# Argument Based Moderation of Benefit Assessment

Maya WARDEH, Trevor J.M. BENCH-CAPON and Frans COENEN
*Department of Computer Science*
*The University of Liverpool*
*Liverpool*
*UK*

**Abstract**: Error rates in the assessment of routine claims for welfare benefits have been found to very high in Netherlands, USA and UK. This is a significant problem both in terms of quality of service and financial loss through over payments. These errors also present challenges for machine learning programs using the data. In this paper we propose a way of addressing this problem by using a process of moderation, in which agents argue about the classification on the basis of data from distinct groups of assessors. Our agents employ an argument based dialogue protocol (PADUA) in which the agents produce arguments directly from a database of cases, with each agent having their own separate database. We describe the protocol and report encouraging results from a series of experiments comparing PADUA with other classifiers, and assessing the effectiveness of the moderation process.

## Introduction

A significant problem in the assessment of claims to welfare benefit is the high error rate encountered. Groothuis and Svensson [1] drew attention to this in connection with the Netherlands General Assistance Act, and reported experiments which suggested that an error rate of more than 20% was typical. The problem is international: The US National Bureau of Economic Research reports of US Disability Insurance [2]:

''The multistage process for determining eligibility for Social Security Disability Insurance (DI) benefits has come under scrutiny for the length of time the process can take – 1153 days to move through the entire appeals process, according to a recent Social Security Administration (SSA) analysis – and for inconsistencies that suggest a potentially high rate of errors. One inconsistency is the high reversal rate during the appeals process – for example, administrative law judges, who represent the second level of appeal, award benefits in 59% of cases. Another inconsistency is the variation in the award rates across states – from a high of 65% in New Hampshire to a low of 31% in Texas in 2000 – and over time – from a high of 52% in 1998 to a low of 29% in 1982.''

Similar observations are made of the UK. An official UK Publication produced by the Committee of Public Accounts [3]: ''Finds that the complexity of the benefits system remains a major problem and is a key factor affecting performance. Skills of

decision makers need to be enhanced through better training and wider experience. Too few decisions are right first time, with a error rate of 50% for Disability Living Allowance. There are also regional differences in decision making practices that may lead to payments to people who are not eligible for benefits.''

A 2006 report from the UK national Audit Office [4] estimated losses from error in Social Security benefits at around £1 billion per annum, and stated that "Errors by officials arise mainly because of the sheer complexity of benefit rules and regulations."

There is then a significant problem, which it is clear current procedures are unable to address satisfactorily. One important feature of the errors is that they are not random: regional differences in decision making practices arise from the complexity of the rules and regulations because the misunderstandings and misinterpretations differ from office to office. Thus one office will tend to decide one class of case wrongly, while a different office will get this right, but fail on another class of cases.

One way of resolving disagreement in assessment – very common in academic circles – is to a have a moderation discussion. The parties in disagreement argue for their position with one another, and so can come to recognise strengths and weaknesses that they have overlooked or under weighted and thus converge on agreed decisions. In this paper we describe how we can exploit an argumentation based dialogue system, PADUA (Protocol for Argumentation Dialogue Using Association Rules), to provide an analogue of this process for social security benefit decisions made in different offices.

In section 1 we will describe PADUA. In section 2 we outline the data to which we will apply it and report some experiments designed to explore the effectiveness of the approach. The final section concludes the paper with some discussion and directions for future work.

## 1. PADUA

Whereas most systems using argumentation are based on belief bases, sets of rules encoding knowledge of the domain (e.g. [5]), PADUA is based on the notion of arguing directly from experience, without the need to form a theory of the domain. In PAUDA, the dialogue participants form their arguments on the basis of a database of decided examples, with the different participants having their own local collection of cases. The difference of opinion to be resolved comes from the fact that experiences differ, and so the set of examples available to the participants may ground different conclusions with respect to a new example.

