as well. In our study, we compare ABCN2 to CN2 and show that using arguments we improve the characteristics of the model. We also report the results that C4.5, Naïve Bayes and Logistic Regression achieve in this domain.

Keywords. Argument Based Machine Learning, Rule learning, Geriatric population, Bacterial infections

1. Introduction

The elderly population is one of a kind and that is also true from the medical perspective. Compared to younger population, people over 65 years of age usually react to a disease in a different way. Many symptoms may not even present or they are masked by others which makes it a very difficult task for a medical doctor to diagnose a condition, to decide a proper treatment or to estimate the patient's risk of death. From a wider perspective, the proportion of elderly in the population is growing rapidly and so are the costs for medical treatment, which presents an emerging economic problem.

Infections in the aging population present an increasing problem in the developed countries. Many patients that present with infection have asociated chronic diseases such as diabetes, heart, lung, kidney, lung or liver disease which makes the treatment even more complicated. The number of nursing home residents is also increasing in this population. Because of the specific living environment, these people are usually more sus-

[1]Correspondence to: Jure Žabkar, Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, SI-1001 Ljubljana, Slovenia. Tel.: +386 1 4768 299; Fax: +386 1 4768 386; E-mail: jure.zabkar@fri.uni-lj.si.

ceptible to bacterial infections. Despite great progress in treating infectious diseases they remain one of the major causes of death in geriatric population. Some diferences in the course of illness can be observed comparred to younger patients. Greater risk for severe bacterial infection is due to the patient being immunocompromised [1], immobile, nursing home resident or comorbidity. In elderly, the infections often present with untypical signs, such as the absence of fever [13,3,11,14], the absence of cough at pneumonia and weakness or changed mental status [21]. These usually cause a delay in making a right diagnose. A proper and efficient antimicrobial treatment is often given too late, and the risk for fatal outcome is increased [10,19].

The motivation for using machine learning in the study is to build a model from data which would help the physician, at the first examination of the patient, to decide the severity of the infection and consequently, whether the patient should be admited to the hospital or could be treated as an outpatient. More, we would like to have an understandable model, not a black box, to see which parameters play a decisive role. Several studies are described in the literature regarding the difficulty of the course of disease [9,7,8]. Fine et al. [8] implemented a prognostic model for adult patients with documented bacterial pneumonia. An overall study, regarding the bacterial infections of different organs and taking into account so many clinical as well as laboratorial parameters has, to our knowledge, not been carried out yet.

The alternative to machine learning would be to implement with the help of domain experts an expert system and use it for diagnose the severeness of infection. The knowledge possessed by experts is usually implicit and they find it extremely difficult to elicit in the form of a set of rules. On the other hand, it was shown that experts can rather easy discuss a certain case, instead of giving a general theory. Therefore, the research from defeasible argumentation [20] proposed an alternative approach to building expert systems. The experts should first give arguments to some specific examples for all possible outcomes. These arguments are then given to an argumentation engine, an expert system that can use these (possible contradictive) arguments to make predictions for new examples, and whenever a decision from the arguments could not be inferred, the experts are again asked for more arguments.

Our approach, Argument Based Machine Learning (ABML) [2,17], combines machine learning and argumentation. ABML is a new approach to machine learning in which the learning examples could be accompanied by arguments. The expert chooses a subset of learning examples and gives reasons in form of arguments, why the class value of the example is as given. We have developed an argument-based rule learning algorithm ABCN2, an extension of the well-known CN2 algorithm, which we applied to medical domain hoping to improve the prediction quality of standard machine learning techniques by using arguments given by experts.

2. Argument Based Machine Learning

Argument Based Machine Learning (ABML) [17,2] is a new approach to machine learning that can learn from examples and arguments. While the standard problem of machine learning from examples is to induce a hypothesis that explains given examples, in ABML some of these examples are argumented, and the problem of ABML is to induce a hypothesis that explains examples using these arguments. The arguments for specific

examples are a special form of expert's background knowledge which he/she uses to substantiate the class value for the chosen example. We believe that it is much easier for the expert to justify the class value of the specific example than to provide some generally applicable rules. We consider this as the main advantage of ABML approach. The other two important advantages of ABML are:

1. Arguments impose constraints over the space of possible hypotheses, thus reducing search complexity,
2. An induced hypothesis should make more sense to an expert as it has to be consistent with given arguments.

