Electrical Power and Energy Systems 104 (2019) 489–501

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

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

Context aware Q-Learning-based model for decision support in the negotiation of energy contracts

J. Rodriguez-Fernandez[a], T. Pinto[b,*], F. Silva[a], I. Praça[a], Z. Vale[c], J.M. Corchado[b,d]

[a] GECAD Research Group, Polytechnic of Porto (ISEP/IPP), Porto, Portugal
[b] BISITE Research Group, University of Salamanca, Salamanca, Spain
[c] Polytechnic of Porto (IPP), Porto, Portugal
[d] Osaka Institute of Technology, Osaka, Japan

ARTICLE INFO

Keywords:
Automated negotiation
Bilateral contracts
Context awareness
Decision support
Electricity markets
Reinforcement learning algorithm

ABSTRACT

Automated negotiation plays a crucial role in the decision support for bilateral energy transactions. In fact, an adequate analysis of past actions of opposing negotiators can improve the decision-making process of market players, allowing them to choose the most appropriate parties to negotiate with in order to increase their outcomes. This paper proposes a new model to estimate the expected prices that can be achieved in bilateral contracts under a specific context, enabling adequate risk management in the negotiation process. The proposed approach is based on an adaptation of the Q-Learning reinforcement learning algorithm to choose the best scenario (set of forecast contract prices) from a set of possible scenarios that are determined using several forecasting and estimation methods. The learning process assesses the probability of occurrence of each scenario, by comparing each expected scenario with the real scenario. The final chosen scenario is the one that presents the higher expected utility value. Besides, the learning method can determine which is the best scenario for each context, since the behaviour of players can change according to the negotiation environment. Consequently, these conditions influence the final contract price of negotiations. This approach allows the supported player to be prepared for the negotiation scenario that is the most probable to represent a reliable approximation of the actual negotiation environment.

1. Introduction

The Electricity Markets (EM) restructuring placed several challenges to governments and to the companies that are involved in generation, transmission, and distribution of electrical energy. The privatization of previously state owned companies, the deregulation of privately owned systems, and the internationalization of companies, are some examples of the transformations that have been applied [1].

Environmental concerns related to the use of fossil fuels have led to an increase in renewable energy generation sources. The considerable increase of distributed generation units makes EM more competitive, and consequently encourages a decrease in electricity prices [2,3]. However, some recurrent problems that are being addressed all over the world must be considered, such as the dispatch ability, limitations in the power system network, and the integration and large participation of small producers in the EM, among others [3]. Despite these problems, some global solutions are being adopted, some examples are the case of evolution of European EM. The majority of European countries have joined together into common market operators, resulting in joint regional EM composed of several countries, which supports transactions of huge amounts of electrical energy and allows the efficient use of renewable based generation in places where it exceeds the local needs [4].

Nowadays several market models exist, with a set of complex rules and particular regulations, creating the need to anticipate market behaviour. Some implemented market types have the clearing mechanism based on the optimization of offers, such as most electricity markets in the U.S. [5] and other based on symmetric or asymmetric bids, as is the case of most European countries [4]. However, electricity trade worldwide is also supported by means of bilateral contracts negotiation [6], which are the scope of this study.

The common behaviour of market players in contracts negotiation is mainly based on the definition of prices and quantities in energy transactions with each competitor. Hence, relevant information concerning competitors' history of previous negotiations can be used to improve the decision-making process, considering the characteristics of

* Corresponding author.
E-mail addresses: tpinto@usal.es (T. Pinto), fspsa@isep.ipp.pt (F. Silva), icp@isep.ipp.pt (I. Praça), zav@isep.ipp.pt (Z. Vale), corchado@usal.es (J.M. Corchado).

https://doi.org/10.1016/j.ijepes.2018.06.050
Received 18 October 2017; Received in revised form 14 May 2018; Accepted 27 June 2018
Available online 25 July 2018
0142-0615/

the moment of negotiation, namely to improve the forecasting of possible contract prices before the negotiation process [6]. It is essential to consider the concept of context awareness, since it influences the prices and quantities of energy to be negotiated. One example is the new ways of participating in EM, such as renewable resources, which has hardly influenced the players' participation in the negotiation process, due to the dependency of environment factors, such as wind or solar intensity, that influence the final price of electricity. Other examples of contexts are the different types of days such as business days, weekends, holidays, or other days with special situations that affects energy consumption. A unique review of context analysis mechanism of EM negotiating players is presented in [7], which proposed a methodology of analysing the past negotiation context to distinguish days and periods with similar characteristics.

Those introduced aspects have significant implications in the increase of the complexity and unpredictability in EM. Hence, the constant change of the EM environment requires the need to understand market's mechanism and how the interaction between the players affects the markets. It has contributed to the increased use of simulation and decision support tools, in order to achieve the best possible results of each market context for each participating entity [2]. Several modelling tools based in Multi-agent software for the study of electricity markets have emerged [8]. Some relevant examples are Electricity Market Complex Adaptive System (EMCAS) [9], Agent-based Modelling of Electricity Systems (AMES) [10], Genoa Artificial Power Exchange (GAPEX) [11], and Multi-Agent System for Competitive Electricity Markets (MASCEM) [12].

Current tools are directed to the study of market mechanisms and interactions among participants, but are not suitable for supporting the decision of the negotiating players in obtaining higher profits in energy transactions. A new multi-agent adaptive learning system – AiD-EM (Adaptive Decision Support for Electricity Markets Negotiations) – has been integrated with MASCEM market simulator with the purpose of providing effective decision support to electricity markets' negotiating players [8]. This decision support system is modelled for different market negotiation types, namely the participation in auction based markets and the automated negotiation of bilateral contracts. The latter negotiation type is addressed in this paper, namely through the DECON system (Decision Support for Energy Contracts Negotiation). DECON implements several methodologies to analyse competitor players' negotiation profiles, enabling the adjustment of the adopted negotiation strategies and tactics in each step of automated negotiation [13]. Techniques such as adaptive learning and game theory [14], which explores the study of algorithms that can learn from and make predictions or decisions on data, allows the assessment of each current negotiation context[1] and to dynamically learn over time [15,16]. Such concepts should be adopted in order to overcome the current gap in the literature related to the lack of analysis of past information about opponents, and the inadequate exploration of the pre-negotiation stage, as identified in the review on automated negotiation presented in [13].