This alternative basis for persuasion dialogues, to enable what we termed *arguing from experience* to solve classification problems, was introduced in ([6], [7]). When presented with a new case the agents use data mining techniques to discover associations between features of the case under consideration and the appropriate classification according to their previous experience. We argued that this has several advantages:

1. Such arguments are often found in practice: many people do not develop a theory from their experience, but when confronted with a new problem recall past examples;
2. It avoids the knowledge engineering bottleneck that occurs when belief bases must be constructed;

3. There is no need to commit to a theory in advance of the discussion: the information can be deployed as best meets the need of the current situation;
4. It allows agents to share experiences that may differ: one agent may have encountered types of case that another has not, or may make mistakes that another does not.

For the moderation application described in this paper it has the additional point that we are dealing with two distinct sets of data, exhibiting different systematic flaws, that we wish to reconcile.

The moves made in arguments based directly on examples contrast with those found in persuasion dialogues based on belief bases, and have a strong resemblance to those used in case based reasoning systems, e.g. [8] and [9], although a generalisation from a set of cases is cited, rather than a single precedent case.

PADUA (*Protocol for Argumentation Dialogue Using Association Rules*) is an argumentation protocol designed to enable participants to debate on the basis of their experience. PADUA has as participants agents with distinct datasets of records relating to a classification problem. These agents produce reasons for and against classifications by mining association rules from their datasets using data mining techniques ([10], [11] and [12]). By "*association rule*" we mean that the antecedent is a set of reasons for believing the consequent. In what follows $P \rightarrow Q$ should be read as "*P* are reasons to believe *Q*". A full description of PADUA is given in [7].

PADUA adopts six dialogue moves:

- *Propose Rule*: allows generalizations of experience to be cited, by which a new association with a confidence higher than a certain threshold is proposed.
- *Attacking moves:*
  - *Distinguish*: When a player *p* plays a *distinguish* move, it adds some new premise(s) to a previously proposed rule, so that the confidence of the new rule is lower than the confidence of the original rule.
  - *Counter Rule*: is very similar to *propose rule* and is used to cite generalizations leading to a different classification
  - *Unwanted Consequences*: Here the player *p* suggests that certain consequences (conclusions) of the rule under discussion do not match the case under consideration.
- *Refining moves:* these moves enable a rule to be refined to meet objections:
  - *Increase Confidence*: a player *p* adds one or more premise(s) to a rule it had previously played to increase the confidence of this rule.
  - *Withdraw unwanted consequences*: a player *p* plays this move to exclude the unwanted consequences of the rule it previously proposed, while maintaining a certain level of confidence.

| Move | Label | Next Move | New Rule |
|---|---|---|---|
| 1 | Propose Rule | 3, 2, 4 | Yes |
| 2 | Distinguish | 3, 5, 1 | No |
| 3 | Unwanted Cons | 6, 1 | No |
| 4 | Counter Rule | 3, 2, 1 | Nested dialogue |
| 5 | Increase Conf | 3, 2, 4 | Yes |
| 6 | Withdraw Unwanted Cons | 3, 2, 4 | Yes |

Table 1 – PADUA moves.

The PADUA protocol defines for each of those six moves a set of legal next moves (i.e. moves that can possibly follow this move). Table 1 summarizes PADUA protocol rules, and indicates whether a new rule is introduced.

For a fuller discussion of the operations of the PADUA system and the rationale for its moves see [7]: for a discussion of different strategies that the participating agents can use, see [6].

## 2. The Experiments

To illustrate experimentally the kinds of dialogues produced by PADUA, we applied PADUA to a fictional welfare benefit scenario, where benefits are payable if certain conditions showing need for support for housing costs are satisfied. This scenario is intended to reflect a fictional benefit Retired Persons Housing Allowance (RPHA), which is payable to a person who is of an age appropriate to retirement, whose housing costs exceed one fifth of their available income, and whose capital is inadequate to meet their housing costs. Such persons should also be resident in this country, or absent only by virtue of "service to the nation", and should have an established connection with the UK labour force. These conditions need to be interpreted and applied [13]. We use the following desired interpretations:

1. *Age condition*: "Age appropriate to retirement" is interpreted as pensionable age: 60+ for women and 65+ for men.
2. *Income condition*: "Available income" is interpreted as net disposable income, rather than gross income, and means that housing costs should exceed one fifth of candidates' available income to qualify for the benefit.
3. *Capital condition*: "Capital is inadequate" is interpreted as below the threshold for another benefit.
4. *Residence condition*: "Resident in this country" is interpreted as having a UK address. Residence exception: "Service to the Nation" is interpreted as a member of the armed forces.
5. *Contribution condition*: "Established connection with the UK labour force" is interpreted as having paid National Insurance contributions in 3 of the last 5 years.