Regarding the first advantage above, it is obvious that constraining the search space should help to overcome the problem of explosive combinatorial space of possible hypotheses. Arguments do not simply reduce search complexity but they rather do it smarter, by directing the search into subspaces where better hypotheses should reside. Regarding the second advantage, we should mention that several hypotheses could explain the given examples well but some may not be understandable to the expert. By including the arguments the induced hypotheses should make more sense to the expert.

2.1. ABCN2

Argument Based CN2 (ABCN2) is a realization of concepts just described. It is an extension of the rule learning algorithm CN2 [4,5] in which a subset of learning examples may be argumented. The details of the algorithm and the formalism of accepted arguments by the method are described in [16,15]. Here we shall give only a brief overview.

2.1.1. Argumented examples

A learning example E in the usual form accepted by CN2 is a pair (A, C), where A is an attribute-value vector, and C is a class value. An argumented example AE is a triple of the form:

$$AE = (A, C, Arguments)$$

As usual, A is an attribute-value vector and C is a class value. $Arguments$ is a set of arguments $Arg_1, \ldots, Arg_n$, where an argument Arg_i has one of the following forms:

$$C \text{ because } Reasons$$

or

$$C \text{ despite } Reasons$$

The former specifies a *positive* argument (speaks for the given class value), while the latter specifies a *negative* argument (speaks against the class value). $Reasons$ is a conjunction of reasons $r_1, \ldots, r_n$,

$$Reasons = r_1 \wedge r_2 \wedge \ldots \wedge r_n$$

where each of the reasons r_i is a condition on a single attribute (e.g. $X = x$, where X is the name of the attribute and x is a possible value for this attribute).

2.1.2. ABCN2 - Algorithm

ABCN2 is based on the version of CN2 that induces a set of unordered rules [5]. The main difference between these methods is in the definition of rule *covering*. In the standard definition (CN2), a rule covers an example if the condition part of the rule is true for this example. In argument based rule learning, this definition is modified to: A rule R *AB-covers* an argumented example E if:

1. All conditions in R are true for E (same as in CN2),
2. R is consistent with at least one positive argument of E, and
3. R is not consistent with any of negative arguments of E,

where rule R is consistent with an argument Arg if the reasons of Arg are present among conditions of R.

We mentioned that the first requirement for ABML is that an induced hypothesis explains argumented examples using given arguments. In rule learning this means that each argumented example must be covered by at least one rule that AB-covers the example. This is achieved simply by replacing covering in original CN2 with AB-covering. However, although replacing the "covers" relation (from CN2) with "AB-covers" in ABCN2 ensures that both argumented and non-argumented examples are AB-covered, we improved the initial CN2 algorithm so that induced rules explain as many as possible non-argumented examples by arguments given for the argumented examples (see [16,15]). However, in the ABCN2 algorithm there are still three important parts inherited from CN2 that render learning from argumented examples less effective as it could be. These are: examples removing strategy (after a rule is learned), evaluation function, and classification from rules. In the remaining of this section we will explain why these parts are problematic and how did we improve them.

2.1.3. Removing strategy

After CN2 learns a rule, it removes examples covered by this rule and recursively continues learning on the remaining examples. This approach assumes that algorithm induces the best possible rule for given examples - there exist no rule that would be evaluated better than this rule and cover the same examples. This assumption might be true for the original CN2, but for ABCN2, where we first learn from argumented examples (learning is constrained by arguments of argumented example), this assumption is likely to be incorrect. A rule, learned from an argumented example, can be seen as the best possible rule covering this example. No-one can assure that this rule is also the best rule for other examples covered by this rule. For instance, it could happen that CN2 finds a better rule for some of these examples. Therefore, removing examples after learning from argumented examples might prevent classical CN2 from learning some good rules. In [17] we developed a probabilistic covering strategy.

