Noname manuscript No.
(will be inserted by the editor)

Multiagent Based Classification Using Argumentation From Experience

Maya Wardeh · Frans Coenen · Trevor Bench-Capon

Received: date / Accepted: date

Abstract An approach to classification using a multiagent system using an *Argumentation from Experience* paradigm is proposed. The technique is based on the idea that classification can be conducted as a process whereby a group of agents "argue" about the classification of a given case according to their experience as recorded in their individual local data sets. The paper describes mechanisms whereby this can be achieved, which have been realised in the PISA framework. The framework allows both the possibility of agents operating in groups (coalitions) and migrating between groups. The proposed multiagent classification using the Argumentation from Experience paradigm has been used to address standard, ordinal and unbalanced classification problems with good results. A full evaluation, in the context of these applications, is presented.

Keywords Multiagent Classification · Argumentation · Classification Association Rules

1 Introduction

Argumentation is the study of how to draw conclusions based on reasons for and against particular propositions, and has been used in AI and multiagent systems to support persuasion, deliberation and negotiation (see [5] for an overview). Argumentation often takes the form of a debate or dialogue between parties representing the different points of view. For example, the US and the English legal systems are based on an adversarial approach in which opposing lawyers representing the parties to the dispute put forward arguments as to why the case should be decided for their clients. In this way the

M. Wardeh, F. Coenen, T. Bench-Capon
Department of Computer Science, The University of Liverpool,
Liverpool, L69 3BX, UK
E-mail: maya.wardeh@liverpool.ac.uk, coenen@liverpool.ac.uk, tbc@liverpool.ac.uk

arguments pro and con can be proposed, critiqued by counter arguments and evaluated. The work described in this paper applies argumentation in the context of classification. We have a dialogue between a number of participants regarding the classification of a given case (each agent charged with advocating a particular classification). In the case of binary classification this will take the form of a two-party dialogue. In the case of multi-class classification this will take the form of a multi-party dialogue, with agents representing each possible classification. This approach has been implemented in the PISA system, with the the participants realised as software agents, operating within a Multi-Agent System (MAS). With PISA we can thus classify instances in a variety of domains through dialogues conducted by sets of software agents.

The motivational application setting is a distributed data mining scenario where, for one of a number of reasons, the data should not be brought together into a single "data warehouse". In such a situation, each agent has its own personal data repository, containing records different from those held by other agents (i.e. the data sets are disjoint), which the agent can mine so as to generate arguments for its own classification and against the classifications proposed by other agents[1]. We refer to this classification paradigm as Multi-Agent Argumentation Based Classification from Experience (MABCE), since an agent's particular "experience" is reflected by its individual data repository. The locality of data may be necessary where, for example, individual records (perhaps because they contain personal information) are not permitted to be shared across agents. Using PISA only aggregated data is exchanged, and so individual privacy is respected. It is also desirable when the problem contains some relatively uncommon exceptions to a general rule of classification. These can be mined from an individual data set, when they would be lost if the data were pooled into a single large set.

One exemplar application domain used in this paper is *lay adjudication* such as that associated, in many countries, with the award of welfare benefits. In this setting adjudicators will typically deal with many cases, and will develop particular habits of classification. Error rates in such decision making are high, and this is often because rarely encountered exceptions are overlooked, and because some bad habits of interpretation can become ingrained in a particular group of adjudicators. Such welfare benefits are typically decided by a range of adjudicators working in several different regional offices, and different adjudicators and different offices will tend to encounter different types of case (e.g. some particular lung diseases are much more common in mining areas; some occupations requiring special treatment, for example trawler fishing, will be rarely encountered in inland areas, etc) and so the different offices will tend to develop different bad habits and blind spots. The high error rate encountered in the assessment of claims to welfare benefit is a significant problem [18,27,38,39]. The proposed MABCE system addresses this issue by allowing a dialogue between two or more agents representing different offices, with a

[1] Of course, the agent may also be able to find arguments *against* the classification it is advocating. These are, however, not used, except when considering whether to concede a point: it is the role of the other agents to put forward arguments against this classification.

view to moderating their decisions. This will then enable the different perspectives to be considered, and mistakes to be corrected. Another exemplar application is academic moderation. For marking projects and dissertations it is usual to use two or more independent markers. When there is disagreement in the initial assessments the assessors meet and attempt to justify their different marks. This process enables strengths and weaknesses initially over looked or under weighted to be identified and reconsidered. Usually a consensus will be possible: otherwise the reasons put forward in the moderation will need to be evaluated by a third party, whose task is eased by having the reasons made explicit in this way.

This paper provides a full overview of the proposed framework for Arguing from Experience and its application to multi-agent classification. Among the issues that must be considered when developing such a framework are:

1. The nature of a framework that will enable the envisioned argumentation from experience process whereby a collection of agents can reach an agreement about the classification of cases in some domain.
2. The means by which the discussion between different agents is facilitated.
3. The mechanism by which the agents can *agree* or *disagree* with each others arguments; and the effects of the underlying agreement model on the argumentation process.
4. The mechanisms required to address unbalanced and ordinal classification problems as well as standard classification problems.

Mechanisms to address these issues as realised in the PISA (Pooling Information from Several Agents) multiagent system are described and their evaluation form the subject of the remainder of this paper. The framework has been realised as part of the PISA (Pooling Information from Several Agents) multiagent system which is also described. The rest of this paper is organized as follows. Section 2 presents a review of some related previous work. In Section 3 an overview of the proposed argumentation framework is given, and the PISA realisation of this framework is described in Section 4. The operation of PISA is illustrated with a worked example in section 5. An extensive evaluation of MABCE, and PISA, using a variety of different datasets is presented in Section 6. Finally, a discussion and some conclusions are offered in Section 7.

2 Previous Work

Agent based techniques are widely applied to classification and machine learning tasks. Examples of the most mature systems, related to the work described in this paper, include the JAM[45] and BODHI[19] systems which provide agent based meta-learning strategies for classification. The literature also provides evidence that multi-agent approaches to classification can yield better results than other approaches ([44, 45, 47]). Multi-agent based classification typically involves a number of agents, each provided with a local dataset. In

some cases [34,37] the agents in the system have the same data, whereas in others, as in our work, the data sets are different (e.g. [45,19,47]). Additionally a number of approaches for sharing data in learning systems have also been proposed: [25] suggests an exchange of training examples among agents, [24] deals with limited information sharing in distributed clustering. In the PISA system each agent has a distinct local dataset, and arguments rather than examples or data are exchanged.

The main issue in multi agent classification is often considered not to be the classifier generation algorithms themselves, but the most appropriate mechanisms to allow agents to collaborate [26]. PISA provides such a mechanism for agent collaboration through argumentation. Other examples of agent based collaboration include [13], where the agents gradually join their datasets together according to a fixed distributed algorithm. Another example is [52] where groups of classifier agents learn to organise their activity so as to optimise global system behaviour. Both approaches assume that the agents cooperate to achieve a joint learning goal.

Agent technology has also been employed in *meta-learning*, the generation of a "global" classifier by combining a number of locally generated "base" classifiers [3,45]. One particular example of a meta-learning technique is ensemble learning (see [20] for a survey, and [21] for an experimental comparison of three techniques). Ensemble techniques have been shown to achieve good performance, especially in fields where the development of a powerful single learning system requires considerable effort [53]. Ensemble learning has been applied in the context of MAS. Generally, multiagent ensemble learning can be divided into two categories:

1. *Competitive ensemble learning*, where agents work asynchronously on the same problem and the decision of the best agent is the group decision.
2. *Cooperative ensemble learning*, where the group decision is a fusion or aggregation of the individual decisions of all agents involved.

PISA supports both techniques; the overall argumentation process presents a competitive ensemble approach, whereas the inter-group decision making procedure (discussed in Sub-section 4.2) is akin to the cooperative ensemble approach. In [41] it has been shown that an effective ensemble learning system may benefit from the combination of both techniques. These findings were also supported by the results of the experiment using PISA described in Section 6.

The proposed multiagent arguing from experience paradigm uses an Association Rule Mining (ARM) technique to produce arguments. ARM is a well established data mining technique developed in the early 1990s and first applied to super-market basket analysis [1,2]. It is focused on the discovery of relationships, called Association Rules (ARs) of the form $X \Rightarrow Y$ (where X and Y are disjoint subsets of some global set of attributes defined by the data). ARs are generated from the frequently occurring subsets of attributes, called *itemsets*, in a given binary valued input data set. An itemset is said to be *frequent* if its occurrence count (expressed as a percentage of the total number of records) and called its *support*, is greater than a specified *support threshold*.

For each frequently occurring subset, with a cardinality greater than one, two or more ARs can be generated. For example given a frequent subset $\{a, b\}$ the associations $a \Rightarrow b$ and $b \Rightarrow a$ can be generated. The relevance of an association is given by its *confidence* value, a percentage value calculated by dividing the support of the frequent item set from which the AR was generated by the support for the AR's antecedent. If the two supports are the same the confidence value will be 100%, indicating that every time the antecedent occurs, so does the consequent. PISA operates by applying a preference relation over ARs calculated from the support and confidence metrics. Note that all PISA agents conform to the same global support and confidence thresholds (it would not make sense for the agents to have different thresholds since it would bias the system in favour of the classifications advocated by agents with the more generous thresholds.).