These conditions fall under a number of typical types: conditions (2 and 3) represent necessary conditions over continuous values while conditions (4 and 5) represent a restriction and an exception to the applicant's residency, condition (1) deals with variables depending on other variables and condition (6) is designed to test the cases in which it is sufficient for some *n* out of *m* attributes to be true (or have some predefined values) for the condition to be true. These conditions have been used in several previous experiments in AI and Law [14], [21~~+~~5].

We now suppose that this benefit is assessed in two different offices, covering different regional areas, and each producing errors through a different misinterpretation. We ran three experiments:

1. An experiment to test the extent to which classification would be improved by moderation using PADUA. This was done using a 10 fold cross validation test. A number of other classifiers were also applied to the data to provide a comparison.

2. A McNemar test to show the significance of the differences between classifiers.
3. We then performed a more detailed analysis of the performance of PADUA in order to discover some interesting properties of the moderation dialogues.

For tests we generated two sets of data. Each record comprises 13 fields, the information relevant to the above tests being surrounded by other features which should be irrelevant to the determination of the case.

Both contained 500 cases which should be awarded benefit and 500 cases which should be denied benefit. Cases can fail on any one of five conditions, and the failing cases were evenly divided across them. One dataset was completed by the addition of 500 cases which should fail on the age condition, but which in fact awarded benefit to men over 60, and the other with 500 cases which should have failed the residence condition, but which interpreted the exception too widely, allowing benefit to members of the Merchant Navy and the Diplomatic Service.

### *2.1. Cross Validation and Comparison with Other Classifiers*

The baseline was the number of correct cases in the dataset: namely the 66.7% accuracy which had been achieved by the original decision makers. Five other classifiers were used, operating on the union of the two data sets. These other classifiers were:

1. *TFPC*: TFPC, Total From Partial Classification ([16], [17]), is a Classification Association Rule Mining (CARM) algorithm founded on the TFP (Total From Partial) Association Rule Mining (ARM) algorithm ([18],[19]); which, in turn, is an extension of the Apriori-T (Apriori Total) ARM algorithm.

   TFPC is designed to produce Classification Association Rules (CARs) whereas Apriori-T and TFP are designed to generate Association Rules (ARs). In its simplest form TFPC determines a classifier according to given support and confidence thresholds. The nature of the selected thresholds is therefore the most significant influencing factors on classification accuracy. A more sophisticated version of TFPC uses a hill climbing technique to find a best accuracy given start support and confidence thresholds.

2. *CBA*: CBA (Classification Based on Associations) is a Classification Association Rule Mining (CARM) algorithm developed by Bing Liu, Wynne Hsu and Yiming Ma [20]. CBA operates using a two stage approach to generating a classifier:
   1. Generating a complete set of CARs.
   2. Prune the set of CARs to produce a classifier.
3. *CMAR*: CMAR (Classification based on Multiple Association Rules) is another CARM algorithm developed by Wenmin Li, Jiawei Han and Jian Pei [21]. CMAR also operates using a two stage approach to generating a classifier:
   1. Generating the complete set of CARs according to a user supplied:
      - Support threshold to determine frequent (large) item sets, and
      - Confidence threshold to confirm CRs.
   2. Prune this set to produce a classifier.
4. *Decision Trees*: Classification using *decision trees* was one of the earliest forms of data mining. Ross Quinlan's C4.5 is arguably the most referenced decision tree algorithm [22]. One of the most significant issues in decision tree generation is

deciding on which attribute to *split*. Various algorithms have been proposed in the literature. Two are used here:

- Most frequently supported (or Random) Decision Trees (RDT):
- Information Gain Decision Trees (IGDT).

The first selects the first attribute in a list of attributes order according to its support frequency within the entire data set. Information gain [23] is one of the standard measures used in decision tree construction.