2.1.4. Evaluation function

The evaluation function in rule learning algorithms is used to determine the goodness (or quality) of a rule. This measure of goodness should determine the rule's potential to predict yet unseen examples. In the original CN2, rules are evaluated using Laplace formula for probability. Due to a search through a huge space of possible hypotheses, this evaluation method usually gives optimistic estimates of probability [18]. In the case of

ABCN2 rules learned from argumented examples are selected from less hypotheses than rules induced with standard CN2 algorithm, and thus the quality of a rule learned from an argumented example is relatively under-estimated when compared to a rule learned from standard CN2. We developed a novel evaluation method based on extreme value theory [18] that accounts for multiple comparisons in the search. Using this method, the evaluations of rules learned from arguments are not under-estimated any more. Bearing this fact, the quality of a rule becomes now a very important factor in classification.

2.1.5. Classification from rules

Most of the methods for classification from rules base on the distribution of covered examples by these rules. However, similarly to Laplace evaluation function, the number of positive examples in the distribution tends to be optimistic. As our evaluation function, described in the previous section, accounts for the number of tried hypotheses, it would make sense to use the quality of a rule (instead of distribution) in classification. We developed such method based on the Minimax theorem [12], for a detailed explanation of this classification method see [17].

3. Experiments

3.1. Data

The data for our study was gathered at the Clinic for Infectious Diseases in Ljubljana, from June 1st, 2004 to June 1st, 2005. The physicians included only the patients over 65 years of age with CRP value over 60 mg/l, which indicated a bacterial etiology of the infection. The patients were observed 30 days from the first examination or until death caused by the infection. The data includes 40 clinical and laboratorial parameters (attributes) acquired at the first examination for each of 298 patients (examples). The infections are distinguished with respect to the site where bacteria is found or on the clinical basis (respiratory, urinary tract, soft tissues, other). The continuous attributes were categorized by the physician. The distribution of the class values is the following:

- 34 examples (11,4%) for 'death = yes'
- 263 examples (88,6%) for 'death = no'

3.2. Arguments

The argumentation was done by the physician who was treating the patients and could by her expert knowledge state several positive and negative arguments to 32 examples, where all argumented examples were from class *death = yes*, namely she gave the reasons she believed caused death for each selected patient. A sample argumented example is shown in Table 1.

One could, at this place, ask himself an interesting question about these arguments, whether they would, if used as rules, describe the domain sufficiently well. We built a simple classifier from given arguments and tested it on the same data set; for each case, we counted the number of applicable arguments for class *death = yes* and compared this number to the number of arguments for class *death = no*. The accuracy of a such classifier is only slightly above 40%, therefore there is still a large space available for machine

Attribute	Value	
GENDER	Z	**Positive arguments**
AGE_YEARS	92	
AGE	C	DEATH=YES because RESPIRATORY_RATE_D=“>=16”
NURSING_HOME_RESIDENT	NO	DEATH=YES because SATURATION_D=“<=90”
COMMORBIDITY	0	DEATH=YES because BLOOD_PRESSURE_D=”<=100”
DIABETES	NO	DEATH=YES because TEMPERATURE_D=“>37.9”
HEART	NO	DEATH=YES because LEUKOCYTES_D=“>=12”
KIDNEY	NO	DEATH=YES because CREATININE_D=“>=100”
LIVER	NO	DEATH=YES because BLOOD_UREA_D=“>=13”
LUNG	NO	DEATH=YES because NA_D=“>147”
IMMUNITY	NO	DEATH=YES because AGE_YEARS is high
CENTRAL_NERVE_SYSTEM	NO	DEATH=YES because WEAKNESS=YES
MOBILITY	YES	DEATH=YES because CONSCIOUSNESS=DISSORIENTED
CONTINENCE	YES	
BEDSORE	NO	**Negative arguments**
CATHETER	NO	
IMPLANT	NO	DEATH=YES despite MOBILITY=YES
VOMITING	NO	DEATH=YES despite CONTINENCE=YES
DIABLOODPRESSUREHEA	NO	DEATH=YES despite TROMBOCYTES_D=“>=100”
WEAKNESS	YES	DEATH=YES despite HEART_RATE_D=“<100”
CONSCIOUSNESS	DISSORIENTED	DEATH=YES despite RODS_D=“<10”
TROMBOCYTES_D	>=100	DEATH=YES despite CRP_D=“<150”
TEMPERATURE_D	>37.9	DEATH=YES despite COMMORBIDITY=0
RESPIRATORY_RATE_D	>=16	
SATURATION_D	<=90	
HEART_RATE_D	<100	
BLOOD_PRESSURE_D	<=100	
LEUKOCYTES_D	>=12	
RODS_D	<10	
CRP_D	<150	
CREATININE_D	>=100	
BLOOD_UREA_D	>=13	
GLU_D	<15	
NA_D	>147	
INFECTION_TYPE	RESPIRATORY	
DEATH (class value)	YES	

Table 1. A sample argumented example from the infections database.

learning to improve. However, please note that this experiment is not used to validate the expert knowledge. To do that, at least the arguments to examples from the opposite class should be given as well. Our intention is merely to show that the knowledge given by the arguments is not perfect nor complete though it can still help to improve learning.