There has been some work on the application of argumentation techniques to classification, notably the work of [12, 42, 46]. [12] present a multiagent *argumentation based* system system to reach agreements regarding cooperation and goal satisfaction founded on a *central facilitator* agent. In [42] an argumentation framework for learning agents is articulated. This framework has some similarities with the one proposed here in that it takes the experience, in the form of past cases, of agents into consideration. However, Case-Based Reasoning, rather than data mining, techniques are used to generate arguments. [42] also presented a framework for multiparty argumentation to enable a committee of agents to jointly deliberate about given cases where, unlike PISA, the communication between the arguing agents is direct (there is no mediator agent). An earlier example of un-mediated multi party argumentation can be found in [46], where turn taking is tokenised. When an agent receives an "attack", it informs the attacker whether it accepts the counter argument (and changes its prediction) or not. When an agent has the token it can answer to attacks by generating counter attacks. The communication between the agents continues in the same manner until they all agree on a prediction, or until a given number of rounds has passed during which no agent has generated any counterargument. If at the end of the argumentation the agents have not reached an agreement, then a voting mechanism that uses the confidence of each prediction as a weighting mechanism is used to decide the final solution. PISA has a very different mechanism which, as will be fully described in Section 4, uses an argumentation artefact (the argumentation tree) and a mediator agent to facilitate the argumentation process between a set of agents.

Privacy issues are playing an increasingly important role in emerging data mining applications. Some of the systems discussed above allow, or even require, the agents to reveal or share their local data with other agents to enhance the outcome of the classification process. Privacy preserving data mining, in contrast, permits the agents to communicate only high level statistics about their private data rather than communicating the raw data itself (e.g. [14, 35, 34]). PISA follows a similar pattern in that it maintains the privacy of each agent's local dataset; only generalisations of the data are exchanged during the dialogue, so that the privacy of the underlying data is not compromised.

3 The Argumentation Framework

As indicated above, the PISA framework for MABCE allows a number of agents to "argue" about the classification of a particular case. Each agent (or in some experiments, group of agents) argues for a particular classification and against other classifications. Arguments for or against a particular classification are made with reference to an agent's local data set (reflecting the individual experience of the agent). Each data set comprises a set of records such that each record describes a previously classified case. Each record consists of a set of attribute-value pairs and a single attribute-value pair indicating the classification (class) of that particular record. Note that agents need to have both records classified as belonging to the class they are proposing and records belonging to the classes advocated by other agents. This is because agents need to be able to generate rules to support all classifications, if they are to defend their own classification and to attack the classifications promoted by other agents. Every agent could, if called upon to do so, advocate any classification, but each is given the role of advocating one particular point of view, and so will advance the best rules it can find supporting that point of view. This is unaffected by the existence of rules indicating other outcomes: the role of individual agents is to find and challenge associations, not to do conflict resolution or decision making, which is accomplished by the system as a whole.

Arguments from Experience for a possible Classification (AECs) are the expressed in the form of ARs (as described in Section 2), generated using a ARM algorithm. Thus an AEC comprises an AR of the form $P \Rightarrow Q$, a support value s and confidence value μ, where P and Q are disjoint subsets of the global set of attribute-value pairs, and Q includes an attribute-value pair, c, indicating the classification of the example.

The validity of an argument is assessed according to the support s and confidence values μ. For ARs to be valid their support must be in excess of the specified support threshold ζ. The usage of the confidence measure is two fold. Firstly it represents the degree to which an individual agent believes that the current case should be classified as belonging to class c. Secondly it provides a means of giving weight to the associated AR: arguments with higher confidence are considered stronger than arguments with lower confidence. Good ARs are those whose confidence is a above a specified confidence threshold τ. Thus agents will seek to maximise confidence in AECs supporting the classification they are advocating, and will seek to minimise the confidence value of AECs supporting other classifications, preferably so that they drop below the confidence threshold.

Argumentation proceeds by agents exchanging arguments, in the form of AECs. Each AEC must promote the classification for which the agent is arguing. Either the AEC will represent an AEC promoting the agent's classification and with higher confidence than any so far proposed, or it will attack an AEC proposed by another agent. AECs may attack one another in three different ways.

We denote an AEC promoting a classification c as aec_c. Given two AECs aec_c and $aec_{\acute{c}}$, there are three possible attack relationships between them:

1. $aec_c \rightharpoondown aec_{\acute{c}}$ (Distinguishing attack) if $c = \acute{c}$, $\mu < \acute{\mu}$ and $Q \supset \acute{Q}$. In this case the attacking agent is reducing the confidence value associated with the rule proposed by the attacked agent by adding items to the consequent. The idea here is to attack the classification by indicating that some property normally associated with that classification does not hold, and so reducing confidence.
2. $aec_c \rightharpoonup aec_{\acute{c}}$ (Enhancing attack) if $c = \acute{c}$, $\mu < \acute{\mu}$ and $P \supset \acute{P}$. Here the attacking agent is decreasing the confidence by adding items to the antecedent. This is typically used to reduce confidence by arguing that the current case has particular features which make it an exception to the general rule represented by $aec_{\acute{c}}$
3. $aec_c \rightarrow aec_{\acute{c}}$ (Counter attack) if $c \neq \acute{c}$ and $\mu > \acute{\mu}$. Here the attacking agent is proposing an alternative classification with a higher confidence than that proposed by the attacked agent. This does not contest the reason offered by $aec_{\acute{c}}$, but instead rebuts it with a stronger reason.

We say that aec_c *is intended against* $aec_{\acute{c}}$ if any of these three attack relations between the two AECs exists. We also distinguish between *direct* and *indirect* attacks. A direct attack is an attack against a specific argument placed by another agent, an indirect attack is an attack against an argument which happens to be made by an argument introduced to make a direct attack against some other agent's argument.

4 The PISA System for MABCE

The above argumentation framework has been realised in the PISA (Pooling Information from Several Agents) system. The key idea of PISA is that the proposed dialectical process will enable any number of software agents, each with their local repository of experience (in the form of disjoint tabular datasets), to argue with each other in order to reach an agreement or decision with respect to the classification of some hitherto unseen case. Agents withdraw from the dialogue when they can no longer generate any more arguments. The dialogue ends when there is only one agent left, or a tie situation is reached.

Central to the PISA framework is the *chairperson agent* (CPA). This is a neutral "mediator" agent [40] which performs a variety of administrative tasks to facilitate MABCE dialogues. The CPA has a number of responsibilities: (i) *starting* and *terminating* a dialogue involving a set of participants to classify a given case; (ii) maintaining the *argumentation tree* (the storage structure used by PISA which is discussed further in Sub-section 4.3, (iii) allowing agents to join or leave a dialogue; and (iv) where there is a tie situation, initiating a tie-resolution mechanism. (Although only a simple tie resolution process is used in this paper, a number of possible mechanisms are fully discussed in [50]).

Give a previously unseen instance (φ) that requires classification and a number of PISA participating agents equivalent to the number of possible classifications such that each is assigned the role of promoting one of the possible classifications, the PISA dialogue proceeds as follows:

1. The CPA randomly selects one participant agent, pa_1, from the available set of participant agents PA ($pa_1 \in PA$), to start the dialogue and pa_1 proposes an argument $aec_1 = aec_{c_{p1}}$ with confidence $\mu_1 \geq \tau$ (where τ is the chosen confidence threshold). A new argumentation tree (Ψ) is initiated with aec_1 at its root. If pa_1 is unable to play an opening move (because it cannot generate an appropriate argument), the CPA selects another participant to commence the dialogue. If all the participants fail to propose an opening argument, the dialogue terminates with failure.
2. In the second round the other participant agents attempt to attack aec_1. If none of the participants can generate an appropriate attacking argument of any kind the dialogue terminates, and the case is classified according to aec_1. Otherwise, the argumentation tree data is updated with all the submitted attacks.
3. Before the beginning of each of the subsequent rounds, the chairperson excludes any *dormant* agents from further participation in the dialogue. A dormant agent is one that has not taken part in the last m rounds of the dialogue. Normally m will be greater than 1, allowing agents to choose to sit out a round for strategic purposes. If only one agent remains then the dialogue is terminated, the remaining agent is the winner, and the case is classified accordingly. Otherwise, any participant who can play a legal move may do so; and the argumentation tree data structure is updated with all the attacks submitted.
4. If two consecutive rounds pass without any new moves being submitted to the argumentation tree, or if some predetermined number of rounds have passed without reaching an agreement, the dialogue terminates. If no winner can be identified, a tie-break mechanism is invoked. Otherwise, the case under discussion is classified according to the classification proposed by the winner.

4.1 Agreement Model

Every agent $pa \in PA$ has an *agreement model* to decide whether to accept an argument placed by another agent. The *standard agreement model* is founded on two confidence thresholds, (τ is the confidence threshold for proposing an agument):

- π_{up}: The confidence threshold for accepting arguments for a classification, $\pi_{up} \geq \tau$
- π_{down}: The confidence threshold for accepting arguments against a classification, $\pi_{down} \leq \tau$

These thresholds mean that if an argument is sufficiently strong the agents will accept that clasification, and if an argument reduces the confidence in an association giving the classification it is proposing sufficiently the agent will accept that the association must be rejected. The standard agreement model assumes that PISA agents will accept, where possible, arguments proposed by any of the opponent agents (and groups); and will launch attacks only against arguments that they cannot agree with. However, agents may prefer, for strategic reasons, to agree with certain other participants and not with the rest. This style of agreement is referred to as "*Biased Agreement*". This can usefully be deployed in a ordinal multi-class classification scenario.

In ordinal classification it is assumed that the set of class labels can be ordered in some manner, whereas traditional classification paradigms usually assume that the different classification values are unordered. For many practical applications classification labels tend to exhibit some form of order (e.g. the weather can be cold, mild, warm and hot). Given ordered classes, one is not only concerned to maximise the classification accuracy, but also to minimise the distance between the actual and the attributed classifications. In the case of ordinal classification we do not simply wish to agree or disagree with arguments placed by other agents but want to weight this agreement according the "proximity" of the classification argued for by another agent with respect to our own position. To do this a *Biased Agreement Model* may be adopted. In this case that an agent *pa* will have its own class c_{pa} and an ordered list of alternative classes C_{pa} that it is prepared to agree with because these are close to its own position. We can identify two types of biased agreement: (i) No-Attack biased agreement (NA-BIA): *pa* agrees with any aec_c if $c \in C_{pa}$ and $\mu_{aec_c} \geq \tau$. (ii) Threshold Check biased agreement (TC-BIA): *pa* agrees with aec_c if $c \in C_{pa}$ and $\mu_{aec_c} \geq \pi_{up}$.