The cross validation was achieved by running the experiment 10 times, each time leaving out a randomly selected 10% of the available data. For PADUA, two runs were performed, one in which the agent with Dataset 1 (DS1) was the proponent (i.e. argued for award of benefit), and one in which the agent with Dataset 2 (DS2) was the proponent. The results are presented in Table 2.

| Trial | DS1% | DS2% | TFPC % | CBA% | CMAR% | RDT% | IGDT% |
|---|---|---|---|---|---|---|---|
| 1 | 95.125 | 94.375 | 64.33 | 64.33 | 63.87 | 95.4 | 91.8 |
| 2 | 95.5 | 92.625 | 64.33 | 64.33 | 63.87 | 95.93 | 90.87 |
| 3 | 95.125 | 92.5 | 64.33 | 64.33 | 64.07 | 96.6 | 91.8 |
| 4 | 95.5 | 92.75 | 64.33 | 64.33 | 63.87 | 95.87 | 90.87 |
| 5 | 96 | 88.875 | 64.33 | 64.33 | 64.07 | 95.87 | 92 |
| 6 | 96.75 | 93.25 | 64.33 | 64.33 | 63.87 | 95.93 | 92 |
| 7 | 96.375 | 94.125 | 64.33 | 64.33 | 63.87 | 95.8 | 92 |
| 8 | 96.25 | 93.25 | 64.33 | 64.33 | 63.87 | 95.8 | 91 |
| 9 | 94.125 | 93.875 | 64.33 | 64.33 | 63.87 | 95.8 | 91.33 |
| 10 | 94.375 | 93.875 | 64.33 | 64.33 | 63.87 | 95.8 | 91.2 |
| Summary | 95.83 | 92.72 | 64.33 | 64.33 | 63.91 | 95.88 | 91.487 |

Table2 – Experiment 1 results.

From this we can see that the three association rule classifiers perform less well than the baseline. In contrast PADUA, and the decision tree based classifiers perform significantly better, attaining above 90% accuracy in all cases. While, however, the decision tree classifiers perform rather consistently throughout the ten trials, there is more variation in PADUA, especially for DS2, suggesting that its performance is more sensitive to the exact sample available to the agents. This will be considered in more detail in section 2.3 below.

Overall we find the level of performance encouraging. For comparison with other AI and Law systems, Bench-Capon [14] reported an accuracy of 98%, but that was based on training set of correctly decided cases. Ashley and Brunighaus [~~15~~24] reported a success rate of 91.4% for IBP, and Chorley and Bench-Capon [~~24~~25] a success rate of between 91% and 93% for AGATHA, both applied to noise free examples of US Trade Secret Law. One effort to explore how learning is affected by noise is [~~25~~15]. They used used a very similar dataset and introduced randomly, rather than systematically misdecided cases. Their results gave a success rate of 92% with 20% noise falling to 85% with 40% noise for Argument based Explanation, and 89% for 20% noise falling to 83% for 40% noise for a conventional rule induction algorithm

CN2. It seems therefore, from this previous work, that the level of accuracy attained by PADUA is towards the top end of what can be expected from successful classification systems in AI and Law.

### *2.2. McNemar Test*

The McNemar test is a non-parametric test designed to explore the hypothesis that one classifier is significantly better than another. As might be expected from the results shown in Table 2, PADUA DS1 and DS2 were significantly better than the three association rule classifiers and RDT, but not significantly better or worse than IGDT.

For this test PADUA operated on a set of newly generated cases (500 positive, 500 negative as before and 250 wrongly decided, appropriate to each database). This data was then used as a test set for the other algorithms the original data supplying the training set.

As part of the test, we generate detailed information as to which cases are misclassified by one or both of the classifiers under consideration. These results for DS1 are shown in Table 3a and those for DS2 in Table 3b: n00 are cases misclassified by both, n01 are cases misclassified by PADUA only, n10 are cases correctly classified by PADUA and misclassified by the comparator, and n11 are cases correctly classified by both:

| | DS2 | TFPC | CMAR | CBA | RDT | IGDT |
|---|---|---|---|---|---|---|
| n00 | 5 | 8 | 8 | 145 | 7 | 10 |
| n01 | 146 | 139 | 139 | 2 | 140 | 137 |
| n10 | 142 | 318 | 364 | 461 | 62 | 129 |
| n11 | 1207 | 1035 | 989 | 892 | 1291 | 1224 |