In the literature, is possible to find some tools that support bilateral contracts negotiation such as EMCAS [17], General Environment for Negotiation with Intelligent Multi-Purpose Usage Simulation (GENIUS) [18] and the Multi-Agent Negotiation and Risk Management in Electricity Markets (MAN-REM) [19]. EMCAS is a multi-agent simulator that is able to simulate electricity market bilateral contracts, established between a demand agent and a generation company agent [17]. The generation agents decides the price of the demand agents' proposals that may or may not be accepted by the proposers. GENIUS is a multi-agent simulator that facilitates and evaluates the strategies of automated negotiators [18]. The tool supports domain-independent bilateral negotiations and considers three negotiation phases: Preparation (negotiation protocol and domain), Negotiation, and Post-negotiation (negotiation analysis). MAN-REM simulates the bilateral contracts negotiation through the combination of small multi-agent simulators. The tool models the buyer, seller, trader (distribution), and market operator (validation) agents. Three negotiation phases are considered: Pre-Negotiation (contract's preferences and response to counter-offers definition), Actual Negotiation, and Post-Negotiation (final agreement) [19]. The analysed tools presents a lack of exploration of the pre-negotiation phase, only focusing the actual negotiation. The GENIUS simulator has the most complete exploration of the pre-negotiation phase, but also lacks opponents analysis. In summary, although some advances have been made regarding the pre-negotiation phase, several problems are yet far from being adequately addressed, such as the definition of models to choose the most appropriate parties to negotiate with, and how relevant information regarding competitors' history of previous negotiations can be used to improve the decision making process, namely regarding the choice of the most suitable negotiation strategies and tactics. The absence of automated negotiation models directed to negotiations between electricity market players also brings out several relevant challenges that must be addressed promptly in order to provide market players with adequate decision support solutions to enable market players to adapt to the constantly changing electricity market environment, and learn how to take the most advantages out of market participation.

In order to overcome these limitations, this paper presents a new learning model which has the aim of supporting the decisions of players in the pre-negotiation of bilateral contracts, achieving an advantageous position that allows to identify the ideal negotiators to trade with, enhancing the outcomes of the negotiation process. This method is based on the application of reinforcement learning algorithm (RLA), namely an adaptation of the Q-Learning algorithm, to learn which is the forecasting method that is able to provide a potential contract price that is closer to reality. The proposed algorithm determines the best method depending on the negotiation context. The forecast scenarios are determined using several different methods, such as data mining techniques [20], artificial neural networks (ANN) [21], support vector machines (SVM) [8], fuzzy logic [22], among other methods [8], where each methods suggests an expected price for each amount of energy. However, no method presents a better performance than all others in every situation, only in particular cases and contexts [8]. Thus, these contract prices forecasting are submitted to some error degree. Because of that, the quality of definition of the best forecast method is essential for supporting the decision process. The proposed model is implemented and integrated in the DECON decision support system to enable its experimentation and validation.

After this introductory section, Section 2 presents a discussion on the need of decision support tools for bilateral negotiation in EM, and an overview of the developed methodology for DECON. Section 3 provides the proposed learning method to estimate bilateral contract prices using a Q-Learning based approach. Section 4 presents a case study that shows the experimental results of the proposed methodology, using the alternative negotiation scenarios furnished by DECON and historic bilateral contracts data. Finally, Section 5 presents the most relevant conclusions of this work.

2. Bilateral contracts negotiation

Bilateral contract is a EM rigid model that enables players to directly negotiate with each other, establishing a fixed price for a quantity of energy for an agreed period. When a player wishes to participate in the bilateral market, it contacts potential players offering his power and price proposal. The target players analyse the proposal and, if interested, they can accept it or try to renegotiate it. Before reaching an agreement, the supplier must be sure that it is feasible to deliver energy in the buyer's location, and for that the system operator's feedback is

[1] Negotiation context refers to characteristics or circumstances under which the negotiation process occurs, *e.g.* if it is a business day or weekend, the season, the current global consumption, the current amount of generation, etc.

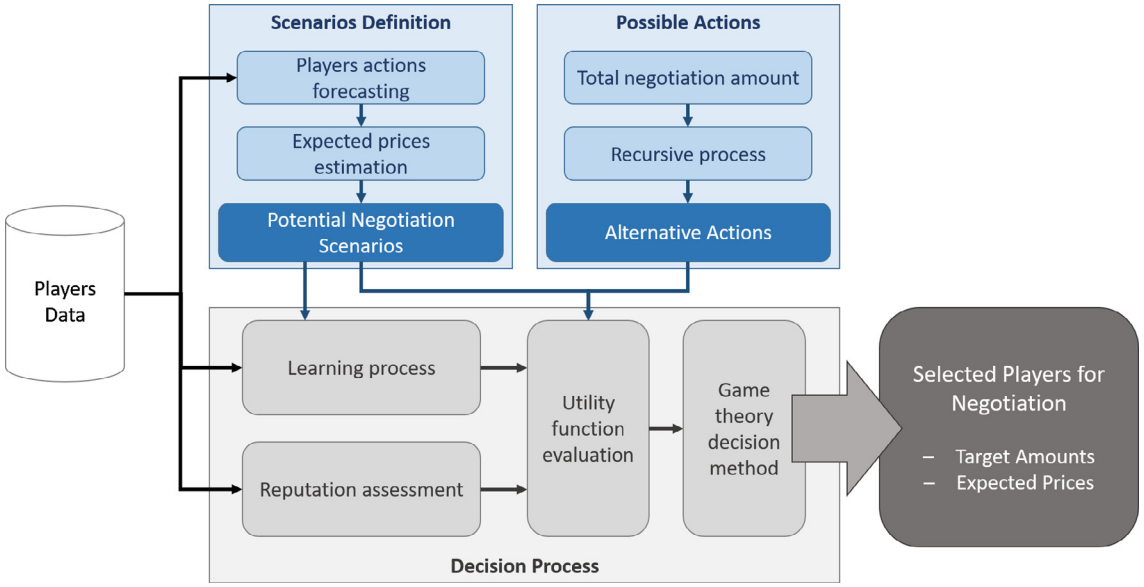

Fig. 1. Pre-negotiation decision support process [14].

needed.