4.2 Participation Groups: Coalitions and Teams in PISA

PISA also allows agents to operate in groups. In its simplest form this allows a number of agents to propose each classification. In this case the agents form *Participation Groups* such that each group represents a possible classification. These groups persist throughout a PISA dialogue. However, in a more advanced setting, agents may leave a group if they change their classification objective (see below). Teamwork has been a focus of much research in the fields of distributed AI and MAS. Several authors identify an *agent team* as consisting of a number of cooperative agents which have agreed to work together toward a common goal [29]. The advantage offered by teams is that the combined resources of the team can be directed at some common goal (which may have a higher collective utility), rather than furthering the utility of the individual members. Thus, the notion of participation groups in PISA can be likened to teamwork.

In PISA participation groups there are two roles: team leader and regular member. There are various mechanisms that we may adopt to identify a team

leader and these are discussed in [50]; the most straightforward approach is to simply select the agent with the largest data set (i.e. the most extensive experience). A similar notion of leadership can be found in (for example) [6,31] where the leading agent acts as a representative and intermediary for the group as a whole. The regular members, as well as the leader, generate arguments as described above. The leader then selects one of the suggested arguments. The leader also : (i) guides the *inter-group dialogue process*, (ii) may redirect attacks at opponents other to those suggested by individual group members, or (iii) insist on a particular strategy. The intuition behind groups in PISA is that each of the group members will generate the best possible argument according to their experience/strategy. This then allows the group to benefit from the different arguments suggested by its members. Further details regarding the nature of groups in PISA can be found in [50]).

Agents can also form *dynamic coalitions*. Dynamic coalitions are temporary groupings whereby two or more PISA agents agree to cooperate against one or more other opponents. Agents in a coalition do not attack each other, but concentrate on attacking arguments placed by agents outside the coalition. The objective of a coalition is to attempt to eliminate agents representing dominant classes from the dialogue. Once the agents in question have been removed (when they have not participated for a number of rounds), the coalition is dismantled and the agents go on to argue for their own particular positions in the normal PISA manner. We propose two ways on which agent coalitions may be dismantled:

1. *Dynamic Coalition Termination 1 (DCT1)*: the coalition is dismantled if the agent supporting the dominant class does not participate in the dialogue for two consecutive rounds. Here that agent is not removed from the dialogue, and the same coalition may be formed again against the same opponent.
2. *Dynamic Coalition Termination 2 (DCT2)*: the coalition is dismantled if the agent supporting the dominant class does not participate in the dialogue for two consecutive rounds. Here, however, the agent is removed from the ongoing dialogue (and so the case will not be classified according the class label it supports).

4.3 Argumentation Tree

The central PISA data structure is the *Argumentation Tree* (Ψ). This tree acts as a mediating artefact for the dialogue [40]. The nodes in the tree represent arguments, and the arcs attacks (child nodes attack parent nodes). The tree uses a four colour coding to mark the status of the arguments played so far. Nodes are either green or blue when introduced: green if they propose a new AEC; or blue if undermining (making a distinguishing or enhancing attack on) an existing AEC. Blue and green nodes remain blue or green until they are defeated: thus blue and green nodes in the argumentation tree indicate undefeated nodes and so are potentially winning arguments. Red nodes are those

From	To			
	G	R	B	P
G		If attacked by an undefeated node		If indirectly attacked by another undefeated green node with higher confidence
R	If all attacking nodes have been defeated and original colour was not blue		If originally green and all attacking nodes have been defeated and original colour was blue	If directly attacking nodes are defeated, but at least one indirect attack remains
B		If attacked by undefeated nodes		
P	If all current attacks (direct and indirect) successfully defeated	If attacked by undefeated nodes		

Table 1 Colour changing regime.

directly under attack and purple nodes are those indirectly attacked. Nodes change their colour according to Table 4.3. Blue and green nodes represent distinct types of arguments (against and for a classification, respectively) and so a green node cannot change to blue and vice versa. The same applies to changing from purple to blue and vice versa, as each purple node represents an argument for a classification, which has been undermined by an undefeated stronger (higher confidence) argument for a different classification. Arcs are labelled as being *explicit*, indicating direct attacks; or *implicit*, indicating indirect attacks. The issue of which agent is being addressed by an utterance that arises in multiparty dialogues is resolved via the direct links. An argument is addressed to the argument it attacks (or counter attacks), except for the opening argument which is addressed to all other participants. Legal PISA arguments are those that change the colouring of the tree (Table 4.3). Additionally, PISA applies certain rules to prevent the repetition of arguments so that "cyclic" behaviour cannot occur (a summary of these rules can be found in [49]).

The argumentation process terminates when it is no longer possible for any of the participating agents to pose new arguments. The outcome is measured according to the number N of undefeated (green or blue) nodes and the set of classes, C, that are represented (note that $|C| \leq N$), as follows:

1. If $N = 0$: The dialogue has failed (there has been no dialogue!).
2. If $N = 1$: There is only one undefeated node in which case their is a clear "winner" and the current case is classified according to label c ($C = \{c\}$).
3. If $N > 1$ and $|C| = 1$: All undefeated nodes argue for the same class and the current case is classified according to label c ($C = \{c\}$).
4. If $N > 1$ and $|C| > 1$: A *Tie* situation exists.

Where a tie situation exists PISA implements a tie resolution mechanism. We have identified various possible tie resolution mechanisms [50]. For example we can repeat the MABCE process with the tied parties, adopt a voting strategy or simply adopt a random resolution. The latter is akin to using a *default rule* and has, for simplicity, been used in the remainder of this paper.

5 Worked Example

To ilustrate the operation of PISA a worked example is presented in this section. In the example PISA is applied to a housing benefit scenario (the exemplar motivational application for the proposed MABCE approach) where a Retired Persons Housing Allowance (RPHA)[2] is payable to persons who are of retirement age, 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 the UK or absent only by virtue of "service to the nation" (e.g. armed forces), and should have an established connection with the UK labour force. These legislative conditions need to be interpreted and applied by those adjudicating claims for RPHA benefits, and so these abstract descriptions need to be expressed in terms of ascertainable facts, typically using a set of guidelines (e.g. [8,9]). The following interpretations were used:

1. *Age condition*: the pensionable age is 60+ for women and 65+ for men.
2. *Income condition*: means that housing costs should exceed one fifth of candidates' available income to qualify for the benefit.
3. *Capital condition*: is interpreted as below the threshold set for another existing benefit.
4. *Residence condition*: is interpreted as having a UK address.
5. *Residence exception*: is interpreted as being a member of the armed forces.
6. *Established contribution condition*: is interpreted as having paid contributions in 3 of the last 5 years.

The outcome of an application can fall into one of four classes:

- *(Fully) Entitled*: Candidates are entitled to a full RPHA allowance if they satisfy all the above conditions.
- *Entitled with Priority*: Candidates are entitled to full allowance with priority if they satisfy the above and also one of the following: (i) they have paid contributions in four out of the last five years and either have significantly less capital than the original limit (interpreted as 1000 less than the original limit) or have significantly less income than the original limit (by 5%), or (ii) they are members of the armed forces and have paid contributions in all five out of the last five years.

[2] Although fictional, this scenario uses a number of typical conditions found in welfare benefits legislation.and has been used in several AI and Law applications, e.g. [36].

- *Partially Entitled*: Candidates are entitled to a lower rate of benefit if they satisfy the age condition, and while they do not satisfy the original conditions, they do satisfy one of three somewhat weaker conditions. That is, they either: (i) have only slightly more capital than the original limit (+1000 more than the limit), but have paid contributions in at least 4 out of the last five years, (ii) have slightly more available income (+5%) than the original limit, but have paid contributions in 4 (or 5) years out of the last five, or (iii) are employed in the Merchant Navy and have paid contributions in five out of the last five years.
- *Not Entitled*: The candidate fails to satisfy any of the above.

Our experiment supposed that there are four different offices providing RPHA services in four different geographical regions, each with a dataset of 6,000 benefit records. Each dataset was assigned to an agent. Thus a total of four agents can engage in dialogues regarding the classification of RPHA applicants, each agent advocating one of the four possible classifications described above[3]. The four agents in the following example are referred to as PR (priority entitled), EN (entitled), PE (partially entitled) and NE (not entitled), according to the classification they are defending. Support and confidence thresholds of 1% and 50%, respectively were used when mining ARs. PISA was then applied to the case of: *a 63 years old female applicant, who satisfies all the benefits conditions and has served in the armed forces and has paid her contribution in each of the past five years.* This case should classify as entitled to priority benefits.

The chairperson invites **EN** to propose the opening rule (argument). Accordingly EN suggests the following association ($N1$): *£2000≤capital≤£3000, 15%≤Income≤20%→ Entitled. confidence = 67.54%.* This initial argument is then attacked by the other three agents in the second round (Figure**??**). All three of them are able to find an association with confidence greater that the that proposed by EN, and so they may make counter attacking moves as follows:

- NE - Counter Attack ($N2$): *60≤Age≤65 and 15%≤Income≤20%→Not Entitled. confidence = 69.0%.*
- PE- Counter Attack ($N3$): *Year1 = paid, Year5 = paid→Partially Entitled). confidence = 68.4%.*
- PR - Counter Attack($N4$): *Residency = armed forces, Year1 = paid→Priority Entitled. confidence = 75.4%.*

Note that NE can use the fact that the case under discussion is of a candidate whose age is between 60 and 65 years to attack the EN argument, since all the male applicants in this age group are not entitled to the benefit, and so there is a strong association provided gender is ignored. At this stage PR is "winning" as it has the best un-attacked rule.