Table 3a: Comparison with DS1

| | DS1 | TFPC | CMAR | CBA | RDT | IGDT |
|---|---|---|---|---|---|---|
| n00 | 5 | 48 | 92 | 55 | 10 | 22 |
| n01 | 142 | 103 | 59 | 96 | 141 | 129 |
| n10 | 146 | 214 | 514 | 66 | 31 | 119 |
| n11 | 1207 | 1136 | 835 | 1283 | 1318 | 1230 |

Table 3b: comparison with DS2

What is interesting here is that although both classifiers only succeed only on 86% of cases for RDT and DS1, 81% of cases for DS1 and IGDT and 82% of cases for DS2 and IDG; the mistakes are very different. Less that 0.5% of the cases are misclassified both by DS1 and RDT and only 1% by the worst combination, DS2 and IDGT. This suggests that we could profitably use PADUA and a decision tree method in combination. If cases where there was agreement were believed to be correct, and we used, for example, DS1 and RDT; and referred cases of disagreement to an expert for decision we could reduce error rates to below 0.05%, at the cost of checking some 13.5% of the cases. Since it is current practice to check 10% of decisions, chosen at random, we could thus, by focusing the cases for expert checking, reduce the error rate

with very little additional expert intervention. Moreover, DS1 and DS2 only both misclassify one case in three hundred, although they are both successful in only 80%. Using PADUA alone, but having each case argued for by both agents, therefore, could reduce the error rate to 0.003%, although it would require around 20% of cases to be checked. This, however, might be improved by first using an expert to resolve a proportion of the cases with disagreement (say a quarter, 5% of all the cases) with disagreement, entering the corrected values into the databases, and then rerunning the moderation. Since performance improves as noise is reduced, this should generate fewer disagreements, reducing the overall checking requirement to an acceptable level.

*2.3. Detailed Consideration of DS1 and DS2.*

In this section we will look at the ten cross validation trials for PADUA in more detail. The detailed results are shown in Table 4a, Table 4b.

| Test | Positive | | Negative Age | | Negative Income | | Negative Capital | | Negative Residency | | Negative Contribution Years | | All Female Exception | | All UK Exception | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pro | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| **1** | 6 | 98 | 92 | 100 | 100 | 91 | 96 | 98 | 96 | 98 | 96 | 98 | 88 | 100 | 93 | 72 |
| **2** | 100 | 94 | 92 | 93 | 100 | 95 | 96 | 97 | 96 | 97 | 96 | 97 | 89 | 94 | 95 | 74 |
| **3** | 99 | 98 | 90 | 100 | 100 | 95 | 97 | 93 | 97 | 93 | 97 | 93 | 90 | 100 | 91 | 68 |
| **4** | 100 | 98 | 94 | 100 | 100 | 94 | 97 | 94 | 97 | 94 | 97 | 94 | 85 | 100 | 94 | 68 |
| **5** | 98 | 96 | 94 | 93 | 100 | 93 | 96 | 95 | 96 | 95 | 96 | 95 | 89 | 76 | 99 | 68 |
| **6** | 99 | 98 | 95 | 100 | 99 | 91 | 98 | 93 | 98 | 93 | 98 | 93 | 88 | 100 | 99 | 78 |
| **7** | 98 | 96 | 94 | 100 | 99 | 93 | 96 | 95 | 96 | 95 | 96 | 95 | 92 | 100 | 100 | 79 |
| **8** | 98 | 94 | 95 | 98 | 100 | 91 | 97 | 97 | 97 | 97 | 97 | 97 | 89 | 100 | 98 | 72 |
| **9** | 99 | 96 | 94 | 100 | 99 | 95 | 97 | 94 | 97 | 94 | 97 | 94 | 82 | 100 | 88 | 78 |
| **10** | 97 | 98 | 92 | 100 | 99 | 94 | 96 | 95 | 96 | 95 | 96 | 95 | 82 | 100 | 97 | 74 |

Table 4a – Detailed tests results.

| Trial | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Pro1 | 95.1 | 95.5 | 95.1 | 95.5 | 96 | 96.8 | 96.4 | 96.3 | 94.1 | 94.4 |
| Pro2 | 94.4 | 92.6 | 92.5 | 92.8 | 88.9 | 93.3 | 94.1 | 93.3 | 93.9 | 93.9 |

Table 4b –Summary results.