2.1. Bilateral negotiation markets

Electricity markets are usually composed of several market types [23], based on several different models such as: day-ahead spot; intra-day, both usually auction based; and bilateral contracts. In the scope of electricity markets, bilateral contracts are long-term contracts established between two entities, buyer and seller, for energy transaction, without the involvement of a third entity. The transaction is usually carried out several weeks or months after the contract is made [24] and usually has the following specifications: start and end dates and times; Price per hour (Eur/MW h) and amount of energy (MW), variable throughout the contract and, finally, a range of hours relative to the delivery of the contract. Players can use customized long-term contracts, trading 'over the counter' and electronic trading to conduct bilateral transactions [25]. In MIBEL [26], there are four types of bilateral contracts: the first type are Forward Contracts, that consist in energy exchange between a buyer and a seller for a future date, for the price negotiated at that moment; the second type are Future Contracts, which are similar to Forward Contracts except that they are managed by a third party responsible for ensuring compliance with the agreement; the third type are Option Contracts, that are similar to the Forward and Future contracts with the difference that the two entities only guarantee a buy/sell option; the last one are Contracts for Difference, that allows concerned entities to protect themselves from the energy price change between the agreement establishment date and the agreed exchange date. With the exception of Contracts for Difference, this type of negotiation allows players to control the price at which they will transact energy, in contrast to what happens in spot markets, due to the proposals' instability. In establishing a Forward or Future contract, players are committing themselves to transact energy for a given price at a future time, with the risk of making a transaction at a lower price than the expected and lose competitive power. Option Contracts or Contracts for Difference can avoid this risk. The first allows the player to choose not to go through with the exchange while the second ensures that the transaction is carried out at the market price. However, the first option also has the risk of not guaranteeing whether or not the other party will exercise their option to exchange and the second option does not allow better prices than the market. This way, it is possible to understand the risk associated with the negotiation of bilateral contracts and the need that players have of tools that help them reduce this risk and even optimize their profits.

2.2. Decision support for bilateral contracts negotiation

Bilateral negotiation is a recurrent theme in the literature in several fields, including in AI (artificial intelligence) [27]). A relevant review of automated negotiation methodologies for computational agents with focus in AI has been presented in [13]. In the scope of this work, automated negotiation plays an important role in the decision support for energy transactions, since the supported player can negotiate simultaneously with several competitor players, where each negotiation will involve mainly an iterative exchange of proposals and counter-proposals, regarding the prices for the energy. This complex scenario implies high effort and time for players, and consequently exists the risk or the possibility to breach an agreement (total or partial) by some party. Thus, decision support solutions using automated negotiation models are suitable to analyse competitors' past actions and modelling competitors' profiles, and study the best possible negotiation strategies and tactics to be used throughout the negotiation process, considering different competitor players and negotiation context, to obtain the best possible outcomes for supported player. However, automated negotiation methodologies in the specific scope of electricity market players' negotiation is completely absent from literature.

A relevant multi-agent system approaches the problem of lack of the decision support for automated negotiation for computational agents: Decision Support for Energy Contracts Negotiation (DECON) that has been presented in [14] According to this decision support methodology, it considers two phases for automated negotiation, the pre-negotiation step, and the actual negotiation process. The decision support for the pre-negotiation step is the focus of this paper. To provide as much benefit as possible for the supported player in the undertaken negotiations, this decision support stage aims to identify the most appropriate competitor(s) that should be approached. Additionally, expected prices and energy amounts of each targeted competitor are estimated to increase the decision support for current negotiation process. From Fig. 1, it is possible to observe the framework of DECON.

The DECON system is composed by three main parts which are detailed in [14]. The proposed methodology of this paper addresses the learning process which is based on RLA to perceive which, from all the alternative potential negotiation scenarios, is the most likely to occur under the current context. The defined scenarios are based on the opponents' historic data analysis, using forecast methods such as ANN and

SVM, among others. This way, it guarantees that the suggested action, for the supported player, is the best potential action under the scenario s and context c with the larger probability of occurrence, as detailed in Section 3. Finally, depending on the risk that supported player is willing to take and using a reputation profile model, regarding the subject competitor players, several decision methods are included with the application of game theory [28].

3. Proposed methodology

The behaviour of electricity markets players is normally based on strategies whose purposes are to define energy price and transacted amounts. Consequently, it is essential that the negotiators could be able to predict the expected prices, resulting from potential negotiations. Using historical data obtained from previous agreements, several forecasting methodologies are applied in DECON system to prognosticate the expected established contract price for each player, for different transacted amounts. As previously mentioned, no method presents a better performance than all others in every circumstance, only in particular cases and contexts [8]. Therefore, these forecasting methods are subject to some error degree. Hence, it is crucial to determine the best forecast method for supporting the decision process. These issues motivated the development of the present work that proposes to undertake a learning process to recognize the forecast contract price, which present the higher probability of occurrence in each current context.