3.3. Results

Learning and testing was performed by 10-fold cross validation which was carried out 10 times with different random splits of examples into folds. We compared the algorithms ABCN2 and CN2, where both methods used improvements shown in the previous section, so that their comparison directly represents the influence of the arguments added to the learning examples. Both algorithms are then compared to Naïve Bayes (NB), decision trees (C4.5) and logistic regression (LogR). Algorithms are compared with regards to classification accuracy, area under ROC (AUC) and Brier score. All methods and tests were implemented within Orange toolkit [6]. The results are shown in Fig. 1- 3.

Observing classification accuracy, the measure that determines the percentage of correct classifications, we can see that CN2, ABCN2 and C4.5 achieve similar results while NB and LogR perform significantly worse (Fig. 1). Although classification accuracy is important it should be accompanied by other estimates especially because the majority

classifier itself is quite accurate in this domain. Therefore we also measure AUC and Brier score, which are relevant as all methods uses can also give the probability of predicted class. AUC measures how well can the method rank examples; it is the probability that for two randomly chosen examples with different classes, the method will predict higher probability for class value of the first example to the first example than to the second example. Figure 2 shows that, according to AUC, ABCN2 significantly outperforms all other methods. The same effect also comes out in Brier scores (Fig. 3), which measure the average quadratic error of predicted probability.

3.4. Discussion

ABCN2 achieved better results than CN2 according to all three measures by using arguments given by expert. The question is how do the induced hypotheses from both measures differ and why ABCN2 is the better method. To examine the hypotheses, we induced a set of rules from the whole data set with ABCN2 and CN2. As the arguments were given only to examples with class value *death=yes*, the induced rules for *death=no* were the same for both methods. Both methods induced 14 rules for class *death=yes*, however there were two important differences between these two sets of rules. First, due to the restriction of hypotheses space with arguments, about half of the rules were different. While inspecting the rules that were the same in CN2's and ABCN2's set, we noticed that the estimations of qualities of these rules did sometimes differ. For example, the rule:

IF trombocites<100 AND mobility=no THEN death=yes

was present in the both rule sets. It covers 6 examples with class value *death=yes* and 1 with *death=no*, which would mean that the probability of *death=yes* is 6/7 = 0.86. However, the evaluation function based on extreme value distributions [18] used in CN2 estimated the probability of this class (given that the conditions are true) as 0.47, which is much less than 0.86. This happened because there is a high probability that such a rule would be found by chance. On the other hand, when learning with ABCN2, the evaluation of the same rule is 0.67. In CN2, this rule was obtained by searching the whole space unguided by expert knowledge while in ABCN2 the rule is built from the argument 'death=yes BECAUSE trombocites<100'. Therefore, as the search space in ABCN2 is therefore smaller, which means that the probability of finding such a rule by chance is lower, the expected quality of rule is higher.

In the above paragraph we have shown the importance of the first expected advantage of of ABML: "Arguments impose constraints over the space of possible hypotheses, thus reducing search complexity". Regarding the second advantage, that induced rule should make more sense to an expert, we asked our expert to examine the rules and compare them. Unfortunately, she could not choose which rules are more understandable to her. We believe that this occurs due to the large number of arguments with only one reason given to each example, while our restriction is that the rule must be consistent with at least one positive argument. The rule must, therefore, contain only one of given reasons and can neglect others.