[3] Of course each dataset will contain reasons to support each of the four classifications. The agents will use the data to put forward arguments supporting their classification and to argue against (or accept) the arguments placed by other agents.

In Round 3 all four players make moves, again proposing associations with higher confidence that the currently winning rule:

- EN, PE and NE proposes a new rule to attack the current best argument (undefeated node ($N4$)):
  1. EN - Proposes a new AEC ($N5$): *Gender = female, 60≤Age≤65, £2000≤capital≤£3000, 15%≤Income≤20%→Entitled. confidence = 78.6%.*
  2. NE - Counter Attack($N6$): *60≤Age≤65,£2000≤capital≤£3000 and 15%≤Income≤20% →Not Entitled. confidence = 75.99%.*
  3. PE - Counter Attack($N7$): *Year1 = paid, Year2 = paid, Year5 = paid→Partially Entitled). confidence = 76.0%.*
- PR strengthens the position of $N4$ by adding an additional condition: ($N8$):*Residency= armed forces, 15%≤Income≤20%, Year1=paid→ Priority Entitled. confidence = 76.2%.*

Now EN has the rule with the currently highest confidence, and so is back in the lead. Note that NE has again played an AEC based on the age of the candidate to try and persuade the other participants to not issue any benefit to this candidate, but this is the last move this participant is able to play to attempt to establish its position. In the fourth round NE and EN make no moves, because NE has nothing available and EN is currently ahead and cannot do better that $N5$. The other two agents can, however, make moves against the current winning position as follows:

- PE - Counter Attack ($N9$): *Year1 = paid, Year2 = paid, Year4 = paid, Year5 = paid→Partially Entitled. confidence = 80.4%.*
- PR - Counter Attack ($N10$): *Residency=armed forces, 15%≤Income≤20%, Year1 = paid→. confidence = 87.3%.*

The AEC proposed by PE is based on the fact that people who have paid contributions in four of the last five tears are often not classified as entitled: either they fail some other condition, or they qualify for Priority Entitlement. The last round concludes this example, since none of the other agents can challenge this AEC. Note that PR has managed to win the dialogue, having found an association with significantly higher confidence that that originally proposed, and so the resulting classification is correctly identified as priority entitled. In the example the agents all adopted a positive strategy, proposing associations supporting the classification which they are advocating. Had they been adopting a more critical strategy, distinguishing attacks based on gender would, for example, have reduced the confidence of the rules proposed by NE, and, had it been necessary, distinguishing attacks could have been found to reduce the confidence of $N9$.

Let us now assume the situation where three of the offices are located in areas where there are not many applicants from members of the armed forces such that the percentage of applicants from the forces is no more than 10% (600 cases of the 6000 records). In contrast the fourth office, which serves a number of military bases, receives a very high percentage of applications from

members of the armed forces, equalling 70% of the records in its dataset (4200 cases out of 6000)[4]. Let us re-apply the same scenario as in the previous example (using the same setup). Here we assume that the PA agent is located in the fourth office. The chairperson again invites EN to propose the opening rule (argument). EN suggests the following association ($N1$): *£2000≤capital≤£3000, 15%≤Income≤20%, Year1 = paid→Entitled. confidence = 62.4%.*

The initial rule is attacked by the other three agents in the second round, as follows:

- NE distinguishes EN's argument by demonstrating ($N2$) that adding the attribute 60≤age≤65 only gives Entitled with a confidence of 19.9%.
- PR and PE propose counter attacks as follows:
  - PE ($N3$):*Gender = Female, Year4= Paid, Year5 = paid→Partially Entitled. confidence = 72.5%.*
  - PR ($N4$):*Residency = armed forces, Year1 = paid, Year2 = paid→Priority Entitled. confidence = 77.8%.*

At this stage PR is winning as it has the only undefeated argument ($N4$). In Round 3 all four players make moves:

- NE distinguishes PR's argument ($N4$) by demonstrating ($N5$) that adding the attributes 60≤age≤65 and Gender = Female gives Entitled with a confidence of only 26.5%.
- PE and EN propose counter attacks against the current best rule ($N4$):
  - EN ($N6$):*Gender = female, 60≤age≤65, £2000≤Capital≤£3000 and 15%≤Income≤20%, Year1 = Paid→Entitled. confidence = 78.6%.*
  - PE ($N7$):*Gender = Female, Year3= Paid, Year4= Paid, Year5 = paid→Partially Entitled. confidence = 76.0%.*
- PR improves on $N4$ with $N8$: *Residency=armed forces, Year1 = paid, Year2 = paid, Year3=paid→Priority Entitled. confidence = 79.9%.*

At the end of Round 3 PR is still in the lead. In Round 4 all agents make moves:

- NE distinguishes PR's argument from Round 3 by demonstrating ($N9$) that adding the attribute £2000≤Capital≤£3000 only gives Entitled with a confidence of 28.2%.
- EN proposes a new AEC to attack the current best argument ($N8$) - $N10$: *60≤Age≤65, £2000≤Capital≤£3000, Year1= Paid, Year2= Paid, Year3= Paid→Entitled. With confidence = 81.2%.*
- PE proposes a counter attack against $N8$ - $N11$: *Gender= Female, 15%≤Income≤20% , Year2= Paid, Year3= Paid, Year4= Paid, Year5= paid→Partially Entitled. With confidence = 80.4%*
- PR further enhances the confidence of its previous argument ($N8$) as follows ($N12$): *The case has the **additional** feature: Year3= Paid→Priority Entitled. confidence = 87.3%.*

[4] This situation is by no means improbable: all applications from persons stationed abroad and seeking to relocate in the UK might well go to the same office.

So PR is still in the lead. NE has again played an undermining move. In Round 4 all agents have moves:

- NE distinguishes PR's argument from the first round by demonstrating ($N13$) that adding the attributes 15%≤Income≤20% and £2000≤Capital≤£3000 only gives Entitled with a confidence of 37.35%.
- EN distinguishes PR's argument from the first round by demonstrating ($N14$) that adding the attribute £2000≤Capital≤£3000 only gives Entitled with a confidence of 34.2%.
- PE Proposes counter attack ($N15$): *Gender = Female, 15%≤Income≤20%, £2000≤Capital≤£3000, Year2= Paid, Year3= Paid, Year4= Paid, Year5 = paid→Partially Entitled. confidence = 88.5%.*
- PR attacks $N9$ ($N16$): *Gender=Female, Residency=armed forces, £2000≤Capital≤£3000, Year1= paid, Year3= paid, Year4= paid→Priority Entitled. confidence = 87.3%.*

Note that both EN and NE have made distinguishing attacks against PR's move from round 1, each using different attributes from the case under discussion. However PR is still in the lead with $N12$. The final round of the dialogue is as follows:

- NE distinguishes PR's argument from the previous round by demonstrating that adding the attribute 60≤Age≤65, and 15%≤Income≤20% only give Entitled with a confidence of 32.2%.
- EN distinguishes PR's argument from the first round by demonstrating that adding the attributes Year2 = paid and 15%≤Income≤20% only gives Entitled with a confidence of 28.3%.
- PE distinguishes PR's argument from the previous round by demonstrating that adding the attribute Year5 = paid and 15%≤Income≤20% only gives Entitled with a confidence of 44.2%.
- PR attacks $N14$: *60≤Age≤65, Gender=Female, Residency=armed forces, 15%≤Income≤20%, £2000≤Capital≤£3000, Year1= Paid, Year2= Paid, Year3= Paid Year4= Paid, Year5= Paid → Priority Entitled. confidence = 99.54%.*

The last round concludes the dialogue. Note that the argument made by PR has a very high confidence and cannot be distinguished as it makes use of all the attributes in the case under discussion. PR thus wins the dialogue and the case is classified as Priority Entitled. Because the case represents quite special circumstances, it has been necessary to produce a very specific rule to decide it. PE, EN and NE would have been unable to classify this case correctly on their own using a conventional classification approach as they do not have sufficient data on applicants from the armed forces in their dataset to sufficiently support the rule needed to govern this case. Similarly there would have been insufficient examples in a database formed by combining all the individual datasets to support the correct classification. However, PR can also mis-classify cases of a different nature as a result of its high percentage of applications from armed forces personnel. The correct classification can,

however, be found by applying the proposed MABCE process between the four offices using their datasets separately.

6 Evaluation

In the foregoing sections the proposed MABCE framework, as realised in the PISA system, was described and illustrated. This section presents an evaluation of the proposed approach. We commence, Sub-section 6.1, by considering the operation of the system with respect to the standard mutli-class classification problem and compare the operation of PISA with a number of alternative classification paradigms, especially ensemble approaches. We then, Sub-section 6.2, present an evaluation of the operation of PISA when groups of agents, as defined in Sub-section 4.2 above, are used and show how this serves to improve MABCE performance. Both these evaluations use the standard agreement model (Sub-section 4.1), but in Sub-section 6.3 the operation of PISA using the biased agreement model (also described in Sub-section 4.1) is considered for the ordinal classification problem. Recall thatPISA can operate with both participation groups and dynamic coalitions as described in Sub-section 4.2: the use of dynamic coalitions with respect to the unbalanced multi-class classification problem is evaluated in Sub-section 6.4. A major advantage of the MABCE approach, and a reason why the MABCE approach can out perform other approaches, is that it is very resilient to the presence of "noise" (erroneous data in the datasets). This is illustrated in Section 6.5 where the results of experiments using increasing amounts of noise are presented. The ability to cope with noise is important in a number of applications. In particular, the exemplar application of welfare benefits requires this ability, since the high error rates in classifying cases means that sample data will invariably include a substantial number of misclassified cases.