The learning process allows an agent to acquire a skill or knowledge that is not available. In fact, an analysis and appropriate learning can improve the results of the participation of stakeholders. The proposed method uses a learning process based on the assessment of likelihood of occurrence of each alternative negotiation scenario.[2] Thus, this approach allows the supported player to be prepared for the negotiation scenario that is the most likely to occur and perform the action that generates better results. Besides, the contextualization of the learning process is enabled, obtaining expected negotiation scenarios that most reflect each current circumstances and context.

3.1. Context awareness

Context is present everywhere, and unquestionably influences the way information is processed in every situation [29–31]. While context is critical to information processing in all kinds of situations, it is almost fully absent from the modern information technology infrastructure. There is some work done in providing computer systems with context awareness [32], namely in multi-agent simulation [33]. However, the concept of context awareness is very far from being widely used in the computer system's area. The fact that this is an important issue to consider and its lack of consideration in decision support systems made it essential to include context analysis in this work. The analysis and definition of different contexts of negotiation are performed to support an adequate acting of negotiating players, adapting their actions to best suit the context they are encountering at each moment.

In the considered model, the context analysis consists in analysing the past observations, regarding important contextual aspects that affect the negotiation process, *e.g.* day, hour, season, electricity market price, amount of transacted power in the market, the wind intensity verified in that period of the day (this is important because it affects the production of wind plants, and therefore the total negotiated amount of power), the solar intensity, the type of the day (whether it is a working day or weekend; if it is a holiday, or a special situation day, *e.g.* a day of an important event, such as an important game in a certain sport, which affects the energy consumption in that day, both because of the consumption in the stadium, and for increasing the number of people with the TV on to watch it).

The analysis of these data is performed by means of a clustering process, which groups observations according to their similarity. Each cluster (group) corresponds to a specific and distinct context [7]. Once the contexts are defined, when a new observation occurs (*e.g.* a new established energy contract), the contextual information is used to classify this new observation into the most similar group (context) from those defined in the first stage. In this way it is possible to determine the context to which the new event most relates to.

3.1.1. Clustering

The clustering mechanism analyses the characteristics of each period throughout the days, and assigns each period of each day to the cluster that presents the most similar characteristics. The clustering is performed using the K-Means clustering algorithm [34]. The K-Means clustering methodology considers a set of observations $(x_1, x_2, ..., x_n)$, where each observation is a d-dimensional real vector, and n is the number of considered observations. The clustering process aims at partitioning the n observations into $k(\leqslant n)$ clusters $C = \{C_1, C_2, ..., C_k\}$ so that the Within-Cluster Sum of Squares (WCSS) is minimized (1).

$$\min \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{1}$$

where μ_i is the mean of points in C_i, *i.e.* the cluster *centroid*.

The dimension of the vector that characterizes each observation x_p, $p \in \{1, ..., n\}$ is equal to the sum of five vectors of equal size, each containing the observable information regarding each of the five characteristics presented before in this section (in the bullet points); *i.e.* $x_p = \{mp_p, pw_p, ws_p, si_p, sp_p\}$, where mp_p represents the set of market price values associated to each hour of each day that compose each observation, pw_p represents the amounts of transacted power, ws_p represents the wind speeds that have been verified, si_p, the solar intensities, and sp_p is the indication of special cases, where the value of 0 indicates a business day, 1 signifies a weekend day, 2 represents a holiday, and 3 a special situation day, *e.g.* days in which relevant events occur during certain hours of a day. The length of each of the five vectors that compose x_p are dependent on the amount of data that is considered as part of each observation.

With the objective of minimizing Eq. (1), the clustering process executes an iterative process between two steps: (i) the assignment step, where each observation x_p is assigned to the cluster $C^{(t)}$ whose mean value yields the minimum WCSS in iteration t, as presented in (2); and (ii) the update step, where the new means of each cluster are calculated, considering the newly assigned observations, determining the new centroid μ_i of each cluster, as in (3).

$$C_i^{(t)} = \{x_p : \|x_p - \mu_i^{(t)}\|^2 \leqslant \|x_p - \mu_j^{(t)}\|^2 \ \forall\, j, \quad 1 \leqslant j \leqslant k\} \tag{2}$$

$$\mu_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{xj \in C_i^{(t)}} x_j \tag{3}$$

The execution of the algorithm stops when the convergence process is completed, *i.e.* when the assignments of observations to different clusters no longer change. By minimizing the WCSS objective, in Eq. (1), the K-Means clustering methodology assigns observations to the nearest cluster by distance. This means that each subject will be grouped in the same cluster as the ones that are more similar.

3.1.2. Classification

The proposed classification model intends to enable identifying the context in which new observations or events should be associated to. The ANN used in this proposed method is a feedforward neural network. Feedforward networks consist of a series of layers. The first layer has a connection from the network input. Each subsequent layer has a connection from the previous layer. The final layer produces the

[2] Negotiation scenario refers to the negotiation setting that the envisaged player will encounter when facing negotiations, *e.g.* the expected available counterparts, their target prices and trading volume.

network's output.

The considered ANN is a Multi-Layer Perceptron (MLP) feedforward neural network, which considers the contextual information about the event in time. The output is the corresponding context. A study that supports this MLP topology is shown in [21].

The training algorithm is backpropagation using the gradient descent method [35]. The squared error function E for the single output neuron is defined as in (4).

$$E = \frac{1}{2}(t-y)^2 \tag{4}$$

where t is the target output for a training sample, and y is the actual output of the output neuron. For each neuron j, its output o_j is defined by feedforward calculation, as in (5).

$$o_j = f\left(\sum_{k=1}^{n} w_{kj}x_k\right) \tag{5}$$

where n is the number of input units to neuron j, and w_{kj} is the weight between neurons k and j. The logistic function is used as activation function f, as in (6).

$$f(z) = \frac{1}{1+e^{-z}} \tag{6}$$

This classification process enables identifying the context that a new observation (contract) belongs to; hence enabling the introduction of the context-aware dimension in the learning process, as proposed in this paper.