Three points in particular can be noted from this data.

- The overall performance is rather consistent, with only trial 5 for DS2 showing a significantly worse performance that the rest. Within the detailed breakdown by types of case, however, there is rather more variation.
- Although PADUA succeeds in classifying more cases correctly, some errors are introduced: rarely does it succeed in classifying 100% of cases correct in the data sets correctly. This is because the high number of misclassified cases in the dataset impairs the ability to form correct rules. In particular, the negative age condition becomes harder for DS1, which misunderstands the exception to that condition.

- It matters who is the proponent. For example when DS1 is arguing for benefit for the misclassified age cases, it can defend itself quite a lot of the time. On the other hand when DS2 is proposing that the benefit be given wrongly in these cases, it almost invariably fails. This is readily explicable because DS2 cannot find any good reasons from its own dataset to award benefit in these cases. This effect does not obtain, however, in the case of Trial 5, when DS2 [reforms unusually badly on this factor. One assumes that this is explained by a lack of correctly classified men between 60 and 65 in the particular selection of data used by DS2 in that trial. A similar effect can be observed when cases with misclassified residency are argued for: misclassifications are more likely to be accepted when DS2, which believes them, is the proponent.

It is this last point in particular that suggests that expert resolution of cases where there is disagreement when different agents act as the proponent is likely to be effective in selecting cases for expert checking.

## Discussion

In this paper we have proposed a novel means of attempting to reduce error rates in decisions on Social Security benefits. This is a significant problem, for which a solution is highly desirable. We have proposed an approach to the problem by means of moderation: an argumentation dialogue between two agents, each using their own cases. We have reported experimental results which shows that this dialogue will result in reducing the misclassifications in the databases, from 33% in the original data to less that 10%, a performance superior to other association rule classifiers and comparable with decision tree classifiers, and previously reported AI and Law systems, even where they have used only correctly decided cases for training. Moreover we have shown that the cases which remain misclassified differ according to which agent acts as proponent and which as opponent. By running the cases with first one agent as proponent and then the second as proponent, we find that we can reduce the number misclassified on both runs to 0.003%, although there is disagreement in 20% of cases. We suggest that this could provide an effective way of identifying cases for expert checking, which would improve significantly on the current practice of checking a random sample. Alternatively PADUA could also be effective when used in conjunction with a decision tree classifier.

Current work is focused on extending PADUA from a two agent dialogue protocol to allow for multiple participants. The multi agent version, PISA, [26], will allow for the moderation dialogue to be conducted with a number of agents. Assuming that particular misinterpretations are confined to a minority, we believe that this will result in a highly accurate consensus. We will explore this hypothesis in future work.

## References

[1] Groothius, M. and Svensson, J. (2000). Expert System Support and Juridical Quality. In Proceedings of Jurix 2000, 1–10. IOS Press: Amsterdam.

[2] From Web Page: http://www.nber.org/aginghealth/winter04/w10219.html.

[3] Getting it right: Improving Decision-Making and Appeals in Social Security Benefits. Committee of Public Accounts. London: TSO, 2004 (House of Commons papers, session 2003/04; HC406).