A number of datasets, drawn from the UCI repository [10], were used for the evaluations as well as the housing benefit data set used in the worked example presented in Section 5. Where necessary continuous values were discretised into ranges. The chosen datasets (Table 2) displayed a variety of characteristics with respect to number of records (R), number of classes (C) and number of attributes (A). Importantly they include a diverse number of class labels distributed in different manners (balanced and unbalanced); thus providing the desired variation in the experience assigned to individual PISA participants.

6.1 Multiagent Classification

In order to provide an empirical assessment of the application of PISA in the context of standard classification problems a series of experiments was conducted designed to evaluate the hypothesis that applying the proposed MABCE process to classification produces results that are better than or, failing that, comparable to those obtained using more traditional classification techniques. The traditional techniques considered were:

Name	R	C	A	Bal
Hepatitis	155	2	19	no
HorseColic	368	2	27	no
Cylnder Bands	540	2	39	yes
Pima (Diabetes)	768	2	9	yes
Mushrooms	8124	2	23	yes
Iris	150	3	4	yes
Wine	178	3	13	yes
Lymphography	148	4	18	no
Heart	303	5	22	no
Dematology	366	6	49	no
Zoo	101	7	17	no
Glass	214	7	10	no
Ecoli	336	8	8	no
Led7	3200	10	8	yes
Chess	28056	18	6	no
Ionosphere	351	2	34	no
Cong. Voting	435	2	17	yes
Breast	699	2	11	yes
TicTacToe	958	2	9	no
Adult	48842	2	14	no
Waveform	5000	3	22	yes
Connect4	67557	3	42	no
Car Evaluation	1728	4	7	no
Nursery	12960	5	9	no
Annealing	898	6	38	no
Autombile	205	7	26	no
Page Blocks	5473	7	11	no
Solar Flare	1389	9	10	no
Pen Digits	10992	10	17	yes

Table 2 Summary of datasets. Columns indicate: domain name, number of records, number of classes, number of attributes and class distribution (approximately balanced or not).

1. *Decision trees*: C4.5 as implemented in [28], and the Random Decision Tree (RDT) as implemented in [15], were used.
2. *Classification Association Rule Mining (CARM)*: The TFPC (Total From Partial Classification) algorithm [16] was adopted because this algorithm utilises similar data structures [17] as PISA.
3. *Ensemble classifiers*: Table 3 summarises the techniques used. We chose to apply Boosting and Bagging, combined with decision trees, because previous work has demonstrated that this combination is very effective (e.g. [7, 43]).

C4.5 and CARM were selected because they represent well understood "centralised" approaches to classification. The comparison with ensemble methods was undertaken because, in many respects, multi-agent classification can be said to operate in a similar manner in that both approaches "pool" results to produce a better classification. For the purposes of running PISA, each training dataset was equally divided among a number of Participant Agents corresponding to the number of classes in the dataset. Then a number of PISA dialogues were executed to classify the cases contained in test sets[5]. The results presented throughout this sub-section were obtained using *Tenfold Cross Validation* (TCV). The standard agreement model was used with $\pi_{up} = 70\%$ and

[5] For each evaluation the confidence threshold used by each participant was 50% and the support threshold 1%, chosen as the thresholds most comonly used in the literature.

Ensemble	Technique	Decision Tree
Bagging-C4.5	Bagging [11]	C4.5 (S=1%)
Bagging-RDT	Bagging [11]	RDT (S=1%)
ADABoost-C4.5	ADABoost.M1 [22]	C4.5 (S=1%)
ADABoost-RDT	ADABoost.M1 [22]	RDT (S=1%)
MutliBoostAB-C4.5	MultiBoosting [51]	C4.5 (S=1%)
MultiBoostAB-RDT	MultiBoosting [51]	RDT (S=1%)
DECORATE	[33]	C4.5 (S=1%)

Table 3 Summary of the Ensemble Methods used. The implementation of these methods was obtained using WEKA [28]. (S=Support, RDT=Random Decision Trees).

Dataset	PISA	Ensembles							Dec Trees		TFPC
		Bagging		ADABoost		MultiBoost		Dec			
		C4.5	RDT	C4.5	RDT	C4.5	RDT		C4.5	RDT	
Hepatitis	**13.3**	18.1	14.8	15.5	21.3	13.5	18.7	16.1	16.1	23.2	18.0
Ionosphere	3.3	7.7	6.8	7.1	10.8	6.3	10.8	7.4	8.6	**2.6**	14.3
HorseColic	**2.8**									3.9	22.8
Congress	1.8	3.0	2.3	2.1	3.0	2.1	3.0	2.8	4.2	**0.0**	9.3
CylBands	**15.0**	42.2	27.1	42.2	34.8	42.2	34.1	39.8	42.2	36.5	30.4
Breast	**3.9**	5.1	4.9	4.8	4.8	4.9	4.9	5.4	4.9	5.1	10.0
Pima	**14.5**	27.2	25.3	25.3	23.8	25.1	24.9	25.7	26.7	16.2	25.9
TicTacToe	2.8	7.2	5.4	**2.2**	20.4	**2.2**	20.4	5.9	15.5	20.8	33.7
Mushrooms	0.4	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.1	1.1
Adult	14.5									**13.1**	19.2
Iris	**2.7**	4.7	5.3	6.0	7.3	6.0	7.3	4.7	4.0	8.0	6.0
Waveform	**2.2**	17.9	11.9	21.5	21.5	13.6	11.9	21.5	21.5	2.4	33.3
Wine	1.2									**0.0**	25.3
Connect4	5.1									**4.3**	34.2
Lympho	**6.2**	18.9	19.6	14.9	29.7	15.5	29.7	19.6	22.9	25.0	24.3
Car Eval	4.1	4.5	**1.2**	2.4	6.3	2.6	6.3	4.3	5.1	5.9	30.0
Heart	5.1	20.1	19.8	22.8	21.1	19.1	19.7	20.4	19.1	**4.7**	46.7
Page Bloc	**2.2**	6.9	6.9	7.1	6.9	7.1	6.9	6.9	7.1	6.9	9.9
Nursery	6.4	2.08	3.09	**0.38**	3.09	0.35	3.09	1.91	2.62	3.72	22.3
Dematology	4.9	4.1	3.6	3.8	15.3	3.3	15.3	**1.6**	6.1	5.3	25.0
Annealing	9.6	1.2	0.7	0.5	1.8	**0.6**	1.8	1.3	1.6	1.7	11.8
Zoo	9.9	7.9	4.9	3.9	19.8	3.9	19.8	6.9	7.9	**0.0**	8.0
Auto	**12.0**	15.1	15.6	14.2	21.5	15.6	21.5	16.1	18.1	17.0	29.0
Glass	**14.7**	27.1	21.5	22.4	29.9	25.2	29.9	29.9	33.2	29.9	33.8
Ecoli	5.2	13.99	15.2	16.4	24.7	14.9	24.7	13.1	15.8	8.8	37.3
Flare	6.1	**2.5**	3.41	3.4	3.4	3.1	3.1	3.1	**2.5**	8.0	14.7
Led7	**12.0**	24.8	24.2	24.9	24.3	24.9	24.3	24.8	24.8	24.3	31.0
Pen Digit	2.8	4.47	**1.4**	1.6	2.5	5.1	1.9	2.5	5.7	1.1	18.2
Chess	**9.1**									18.6	15.7

Table 4 Test set Error Rate (ER) (%). Values in **bold** are the lowest in a given dataset. Dec Tree = Decision Trees, Dec=Decorate and ADABoost=ADABoost.M1

π_{down} = 25%. ADABOOST/ADABoost.m1 and Multiboosting TCVs were executed using 10 iterations eight mass to build the default 100 classifiers. Bagging was executed using the default number of iterations (10). The size of each bag was a 100 cases (default). DECORATE was applied using WEKA default setup.

For each of the included methods (and PISA) three evaluation metrics were recorded with respect to dataset: (i) classification *Error Rate* (ER), (ii) *Balanced Error Rate* (BER) using a confusion matrix obtained from each TCV[6]; and (iii) *execution time*. These three values then provided the criteria for assessing and comparing the classification paradigms.

[6] Balanced Error Rates (BER) were calculated, for each dataset, as follows:

$$BER = \frac{1}{C\sum_{c}^{i=1}\frac{F_{ci}}{F_{ci}+T_{ci}}}$$

where C = the number of classes in the dataset, T_{ci} = number cases correctly classified as ci, and F_{ci} = number ci cases which were incorrectly classified.