3.2. Context aware Q-learning model

The context aware bilateral contract price estimation approach is based on the application of the Q-Learning reinforcement learning algorithm [36], due to dynamic environment such as bilateral negotiations, where an agent learns through attempt and error. Q-Learning is a very popular reinforcement learning method. It is an algorithm that allows the autonomous establishment of an interactive action policy.[3] It is demonstrated that the Q-Learning algorithm converges to the optimal proceeding when the learning Q state-action pairs is represented in a table containing the full information of each pair value [37].

The proposed methodology considers an adaptation of the Q-Learning algorithm to undertake the learning process. The basic concept behind the proposed Q-Learning adaptation is that the learning algorithm can learn a function of optimal evaluation over the whole space of context-scenario pairs ($c \times s$), thus introducing a context awareness component to the standard algorithm. This evaluation defines the Q confidence value that each scenario can represent the actual encountered negotiation scenario s in context c. For instance, an agent that operates in an environment formed by a set of possible contexts where the agent can choose actions within a set of possible actions, so each time that the player performs an action, it receives a reinforcement value. Thus, the only learning source is the agents' own experience, whose goal is to acquire an actions policy that maximizes its overall performance [37]. The Q function performs the mapping as in the Eq. (7).

$$Q\colon c \times s \rightarrow U \tag{7}$$

The U is the expected utility value when selecting a scenario s in context c. The expected future reward, when choosing the scenario s in context c, is learned through trial and error according to Eq. (8).

$$Q_{t+1}(c_t, s_t) = Q_t(c_t, s_t) + \alpha(c_t, s_t)[r(s, c, t) + \gamma U_t(c_{t+1}) - Q_t(c_t, s_t)] \tag{8}$$

[3] Policy refers to the strategy or procedure to determine which actions to try in each moment taking into account the need for the balance between exploration of unknown actions and exploitation of the best actions found so far.

The c_t is the kind of context when performing under scenario s_t at time t:

- $Q_t(c_t, s_t)$ represents the value of the previous iteration (each iteration represents each new contract established in the given scenario and context). Generally, $Q_t(c_t, s_t) = 0,\ \ t = 0,\ \ \forall\ c, s$.
- $\alpha(c_t, s_t)(0 < \alpha \leqslant 1)$ is the learning rate which determines the extent to which the newly acquired information will replace the old information (*e.g.* assuming a value of 0 learns nothing; on the other hand, a value of 1 represents a fully deterministic environment).
- $r(s, c, t)$ is the reward, which represent the quality of the pair context-scenario ($c \times s$). It appreciates the positive actions with high values and negative with low values, all of them are normalized on a scale from 0 to 1. The reward r is defined in Eq. (9).

$$r(s, c, t) = 1 - |RP_{c,t,a,p} - EP_{s,c,t,a,p}| \tag{9}$$

The $RP_{c,t,a,p}$ represents the real price that has been established in a contract with an opponent p, in context c, in time t, referring to an amount of power a; and $EP_{s,c,t,a,p}$ is the estimation price of scenario that corresponds to the same player, amount of power and context in time t. All r values are normalized so that $r(s, c, t) \in [0, 1], \forall\ s, c, t$.

- $\gamma \in [0, 1]$ is the discount factor which determines the importance of future rewards. A value of 0 only evaluates current rewards, and higher values than 0 takes into account future rewards.
- $U_t(c_{t+1})$ is the estimation of the optimal future value which determines the utility of scenario s, resultant in context c. U_t is calculated as in Eq. (10).

$$U_t(c_{t+1}) = \max_s Q(c_{t+1}, s) \tag{10}$$

The proposed adaptation of the Q-Learning algorithm is executed as shown in the flowchart of Fig. 2:

Fig. 2 shows the learning process of the proposed methodology. The several steps of this model can be executed as follows:

- For each c and s, initialize $Q(c, s) = 0$;
- Observe new event (new established contract);
- Repeat until the stopping criterion is satisfied:
  - Select new scenario for current context;
  - Receive immediate reward $r(s, c, t)$;
  - Update $Q(c, s)$ according to the Eq. (8);
  - Observe new context c';
  - $c \rightarrow c'$.

After each update, all Q values are normalized according to the Eq. (11), to facilitate the interpretation of values of each scenario in a range from 0 to 1.

$$Q'(c, s) = \frac{Q(c, s)}{\max[Q(c, s)]} \tag{11}$$

The proposed learning model assumes the confidence of Q values as the probability of a scenario in a given context. $Q(c, s)$ learns by treating a forecast error, updating each time a new observation (new established contract) is available again. Once all pairs context-scenario have been visited, the scenario that presents the highest Q value in the last update is chosen by the learning algorithm as the most likely scenario to occur in actual negotiation under the corresponding context.

4. Case study

This section presents a case study with the goal of demonstrating the performance of the proposed methodology. For this case study, a historical database, concerning the past log of established contracts of different electricity market players, is used to apply the proposed