[4] National Audit Office (2006). International benchmark of fraud and error in social security systems REPORT BY THE COMPTROLLER AND AUDITOR GENERAL | HC 1387 Session 2005-2006 | 20 July 2006
[5] H. Prakken (2006). Formal systems for persuasion dialogue. *The Knowledge Engineering Review* 21 (2006): 163-188.
[6] M. Wardeh, T. J. M. Bench-Capon and F. P. Coenen: PADUA Protocol: Strategies and Tactics. In Proc. ECSQARU, 9th European Conf. on Symbolic and Quantitative Approaches to Reasonong with Uncertainty, LNAI 4724, (2007), 465-476.
[7] M. Wardeh, T. J. M. Bench-Capon and F. P. Coenen: Arguments from Experience: The PADUA Protocol. In Proc. Computational Models of Argument, COMMA'2008. IOS press, pp 405-416.
[8] Ashley, K. D. (1990). Modelling Legal Argument. Bradford Books, MIT Press: Cambridge, Mass.
[9] Aleven, V. (1997). Teaching Case Based Argumentation Through an Example and Models, PhD Thesis, The University of Pittsburgh.
[10] R. Agrawal, T. Imielinski, A.N. Swami: Association rules between sets of items in large databases. In: Proc. of ACM SIGMOD Int. Conf. on Management of Data, Washington, (1993), 207-216.
[11] G. Goulbourne, F. P. Coenen and P. Leng: Algorithms for Computing Association Rules Using A Partial- Support Tree. In: Proc. of ES99, Springer, London, UK, (1999), 132-147.
[12] F. P. Coenen, P. Leng and G. Goulbourne: Tree Structures for Mining Association Rules. In: Journal of Data Mining and Knowledge Discovery, Vol 8, No 1, (2004), 25-51.
[13] T.J.M. Bench-Capon: Knowledge Based Systems Applied To Law: A Framework for Discussion. In: T.J.M. Bench-Capon (ed), Knowledge Based Systems and Legal Applications, Academic Press, (1991), 329-342.
[14] Bench-Capon, T. (1993). Neural Nets and Open Texture. In Fourth International Conference on AI and Law, 292–297. ACM Press: Amsterdam.
[15] Martin Mozina, Jure Zabkar, Trevor J. M. Bench-Capon, Ivan Bratko: Argument Based Machine Learning Applied to Law. Artif. Intell. Law 13(1): 53-73 (2005)

~~[15] Bruninghaus, S. and Ashley, K. D. (2003). Predicting Outcomes of Case-based Legal Arguments. In Proceedings of the Ninth International Conference on AI and Law, 233–242. ACM Press: New York.~~
[16] Coenen, F. and Leng, P. (2005). *Obtaining Best Parameter Values for Accurate Classification*. Proc. ICDM'2005, IEEE, pp597-600.
[17] Coenen, F., Leng, P. and Zhang, L. (2005). *Threshold Tuning for Improved Classification Association Rule Mining*. Proceeding PAKDD 2005, LNAI3158, Springer, pp216-225.
[18] Coenen, F., Leng, P. and Ahmed, S. (2004a). *Data Structures for association Rule Mining: T-trees and P-trees*. IEEE Transactions on Data and Knowledge Engineering, Vol 16, No 6, pp774-778.
[19] Coenen, F.P. Leng, P. and Goulbourne, G. (2004b). *Tree Structures for Mining Association Rules*. Journal of Data Mining and Knowledge Discovery, Vol 8, No 1, pp25-51.
[20] Liu, B. Hsu, W. and Ma, Y (1998). *Integrating Classification and Assocoiation Rule Mining*. Proceedings KDD-98, New York, 27-31 August. AAAI. pp80-86.
[21] Li W., Han, J. and Pei, J. (2001). CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. Proc ICDM 2001, pp369-376.
[22] Quinlan, J. R. (1998). *C4.5: Programs for Machine Learning.*. Morgan Kaufmann Publishers.
[23] Mitchell, T.M. (1997). *Machine Learning*. McGraw-Hill.
[24] Bruninghaus, S. and Ashley, K. D. (2003). Predicting Outcomes of Case-based Legal Arguments. In Proceedings of the Ninth International Conference on AI and Law, 233–242. ACM Press: New York.

[~~24~~25] Alison Chorley, Trevor J. M. Bench-Capon: AGATHA: Using heuristic search to automate the construction of case law theories. Artif. Intell. Law 13(1): 9-51 (2005)

~~[25] Martin Mozina, Jure Zabkar, Trevor J. M. Bench-Capon, Ivan Bratko: Argument Based Machine Learning Applied to Law. Artif. Intell. Law 13(1): 53-73 (2005)~~
[26] M. Wardeh, T. J. M. Bench-Capon and F. P. Coenen (2008) PISA - Pooling Information from Several Agents: Multiplayer Argumentation from Experience. To be presented at AI 2008, Cambridge, December 2008.