Dataset	PISA	Ensembles							Dec Trees		TFPC
		Bagging		ADABoost		MultiBoost		Dec			
		C4.5	RDT	C4.5	RDT	C4.5	RDT		C4.5	RDT	
Hepatitis	**12.0**	27.4	20.6	23.4	33.7	19.9	25.1	24.6	23.4	38.2	36.4
Ionosphere	4.6	7.1	6.6	6.4	11.4	5.3	11.4	7.1	8.2	**2.2**	13.4
HorseColic	**2.8**									3.7	28.6
Congress	2.4	3.4	2.7	2.3	3.2	2.3	2.3	3.1	4.7	**0.0**	9.7
CylBands	**14.5**	46.1	24.5	46.1	35.6	46.1	35.6	40.1	46.1	34.6	32.8
Breast	**4.8**	6.0	6.2	6.2	6.2	6.1	6.2	6.7	6.2	4.7	12.9
Pima	**13.9**	28.9	26.9	26.9	25.2	26.7	26.1	27.2	28.3	24.5	33.7
TicTacToe	**2.1**	6.7	5.4	**2.3**	22.5	2.3	22.5	5.3	16.9	22.9	47.4
Mushrooms	0.6	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.1	1.0
Adult	**8.8**								17.8	39.9	
Iris	**2.9**	4.6	5.3	5.9	7.3	5.9	7.3	4.7	3.7	7.7	6.1
Waveform	**3.9**	18.0	11.9	21.5	21.5	13.6	11.9	21.5	21.5	2.4	33.4
Wine	1.4									**0.0**	24.1
Connect4	11.0									**5.3**	66.7
Lympho	**15.9**	30.1	9.7	25.9	43.4	38.7	43.4	39.4	35.9	47.1	16.1
Car Eval	8.2	11.2	6.8	**4.8**	10.4	5.3	10.4	10.2	16.6	10.7	75.0
Heart	8.3	9.2	8.9	9.9	7.9	**7.9**	8.9	9.4	7.9	9.8	48.0
Page Bloc	**9.5**	21.5	22.9	27.9	21.5	27.9	21.5	22.9	27.9	21.5	19.9
Nursery	5.5	4.1	2.3	1.1	5.8	**0.8**	5.8	4.6	5.9	5.7	40.1
Dematology	8.5	4.7	3.9	3.9	19.4	3.3	19.4	1.8	7.0	**3.3**	61.7
Annealing	16.1	6.8	3.9	**2.6**	4.3	3.3	4.3	7.2	6.8	4.4	33.5
Zoo	13.2	12.8	10.71	10.7	36.5	10.7	36.5	15.7	17.5	**0.0**	17.1
Auto	12.3	**11.4**	15.9	10.6	18.9	15.9	18.9	12.8	17.0	13.6	19.6
Glass	**16.1**	24.6	19.4	24.3	29.6	23.2	29.6	29.6	37.9	29.6	48.6
Ecoli	16.2	36.7	40.2	41.2	51.9	37.9	51.9	24.2	43.4	**9.4**	23.2
Flare	17.2	12.7	12.7	10.9	12.7	12.7	12.6	12.6	12.5	**7.6**	14.7
Led7	**11.8**	24.6	24.1	24.9	24.3	25.1	24.2	24.9	24.7	24.4	31.4
Pen Digit	3.5	4.5	**1.5**	1.6	2.2	4.9	1.9	2.2	5.6	3.7	18.4
Chess	**9.6**									16.4	24.5

Table 5 Test set Balanced Error Rate (BER) (%). Values in **bold** are the lowest in a given dataset. Dec Tree = Decision Trees, Dec=Decorate and ADABoost=ADABoost.M1

The results are presented in Tables 4, 5 and 6. Table 4 compares the performance of PISA with the other classification paradigms in terns of Error Rate (ER). From the table it can be seen that PISA performs consistently well, producing the best result with respect to 12 of the 29 data sets; out performing the other association rule classifier (TFPC), and giving comparable results to the decision tree methods. Additionally, while there was some variation between datasets, PISA produced results that were comparable overall to those produced by the ensemble methods. PISA scored an average overall accuracy of 93.60%, higher than that obtained by any of the other methods tested (e.g. Bagging-RDT (89.48%) and RDT (90.24%))[7]. PISA demonstrated consistent performance with respect to both multi-class and two-class datasets. The results also show that PISA performs well with respect to unbalanced domains (e.g. the Car Evaluation, Nursery and Page Blocks) without the need to pre-process the datasets.

Table 5 shows the recorded BER with respect to each of the given datasets. Similar observations can be made with respect to Table 5 as for Table 4,. From the table 5 it can be seen that PISA again gave good results overall, producing the best result in 14 out of the 29 datasets. Both tables 4 and 5 demonstrate that PISA outperforms the centralised approaches (C4,5, RDT and TFPC) clearly indicating the advantage offered by the MANCE approach over the centralised approach.

Table 6 gives the execution times (in milliseconds) for each of the methods. As might be expected PISA is not the fastest method. However, the recorded performance is by no means the worse (for instance Decorate runs more slowly

[7] These accuracies were calculated from Table4.

Dataset	PISA	Ensembles							Dec Trees		TFPC
		Bagging		ADABoost		MultiBoost		Dec			
		C4.5	RDT	C4.5	RDT	C4.5	RDT		C4.5	RDT	
Hepatitis	115	110	**40**	190	70	200	60	610	**40**	60	213
Ionosphere	437	1130	210	1170	20	1210	20	4090	80	**12**	109
HorseColic	17									**4.8**	108
Congress	34	50	20	20	140	130	20	590	30	**15**	154
CylBands	83	110	130	40	20	40	20	1190	40	**17**	936
Breast	31	110	110	140	110	170	170	330	**8.1**	**8**	11
Pima	75	160	90	80	130	80	110	500	20	21	**11**
TicTacToe	71	80	70	250	30	280	10	620	20	**6.1**	61.4
Mushrooms	313	750	380	110	**50**	60	**50**	6400	80	117	630
Adult	3019									**706**	1279
Iris	42	40	50	60	50	50	10	110	10	13	**2**
Waveform	1243	1840	380	4400	830	1650	560	4730	200	**102**	862
Wine	136									**106**	163
Connect4	4710									**3612**	6054
Lympho	15	80	50	90	10	70	10	140	**5**	**5**	29
Car Eval	74	300	110	370	20	20	310	1580	80	24	**17**
Heart	343	250	80	480	20	430	10	620	20	**5**	183
Page Bloc	159	130	430	430	130	280	130	430	120	**55**	60
Nursery	965	1790	720	3130	60	3760	**10**	1449	110	139	204
Dematology	194	160	40	230	20	20	20	480	20	**7**	169
Annealing	750	1090	120	850	**10**	1170	**10**	3340	50	28	689
Zoo	43	40	10	20	10	30	10	110	10	**5**	85
Auto	210	440	70	320	10	350	10	520	20	**5**	43
Glass	180	260	120	340	**10**	430	**10**	1060	20	**10**	43
Ecoli	139	240	150	360	10	340	10	1510	10	**3**	4
Flare	239	30	20	60	40	20	20	140	**10**	27	23
Pen Digits	1345	2300	460	5810	820	2790	800	2300	290	**80**	1606
Led7	78	730	360	260	130	1150	480	3380	110	90	**25**
Chess	2412									334	**226**

Table 6 Test set execution times (milliseconds). Values in bold are the lowest in a given dataset. Dec Tree = Decision Trees, Dec=Decorate and ADABoost=ADABoost.M1

than PISA with respect to the majority of the datasets), nor is the time taken unacceptable. Additionally, PISA seems to run faster than Bagging and ADABoost with some datasets.

With respect to Tables 4, 5 and 6 the empty "cells" indicate where (for a variety of reasons) the associated algorithm failed to produce a result because the method either "timed out" or because resource errors were encountered. It is interesting to note that PISA did not succumb toany of these resource errors. In some cases the empty fields are because WEKA was unable to process the data.

6.2 Using Groups of Agents

This section reports on the evaluation of the use of groups of agents to enhance classification performance. The ability of PISA to operate with groups of agents supporting a single classification was described in Sub-section 6.2. For the evaluation the above reported experiments were run again using group sizes of 2, 4 and 5 agents supporting each possible class. Figure 1 shows the reduction in the error-rates of PISA when using groups of agents compared to the use of single agents (i.e. groups of one agent). From the figure it can be seen that the use of groups of PISA agents tends to improve performance provided that the initial data set in question is sufficiently large to retain a reasonable amount of data for each agent when distributed across the increased number of participating agents. The evidence indicates that the greater the number of individual databases the greater the number of arguments that can be found, the better the exploration of the problem space. Where there is not enough data to allow each agent a reasonably sized database, such as Congressional

Fig. 1 Reduction in Error Rate for PISA, when using groups of 2, 4 and 5 agents, as a percentage of the original error rate (one agent per group (Table4).

Fig. 2 Reduction in Balanced Error Rate for PISA, when using groups of 2, 4 and 5 agents, as a percentage of the original error rate (one agent per group (Table5).

Fig. 3 Percentage increase in the execution times (milliseconds) for the datasets, when using groups of 2, 4 and 5 agents, compared with the execution time when each group comprises a single agent (Table6).

Voting, Breast, Pima and Led7, dividing the data among 4 or 5 agents increases the error rate, as here each participant is allocated a data set whose size is not sufficient for meaningful application of the group argumentation process. These observations also apply when the balanced error-rate is considered as can be seen from Figure 2.

When using participation groups a computational overhead is incurred. For each round of the dialogue, each member of the group attempts to suggest a move. The group leader then has to choose one of these moves to present in the ongoing dialogue. The experiments reported in Sub-section 6.1 provided information about execution time when groups are not used. Figure 3 shows the increase in execution times (in milliseconds) when using groups of 2, 4 and 5 agents compared to the execution time recorded when using PISA with one agent per class. The figure shows that the run time required for classification with groups of 2 agents takes 1.53 (on average) times longer, 2.37 times longer when using 4 agents per group, and 2.89 longer when using 5 agents.