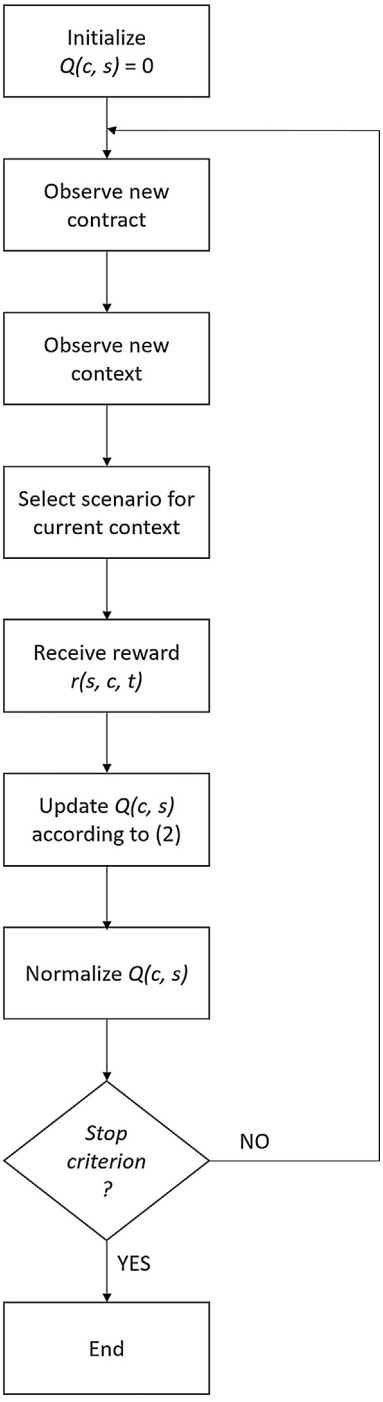

Fig. 2. Flowchart of the proposed learning model.

Table 1
Dataset Overview.

	MIN	AVG	STDEV	MAX
Contracts/Period	128	157	17,78	180
Contracts/Day	147	3753	485,78	4287
Contracts/Player	2	27244	58653,22	288160
Contracts/Player/Period	1	5	6,83	29
Power/Period/Contract	1	69,04	6,25	3575
Power/Player/Contract	1	89,05	223,17	3575
Power/Period	7718	10813	1346,38	14128
Power/Day	8210	258405,89	34317,46	316801
Power/Player	30	1875400,33	4503101,94	26081833

methodology and assess its performance. The used data is based on real data extracted from MIBEL (Mercado Ibérico de Eletricidade), the Iberian Electricity Market [4]. The dataset can be consulted in [38] and is composed by the executed physical bilateral contracts declared in the Spanish System Operator, in the period between 1 July 2007 and 31 October 2008 (16 months/ 488 days). Each negotiation day is composed by 24 negotiation periods (one per hour), in a total of 11712 periods. The negotiations were performed by 132 different players (88 Buyers and 44 Sellers) which established 1797996 contracts. Table 1 presents a detailed overview of the dataset.

The overall goal is to update the Q value of each forecast method (scenario) and context whenever there are new contracts. It is also important to test different combinations of input parameters, such as discount factor and learning rate; to analyse the evolution of Q values; and to have a suitable learning mechanism, which chooses the most likely forecast method to occur (*i.e.* the scenario with a lower forecast error in current context).

Hence, by means of a previous sensitivity analysis, a good balance among the learning parameters has found to be essential to guarantee a good quality of the Q-Learning algorithm results. It is possible to conclude that this balance should be chosen considering the players' expected interaction (*i.e.* the number of established contracts). It has been experimented that, when using high learning rate values (LR), where learning is fast, the value of Q function will only reflect the latest iterations and Q values vary more. This way, it is not as reliable but the algorithm is adapted faster, therefore it is suitable for situations in which the expected number of contracts is small (*e.g.*, contexts where contracts occur infrequently). On the other hand, with smaller LR, the algorithm is not able to adapt as fast, but it is more consistent because the previous historic results have more influence. However, a learning process that is too slow is not also advisable, since the new information observed becomes almost irrelevant for the learning process. Thus, a suitable balance between the consideration of new events and the previous values already learned by the system is crucial. Regarding the other learning parameter, the discount factor (DF), it has been concluded that high values are the most suitable values for a quality convergence, since otherwise the variation is very large at every iteration.

Finally, the last specification has the aim of performing a complete study to analyze the influence of contexts in the undertaken bilateral negotiation. For that purpose, two different simulation studies are conducted: a test without considering the negotiation context (all established contracts were undertaken in the same conditions); and a case considering a set of different negotiation contexts. Both simulations (with and without contexts) consider exactly the same data, the only difference is that the contextual dimension is not considered in the first simulations (all contracts are assumed to be undertaken under the same contextual conditions), while when considering the contexts, each contract is associated to a specific context.

In order to evaluate the results of the case study, a negotiation based simulation environment [39] is used, namely the MASCEM agent based simulator, through the integration of the proposed method in the DECON decision support system.

4.1. Input data review

The first step of this study is to perform an analysis of the input data to validate the Q-Learning algorithm results. This analysis is essential to anticipate the forecast method that is the most likely to occur, in the current context, for each competitor player, for each case in the established contracts log. In this way it is possible to assess the achieved results, by comparing them to what would be expected.

On the one hand are analysed the forecast contract prices that result from 5 different algorithms, generated by DECON, for each subject player, where there is an expected price for each amount of energy (from 1 until 10 MW). The Fig. 3 shows the different scenarios. It is visible that Scenario 1 has high contract prices for low amounts of

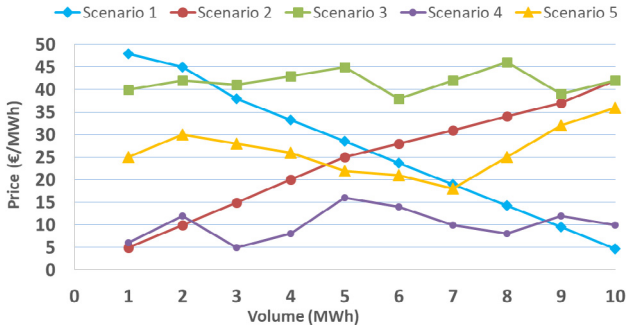

Fig. 3. Forecast contract prices for each scenario.

power and low prices for high amounts too. The opposite occurs in Scenario 2, with low contract price values for low power amounts and large prices for large amounts of energy. Scenario 3 always present a large contract price to any amount of energy. The latter Scenario, Scenario 5, shows intermediate values with exception to the high power amounts, where it shows high prices.