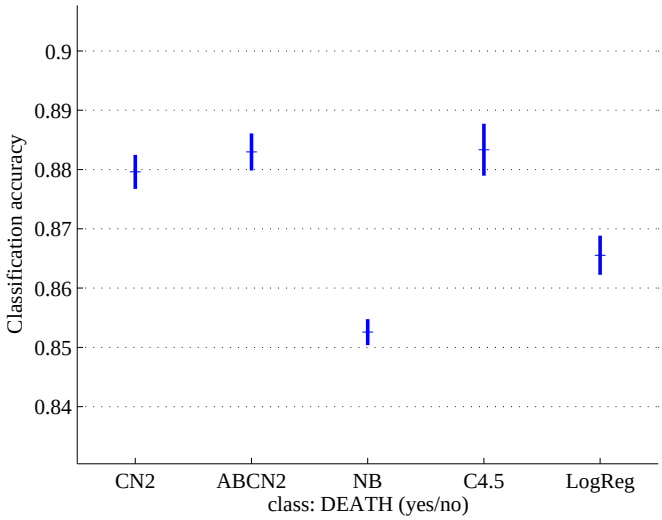

Figure 1. Mean values and standard errors of classification accuracy across tested methods.

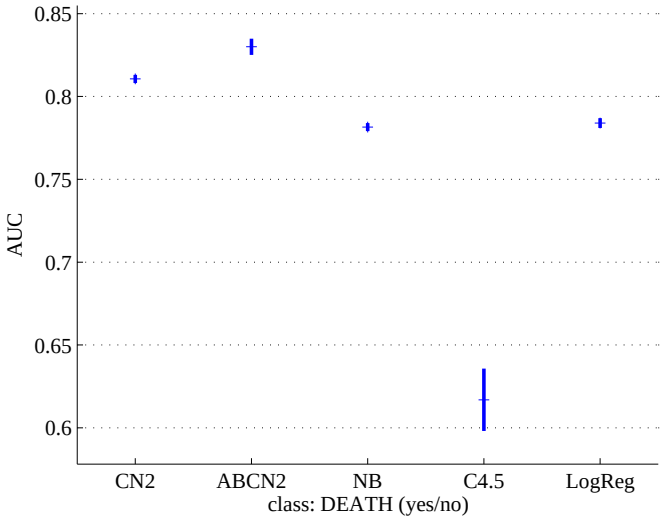

Figure 2. Mean values and standard errors of AUC across tested methods.

4. Conclusion

We described the application of argument based machine learning to the medical domain dealing with severe bacterial infections in geriatric population. Our intention was to show how arguments can be used to guide a machine learning algorithm towards a better model. The use of arguments proved to be a powerful approach which offers a new insight in using the expert knowledge in machine learning. This knowledge is not given

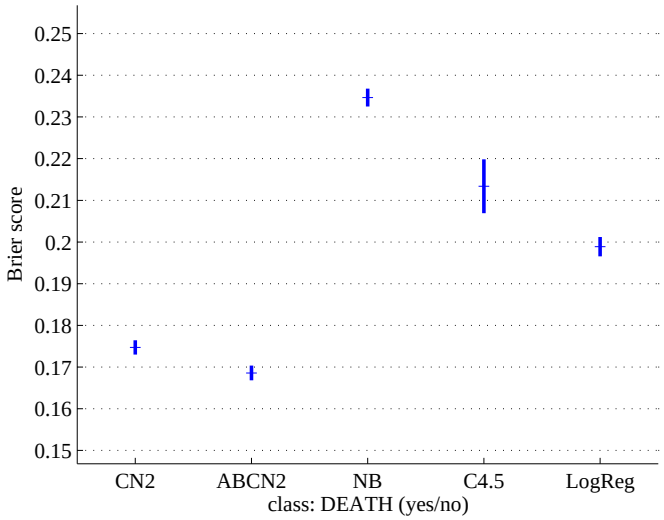

Figure 3. Mean values and standard errors of Brier score across tested methods.

as a general background knowledge but is rather passed to specific examples to reason the class value using available attributes.

We used ABCN2 which is an argument-based version of CN2 algorithm. Our medical domain is the first real-life domain to which ABML has been applied. Several examples were argumented by the medical doctor and used in the learning process. In our experiments we compared ABCN2 to a few machine learning algorithms that are not capable of using arguments, such as CN2, C4.5, Naïve Bayes and logistic regression. The results show several improvements of ABCN2 over other algorithms. ABCN2 significatnly outperforms others in classification accuracy, AUC and Brier score.