6.3 Ordinal Classification

The above experiments used the standard agreement model. An alternative model is the biased agreement model (Sub-section 4.1). As noted above, this is of relevance with respect to Ordinal classification, when the different classifications can be put into a meaningful order. To test the hypothesis that MABCE coupled with a biased agreement model improves the performance of PISA in the context of ordinal classification, a series of TCV tests, using a number of datasets from Table 2, which have ordered classes, were conducted. PISA was run using the two non-standard agreement models NA-BIA (no attacks) and TC-BIA (differential thresholds). The results were compared with the operation of PISA using the standard agreement model. Additionally, to provide a better comparison, the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) rates for the included datasets and methods were calculated. [23] notes that little attention has been directed at the evaluation of ordinal classification solutions, and that simple measures, such as accuracy, are not sufficient. In [23] a number of evaluation metrics, for ordinal classification, are compared. As a result MSE is suggested as the best metric when

Datasets	**ER**			**BER**		
	PISA	TC-BIA	NA-BIA	PISA	TC-BIA	NA-BIA
Lympo	6.21	4.76	**3.38**	15.95	20.73	**13.94**
Car Eval	4.11	5.00	**4.03**	**9.53**	10.09	10.61
Page Bloc	**2.67**	3.64	3.91	13.43	10.42	**10.06**
Nursery	6.37	6.27	**5.83**	11.79	13.57	**7.88**
Dema	**4.96**	7.95	6.87	8.49	8.74	**7.53**
Zoo	9.90	7.92	**6.86**	13.23	14.67	**12.17**
Ecoli	6.03	5.52	**4.34**	16.81	**6.72**	6.91
Datasets	**MSE**			**MAE**		
	PISA	TC-BIA	NA-BIA	PISA	TC-BIA	NA-BIA
Lympo	0.19	0.05	**0.02**	2.07	1.36	**0.84**
Car Eval	0.86	1.22	**0.71**	1.02	1.32	**1.01**
Page Bloc	**1.250**	5.16	4.76	**0.49**	0.78	0.83
Nursery	7.45	7.07	**6.73**	1.61	1.57	**1.46**
Derma	0.14	0.14	**0.10**	1.46	1.37	**1.24**
Zoo	**0.22**	0.23	0.23	2.26	2.26	**1.96**
Ecoli	0.02	0.02	**0.01**	8.23	7.92	**4.63**

Table 7 The application of PISA with datasets from Table2 with ordered classes.

more (smaller) errors are preferred to penalise large errors; while MAE is a good metric if, overall, fewer errors are preferred with more tolerance for large errors. Table 7 provides a summary of the results of the experiments. From the table it can be seen that by using a biased agreement model the overall classification performance tends to be improved. This is because agents will support (either passively or actively) the classes proposed by other agents if they are close to their own position. Overall the NA-BIA model produces better results than the TC-BIA model.

6.4 Unbalanced Class Problem

As noted in Sub-section 4.2 PISA supports the concept of dynamic coalitions. Dynamic coalitions allow agents to to temporarily cooperate. The hypothesis is that dynamic coalitions between different agents will produce a better performance with respect some data sets, for example unbalanced data sets. It has been observed (e.g.[30]) that class imbalance (i.e a significant differences in class prior probabilities) may produce an important deterioration in the performance achieved by existing learning and classification systems. This situation is often found in real-world data sets that include infrequent but important occurrences. As will be shown here, using the concept of Dynamic Coalitions the operation of PISA can be enhanced with respect to the unbalanced multi-class classification problem.

To test the hypothesis that the use of dynamic coalitions improves the performance of PISA when applied to unbalanced datasets a series of TCV tests,

Datasets	ER			BER			G-Mean		
	PISA	Coal(1)	Coal(2)	PISA	Coal(1)	Coal(2)	PISA	Coal(1)	Coal(2)
Connect4	5.0	4.2	**3.8**	11.9	9.7	**8.7**	87.5	89.9	**91.0**
Lympo	6.2	5.0	**4.0**	15.9	**11.9**	14.7	69.3	82.6	**92.8**
Car Eval	4.1	**3.7**	4.2	9.5	7.2	**4.5**	79.4	88.4	**92.5**
Heart	5.1	**4.9**	**4.9**	8.3	**2.5**	3.2	84.4	87.7	**89.9**
Page Bloc	2.2	1.4	**1.1**	13.4	**7.9**	9.6	68.3	**85.4**	84.0
Derma	4.9	3.9	**3.6**	8.5	4.9	**4.5**	75.8	84.3	**90.1**
Annealing	9.6	4.2	**4.0**	16.1	7.7	**4.2**	63.6	86.2	**91.5**
hline Zoo	9.9	8.0	**7.0**	13.2	8.3	**3.9**	67.2	85.4	**85.5**
Auto	12.0	6.4	**5.8**	12.3	**6.5**	6.7	79.7	87.9	**90.9**
Glass	14.7	12.0	**5.7**	16.1	7.45	**5.8**	80.1	**93.6**	93.2
Ecoli	6.0	**5.2**	5.6	16.2	10.9	**3.9**	74.2	87.3	**96.0**
Flare	**6.1**	7.1	6.9	17.2	5.6	**5.2**	77.4	91.2	**95.8**
Chess	9.1	8.5	**6.3**	9.6	5.9	**5.3**	76.7	91.3	**92.2**

Table 8 The application of PISA with imbalanced *multi-class* datasets from Table2.

Datasets	Time			Datasets	Time		
	PISA	DCT1	DCT2		PISA	Coal(1)	Coal(2)
Connect4	**4710**	5376	5818	Lympo	**15**	65	55
Car Eval	**74**	163	158	Heart	**343**	531	612
Page Bloc	**159**	207	222	Derma	**194**	199	207
Annealing	**750**	980	881	Zoo	**43**	93	85
Auto	**210**	336	293	Glass	**180**	198	211
Ecoli	**139**	186	181	Flare	**2393**	2291	6267
Chess	**2412**	3305	3393				

Table 9 The execution time of the application of PISA with imbalanced *multi-class* datasets from Table2.

using a number of datasets from Table2 which feature unbalanced class distributions, were undertaken. For the evaluation we assumed that the agents representing rare classes were in coalition at the commencement of the dialogues. The results were compared against the use of PISA without any coalition strategy. Four measures were used for the comparison: (i) error rate, (ii) balanced error rate, (iii) run-time and (iv) geometric mean (g-mean)[8]. This last measure was used to quantify the classifier performance in [4]. Tables 8 and Table 9 present the results obtained. The column labels "DCT1" and "DCT2" refer to the two coalition termination mechanisms presented in Sub-section 4.2. From the tables it can be seen that both coalition techniques boost the performance of PISA when applied to unbalanced class datasets, with very little additional cost in time because the coalitions are dismantled when no longer required.

[8] The geometric mean is defined as $g - mean = (\prod_{i=1}^{C} p_{ii})^{\frac{1}{C}}$ where p_{ii} is the class accuracy of class i, and C is the number of classes in the dataset. This measure represents the trade-off between the accuracies of the different classes: in order to achieve a high G-mean value a large portion of the minor samples must be classified correctly, even at the cost of misclassifying some major class examples).

Noise	PISA	RDT	C4.5	TFPC
0	98.47	94.44	68.19	92.56
2	97.64	90.56	67.75	91.81
5	97.36	93.47	62.92	89.72
10	96.53	92.92	60.97	86.81
20	95.69	91.94	60.56	80.83
40	94.44	90.31	56.35	69.86
50	93.75	88.36	61.81	45.83

Table 10 Accuracyb(%) v. Noise.

6.5 Analysing the Effect of Noise

It was conjectured that the reason why MABCE, and its realisation in PISA, operates so effectively compared to more standard approaches to classification was that by distributing the data across a set of agents or groups of agents the effect of noise and the presence of anomalies can be minimised. To analyse the effect of noise the housing benefits data set, described for the worked example given in Section 5, was used. A total of 2400 records were generated, 30% of these (720 records) were set aside as the test set. The remaining 70% were and equally distributed over four agents (one per classification), so that each agent had a data set comprising 420 records. The model used to introduce noise was the same as that reported in [36]; for an N% noise level in a dataset of D instances, $(N * D)$ instances were randomly selected and the class label changed to some other randomly selected value (with equal probability) from the set of available classes. The noise levels used in this study were: 2%, 5%, 10%, 20%, 40% and 50% . The noise was, of course, introduced in the training sets only and not to the test sets.

Table10 shows the affect of adding noise to the housing benefit dataset on the accuracy of each classifier. From the table it can be seen that the best overall classifier, in the presence of large amounts of noise, was PISA with an accuracy level starting with 98.47% for clean (no noise) data and dropping to 93.75% when a 50% noise level is introduced. The experiment clearly indicates that PISA, and by extension MABCE in general, copes extremely well with noisy data compared to the other classifiers tested. Note that the other data mining technique is particularly susceptible to large amounts of noise.

7 Discussion and Conclusions

A mechanism to achieve Multiagent Argumentation-Based Classification from Experience(MABCE) has been described. The mechanism has been realised in the PISA system. PISA allows any number of software agents to engage in argumentation dialogues concerning the classification of a case whereby the participating agents *mine* their arguments directly from local background datasets representing their experience. The dialogue progresses in a round-by-round manner. During each round agents can elect to propose an argu-

ment advocating their own position or attack another agent's position. The arguments are mined and expressed in the form of Association Rules (ARs), which are viewed as generalisations of the individual agent's experience. The exemplar application is benefits adjudication where geographically dispersed benefits offices can "argue" about particular cases by pooling their experience without specifically sharing individual data instances.

In the context of the field of argumentation the proposed approach has general applicability as it does not require the generation of specialised knowledge, rules or case bases (as in the case with other argumentation systems) or reference to domain experts. In the context of classification, PISA also provides an explanation, with reference to the argumentation tree, of how the final classification was arrived (see the worked example presented in Section 5. Also regarding classification, and with reference to the evaluation described in Section 6, the MABCE approach can outperform more standard approaches, and can produce even better performance if there is enough data available to allow agents to work in groups. Further advantages can be obtained if the biased agreement model is used with respect to ordinal classification problems and when dynamic coalitions are used with respect to unbalanced data classification problems. The reason for this is that by distributing the data across a set of agents or groups of agents the effect of noise and the presence of anomalies can be minimised as demonstrated by the experiments described in Sub-section 6.5. These advantages are all clear reasons as to why the proposed MABCE may be desirable rather than centralised approaches.

Overall the advocated MABCE paradigm, and its realisation in PISA, presents an approach to classification that is both novel and effective, and that has real application to common problems, particularly situations where the data may not be entirely reliable.