On the other hand, it is also analysed the historic log of negotiations. The Fig. 4 presents the previous established bilateral contracts with information about the negotiated price and power, without context awareness. It can be observed that the first contracts always have large prices for almost all power amounts, but at the end (from contract 37 onwards) starts to change the trend, with low contract prices for high power amounts.

By matching Figs. 3 and 4, it can be seen that the best expected scenario until contract 36 is scenario 3, as it defines a high price regardless of the defined traded amount. From contract 37 onwards, the best expected scenario is scenario 1, as it defines high prices for low amounts of traded energy, and low prices for large volumes. This is supported by Table 2 which shows the comparison of the average error between the predicted price by each scenario and the actual verified price. The error evaluation is measured using the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Standard Deviation (STD).

From Table 2 it is possible to confirm that the scenario that presents the most accurate prediction (lower error) until contract No. 37 is scenario 3. In the total of all contracts, the scenario with the smaller prediction error is scenario 1, although with only a slight difference from scenario 3.

4.2. Results

This subsection presents the results of the implemented learning model for each competitor player. As previously mentioned, to validate the Q-Learning results, it is essential to compare the data input with the obtained results, which must be in accordance with the expected scenario, throughout the iterations. Since the reinforcement learning

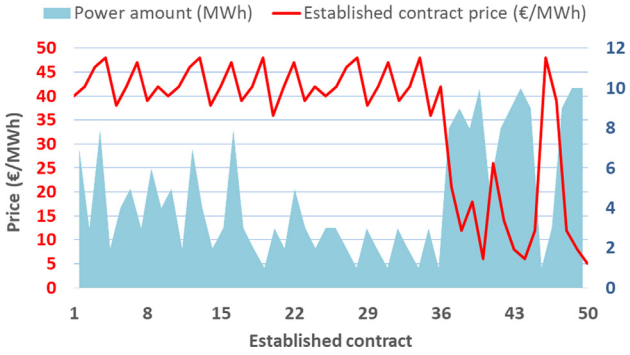

Fig. 4. Historic log of established contracts.

Table 2
Comparison of the prediction error of the five considered scenarios.

No. contracts	Scenario	MAE	MAPE (%)	STD
37	1	5.27	9.15	5.63
	2	18.85	27.45	10.40
	3	4.83	7.25	4.72
	4	23.44	32.51	12.27
	5	9.54	14.16	8.23
50	1	5.02	8.16	5.61
	2	19.31	26.98	12.31
	3	5.24	9.26	6.81
	4	18.32	25.31	12.84
	5	11.06	16.83	9.16

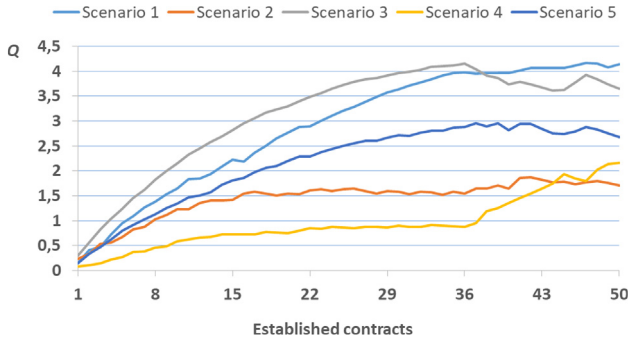

Fig. 5. Q-Learning algorithm evolution throughout established contracts. Parameters: LR = 0.3; DF = 0.8.

algorithm learns throughout established contracts in time, it is also presented the evolution and convergence of Q values over each iteration. This study is conducted for two different test cases: with and without context awareness.

4.2.1. Learning without context awareness

Fig. 5 presents the learning process.

By observing Fig. 5, it is possible to verify the evolution of the Q value throughout the 50 established contracts, for each scenario. A DF of 0.8 and LR of 0.3 are used, to consider a quick learning rate, with the aim of facilitating the fast adaptation to the most recent perceived events. As it can be seen, the most probable scenario is Scenario 3, from the first contract until contract 36, where Scenario 1 surpass the Q value of Scenario 3. To allow a more detailed analysis, the Table 3 presents the normalized Q value for each scenario at every 5 new contracts.

The normalized Q values of Table 3 allow a better identification of the most probable scenario (the value of 1 indicates the recommended scenario by the algorithm, until the last observation). By comparing the results with the previous analyses of the data input, it is visible in

Table 3
Normalized Q values of each scenario throughout 50 iterations.

Contract	Scenario				
	1	2	3	4	5
5	0.76	0.55	1	0.22	0.64
10	0.77	0.57	1	0.27	0.62
15	0.79	0.50	1	0.25	0.64
20	0.84	0.47	1	0.23	0.67
25	0.86	0.44	1	0.23	0.67
30	0.92	0.40	1	0.23	0.68
35	0.96	0.39	1	0.22	0.70
40	1	0.42	0.94	0.34	0.71
45	1	0.44	0.89	0.48	0.68
50	1	0.41	0.88	0.52	0.65

Table 4
Q last values (contract 50) of each scenario.

Scenario				
1	2	3	4	5
4.14	1.71	3.65	2.16	2.68

Table 3 and Fig. 5 that the best scenarios (higher Q value) are the Scenario 3 for first contracts (exactly till contract 36) and, from then forward, is the Scenario 1, as identified in Table 3. Nevertheless, it was predicted that Scenario 4 could also occur for the last contracts. However, the algorithm learned properly due to the past trend of high contract prices and low power amounts, which Scenario 1 also had. These results match the best expected scenarios, as identified in Table 2.

The Table 4 present a summary of the results for each scenario of each analysed case (without context analysis). The presented values are the last learned Q value for each Scenario. Thus, the scenario that presents the highest Q value is identified as the most likely scenario to occur under actual negotiation. From this table it can be concluded that in the end of the 50 contracts, the best scenario has been identified. Scenario 1 is predicted as the best scenario, as identified in Tables 2 and 3, and matching the expected result shown in Table 2.