For further work, it would be very interesting to see how well can an expert alone (without machine learning) classify the examples. We would need to ask an independent expert, who has not seen these examples yet, and ask her to to classify them according to her knowledge. We believe that such experiment would truly show the added value of argument based approach. Moreover, it would be also interesting to see how would the number of of argumented examples influence the results and check how the results change if we select different subsets of argumented examples. In our experiment the number of arguments was quite large, which might not always happen, as argumenting examples is usually time consuming for experts. Another interesting experiment would be to have a few physicians argumenting the examples and compare the models.

Acknowledgements

This work was carried out under the auspices of the European Commission's Information Society Technologies (IST) programme, through Project ASPIC (IST-FP6-002307).

References

[1] Ben-Yehuda A, Weksler ME. Host resistance and the immune system. Clin Geriatr Med 1992; 8: 701-11.

[2] Bratko I, Možina M: Argumentation and Machine Learning. In: Deliverable 2.1 for the ASPIC project 2004.

[3] Castle SC et al. Fever response in elderly nursing home residents: are the older truly colder? JAGS 1991; 39: 853-7.

[4] Clark P, Niblett T: The CN2 induction algorithm. Machine Learning Journal, 4: 261-283, 1989.

[5] Clark P, Boswell R. Rule Induction with CN2: Some Recent Improvements. In Machine Learning - Proceedings of the Fifth European Conference (ESWL-91), pages 151-163, Berlin, 1991.

[6] Demšar J, Zupan B, Leban G (2004) Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.

[7] Farr BM, Sloman AJ, Fisch MJ. Predicting death in patients hospitalized for community acquired pneumonia. Ann Int Med 1991; 115:428-36.

[8] Fine MJ, Smith MA, Carson CA, et al. Prognosis and outcomes of patients with community acquired pneumonia. JAMA 1996; 275: 134-41.

[9] Fine MJ, Auble TE, Yeay DM, et al. A prediction rule to identify low-risk patients with community acquired pneumonia. NEJM 1997; 336: 243-50.

[10] Fontanarosa PB, Kaeberlein FJ, Gerson LW, Thompson RB. Difficulty in predicting bacteriemia in elderly emergency patients. Ann Emerg Med 1992; 21: 842-8.

[11] Gleckman R, Hibert D. Afebrile bacteriemia: a phenomena in geriatric patients. JAMA 1982; 248: 1478-81.

[12] John von Neumann. Zur Theorie der gessellshaftsspiele. Mathematische Annalen, 100:295-320, 1928. English Translation Fin Tucker AW, Luce RD, Contributions to the Theory of Games IV, Annals of Mathematics Studies 40, 1959.

[13] Marco CA et al. Fever in geriatric emergency patients: clinical features associated with serious illness. Ann Emerg Med 1995; 26:18-24.

[14] Mellors JW et al. A simple index to identify occult bacterial infection in adults with acute unexplained fever. Arch Intern Med 1987; 147: 666-71.

[15] Možina M, Žabkar J, Bench-Capon T, Bratko I: Argument Based Machine Learning Applied to Law. Artificial Intelligence and Law, 2005; In press.

[16] Možina M, Žabkar J, Bratko I: Argument Based Rule Learning. Accepted for publication in Proceedings of ECAI, Riva del Garda, 2006.

[17] Možina M, Žabkar J, Bratko I: Implementation of and experiments with ABML and MLBA, preliminary version. ASPIC deliverable D3.4, 2006.

[18] Možina M, Demšar J, Žabkar J, Bratko I: Why is Rule Learning Optimistic and How To Correct It. Submitted to ECML conference, 2006.

[19] Pfitzenmeyer P, Decrey H, Auckenthaler R, Michel JP. Predicting bacteriemia in older patients. JAGS 1995; 43: 230-5.

[20] Prakken H, Vreeswijk G. Handbook of Philosophical Logic, second edition, volume 4, chapter Logics for Defeasible Argumentation, pages 218-319. Kluwer Academic Publishers, Dordrecht etc, 2002.

[21] Rockwood K. Acute confusion in elderly medical patients. J Am Geriatr Soc 1989; 37: 150-4.

[22] web page: http://www.stat.si/tema_demografsko_prebivalstvo.asp, 2005.

[23] Yoshikawa TT. Epidemiology and unique aspects of aging and infectious diseases. Clin Infect Dis 2000; 30: 931-3.