References

1. Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. Proc. ACM SIGMOD Conf. on Management of Data (SIGMOD'93). ACM Press, pp207-216 .
2. Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. Proc. 20th International Conference on Very Large Data Bases (VLDB94), pp487-499.
3. Albashiri, A., Coenen, F. and Leng, P. (2009). EMADS: An Extendible Multi-Agent Data Miner. Knowledge Based Systems, vol.22(7), pp523-528.
4. Alejo, R., Garcia, V., Sotoca, J., Mollineda, R., Sanchez, J. (2007). Improving the Performance of the RBF Neural Networks with Imbalanced Samples. Proc. 9th Intl. Conf. on Artl. Neural Networks. Springer, pp162-169.
5. Bench-Capon, T., and Dunne, P. (2007) Argumentation in Artificial Intelligence Artificial Intelligence. vol 171(10-15), pp619-41.
6. Bai, Q. and Zhang, M. (2005). Dynamic Team Forming in Self-interested Multi-agent Systems." In *AI 2005: Advances in Artificial Intelligence*, Spring. pp. 674–683 (2005).
7. Bauer, E. and Kohavi, R, (1999). An empirical comparison of voting classification algorithms: Bagging, Boosting and variants. J. Machine Learning, Vol. 36, pp105-139.
8. Bench-Capon, T. J. M. (1991) Knowledge Based Systems Applied To Law: A Framework for Discussion. In Knowledge Based Systems and Legal Applications. Academic Press. pp329-342.

9. Bench-Capon, T. J. M. (1993) Neural Nets and Open Texture. In Proc. 4th Int. Conf. on AI and Law (ICAIL'94). ACM Press: Amsterdam, (1993). pp292-297.
10. Blake, C.L., and Merz.,C.J. (1998). UCI Repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.
11. Brieman, L. (1996). Bagging predictors. In J. Machine Learning, **24**, Springer, pp123-140.
12. Carabelea, C. (2001). Adaptive Agents in Argumentation-Based Negotiation. Proc. ECCAI-ACAI/EASSS'01, AEMAS'01, HoloMAS'01 on Multi-Agent-Systems and Applications II-Selected Revised Papers. Springer, London. pp180-187.
13. Caragea, D., Silvescu, A. and Honavar, V. (2000) Agents that learn from distributed dynamic data sources. In Stone, P. and Sen, S. (Eds.): Proc. of the Workshop on Learning Agents (Agents 2000/ECML 2000), Barcelona, Spain.
14. Caragea, D., Silvescu, A. and Honavar, V. (2003) Decision tree induction from distributed, heterogeneous, autonomous data sources. In Proc. Conf. on Intelligent Systems Design and Applications (ISDA 03), 2003.
15. Coenen, F. (2007). The LUCS-KDD Decision Tree Classifier Software, `http://www.csc.liv.ac.uk/~frans/KDD/Software/DecisionTrees/decisionTree.html`, Dept. of Computer Science, The University of Liverpool, UK.
16. Coenen, F. and Leng, P.H. (2005). Obtaining Best Parameter Values for Accurate Classification. In Proc. ICDM'05, IEEE. pp597-600.
17. Coenen,F., Leng, P.H., and Ahmed,S., (2004). Data structure for association rule mining: T-trees and p-trees. In J. IEEE Trans. Knowl. Data Eng., vol.16(6). pp774-778.
18. Committee of Public Accounts (2003) Getting it right: Improving Decision-Making and Appeals in Social Security Benefits. Committee of Public Accounts. London: TSO, 2104 (House of Commons papers, session 2003/04; HC406).
19. 11. S. Datta, K. Bhaduri, C. Giannella, R. Wolff, and H. Kargupta (2006). Distributed data mining in peer-to-peer networks. In J. Internet Computing, IEEE, vol.10(4). pp18-26.
20. Dietterich, T. (2000). Ensemble methods in machine learning. In: Lecture Notes in Computer Science, vol.1857, Springer, pp1-15.
21. Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. J. Mach. Learn. vol. 40, Springer. pp139-157.
22. Freund, Y. and Schapire,R, (1996). Experiments with a new boosting algorithm. In Proc. ICML'96. pp148-156.
23. Gaudette, L., and Japkowicz, N. (2009). Evaluation Methods for Ordinal Classification. In Yong, G., and Japkowicz, N. (editors), Advances in Artificial Intelligence, LCNS, vol.5549, Springer, pp207-210.
.
24. Ghosh, J., Strehl, A. and Merugu, S. (2002) 'A consensus framework for integrating distributed clusterings under limited knowledge sharing', Proceedings of the NSF Workshop on Next Generation Data Mining, pp.99-108.
25. Grecu, D.L. and Becker, L.A. (1998) 'Coactive learning for distributed data mining', Proceedings of KDD-98, New York, NY, August, pp.209-213.
26. Gorodetsky, V., Karsaeyv, O. and Samoilov, V. (2003). Multi-agent technology for distributed data mining and classification. Proc. IAT'03, IEEE/WIC, pp438-441.
27. Groothius, M. and Svensson, J. (2000). Expert System Support and Juridical Quality. Proc. Jurix 2000, 110. IOS Press: Amsterdam.
28. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten,I. H (2009). The WEKA Data Mining Software: An Update. In SIGKDD Explorations, Vol. 11(1), pp10-18.
29. Horling, B. and Lesser,V. (2005) A Survey of Multi-Agent Organizational Paradigms. Knowledge Engineering Review, Vol. 19(4) Cambridge University Press. pp281-316.
30. Japkowicz, N. and Stephen, S. (2002). The Class Imbalance Problem: A systematic study. J. Intelligent Data Analysis, vol.6(5), pp429-449.
31. Klusch, M. and Gerber, A. (2001). Dynamic coalition formation among rational agents. Intelligent Systems, Vol. 17(3), IEEE. pp42-47.
32. Liu, B., Hsu, W. and Ma, Y. (1998). Integrating Classification and Association Rule Mining. Proc. KDD98, American Association for Artificial Intelligence, pp80-86.

33. Melville, P., and Mooney,R., (2003). Constructing Diverse Classifier Ensembles Using Artificial Training Examples. Proc. IJCAI'03, pp505-510.
34. Modi, P., and Shen, W. (2011). Collaborative multiagent learning for classification tasks. Proc. AAMAS'01. ACM press. pp37-38.
35. Modi, P., and Kim, P. (2005) Classification of Examples by Multiple Agents with Private Features. In Proc. IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'05). pp223-229.
36. Mozina, M., Zabkar, J., Bench-Capon T. and Bratko, I. (2005). Argument based machine learning applied to law, Artificial Intelligence, Vol. 13(1), pp53-73.
37. Mulder, W. Meijer, G. R. and Adriaans, P. W. (2008). Collaborative learning agents supporting service network management. In Proc. SOCASE'08 the 2008 AAMAS international conference on Service-oriented computing: agents, semantics, and engineering Springer-Verlag Berlin, Heidelberg.
38. National Audit Office (2006). International benchmark of fraud and error in social security systems . Report by the Controller and Auditor General, HC 1387 Session 2005-2006, 20 July 2006
39. National Bureau of Economic Research Errors in the Social Security Disability Award Process `http://www.nber.org/aginghealth/winter04/w10219.html`.
40. Oliva, E., McBurney, P. and Omicini, A. (2008). Co-argumentation Artifact for Agent Societies. Proc. 6th Int. Workshop on ARGumentation in Multi-AgentSystems (ArgMAS'07), Springer, LNCS 4946, pp207-224.
41. Olmeda, I., Fernandez, E. (1997). Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction. J. Computational Economics, vol.10, pp317-335.
42. Ontanon, S., and Plaza, E. (2007). An Argumentation-Based Framework for Deliberation in Multi-agent Systems. Proc. ArgMAS'07. pp178-196 (2007).
43. Opitz, D., and Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. In J. Artif. Intell. Research, vol.11, pp169-198.
44. Peng, S., Mukhopadhyay, S., Raje, R., Palakal, M., and Mostafa, J. (2001). A comparison between single-agent and multi-agent classification of documents. Proc. 15th Int. Parallel and Distributed Processing Symposium, pp935-944.
45. Prodromides, A., Chan, P., and Stolfo, S. (2000). Meta-learning in distributed data mining systems: issues and approaches. In Advances in Distributed and Parallel Knowledge Discovery, AAAI Press/The MIT Press, pp81-114.
46. Tambe, M., and Jung, H. (1999). The Benefits of Arguing in a Team. AI Magazine, Vol. 20(4), American association for Artificial Intelligence, pp85-92.
47. Tozicka, J. Rovatsos, M. and Pechoucek, M. (2008) MALEF: Framework for distributed machine learning and data mining Int. J. Intelligent Information and Database Systems, vol. 2 (1). pp6-24.
48. Wardeh, M., Bench-Capon, T. and Coenen, F.P. (2009). An Arguing From Experience Approach to Classifying Noisy Data. Proc. 11th Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK'09), Springer LNCS 5691, ISSN 0302-9743, pp354-365.
49. Wardeh, M., Coenen, F. and Bench-Capon, T. (2010). Arguing in Groups. Proc. Computational Models of Argument (COMMA'10,), IOS Press, pp475-486.
50. Wardeh, M., Bench-Capon, T. and Coenen, F. (2011). Arguing from Experience Using Multiple Groups of Agents. Argumentation and Computation, Vol 2, No 1, pp51-76.
51. Webb, G., (2000). MultiBoosting: A Technique for Combining Boosting and Wagging. J. Machine Learning, vol. 40(2), pp159-196.
52. Weiss, G. and Dillenbourg, P. (1999) What is multi in multi-agent learning?. In Dillenbourg, P.(Ed.): Collaborative-learning: Cognitive and Computational Approaches, Elsevier, Oxford, pp6480.
53. Yu, L., Wang, S.Y., Lai, K.K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach". J. Expert Systems with Applications, vol. 34(2). pp. 1434 – 1444.