4.2.2. Context aware learning

This sub-section presents a test that considers the current context of the negotiation, since it usually influences the negotiation environment of players, as previously mentioned. Table 5 shows the average prediction error achieved by the different scenarios for each context.

From Fig. 5 it is visible that scenarios with lowest prediction error are: Scenario 1 for Contexts 2 and 4, and Scenario 3 for Contexts 1 and 3. Fig. 6 shows the evolution of Q-Learning for each scenario of Context 1.

In Fig. 6, it can be observed that Player 1 has a small number of contracts in Context 1, which does not give enough iterations to the learning algorithm to guarantee a good learning process. Therefore, in this case, it is necessary a high LR value to achieve faster learning. The Scenario 3 was the best forecast method, followed by Scenario 2 and 5. The evolution of the learning process can be seen in more detail in Table 6, which presents the normalized Q values through the iterations.

Table 5
Comparison of the prediction error of the five considered scenarios in each of the four considered contexts.

Context	Scenario	MAE	MAPE(%)	STD
1	1	13.94	18.43	16.34
	2	8.75	12.48	6.36
	3	3.26	4.12	2.36
	4	19.74	27.39	17.47
	5	12.89	17.62	9.20
2	1	3.67	5.26	4.62
	2	19.84	28.53	15.62
	3	3.82	5.48	4.89
	4	26.84	39.98	18.74
	5	10.31	14.53	8.38
3	1	3.91	7.52	4.93
	2	9.93	14.45	8.22
	3	3.74	7.16	4.53
	4	14.42	19.36	12.60
	5	6.22	9.36	6.83
4	1	5.46	8.63	7.31
	2	20.31	33.16	16.82
	3	24.17	38.28	18.45
	4	8.02	12.21	8.24
	5	15.16	21.75	13.82

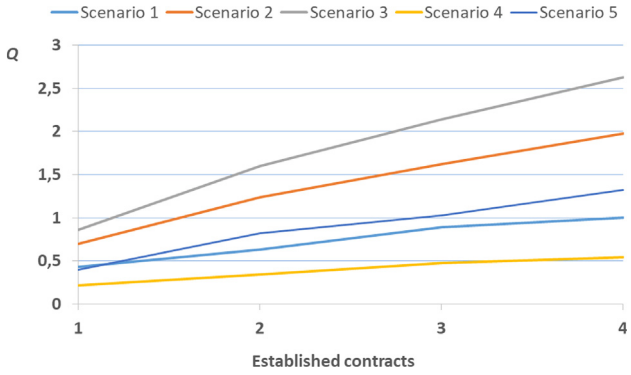

Fig. 6. Q-Learning algorithm evolution throughout established contract in Context 1. Parameters: LR = 0.9; DF = 0.8.

Table 6
Normalized Q values of each scenario throughout iterations in Context 1.

Contract	Scenario				
	1	2	3	4	5
1	0.50	0.81	1	0.26	0.47
2	0.39	0.77	1	0.21	0.51
3	0.42	0.76	1	0.22	0.48
4	0.38	0.75	1	0.21	0.50

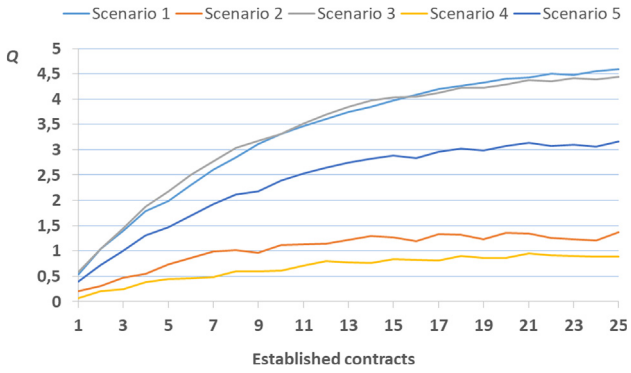

Fig. 7. Q-Learning algorithm evolution throughout established contract in Context 2. Parameters: LR = 0.6; DF = 0.8.

By analysing the context 2 from Fig. 7, it becomes evident that there are two scenarios whose evolution is almost the same (Scenarios 1 and 3). In this case, the number of established contracts is higher, and therefore, it is recommended a lower LR value than in the previous case. It is not able to adapt so fast, but it is more consistent because the previous results have more impact, increasing the results' reliability. The Table 7 shows the normalized Q values throughout iterations among Player 1 and supported player.

In Context 3, presented in Fig. 8, it is visible that the Scenario 3 is clearly the best scenario method. To adapt the learning process to this

Table 7
Normalized Q values of each scenario throughout iterations in Context 2.

Contract	Scenario				
	1	2	3	4	5
5	0.91	0.34	1	0.20	0.68
10	1	0.34	1	0.18	0.72
15	0.98	0.31	1	0.21	0.71
20	1	0.31	0.98	0.19	0.70
25	1	0.30	0.97	0.19	0.69

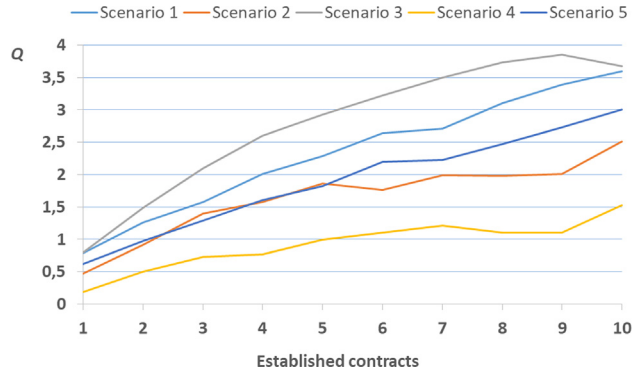

Fig. 8. Q-Learning algorithm evolution throughout established contract in Context 3. Parameters: LR = 0.9; DF = 0.8.

Table 8
Normalized Q values of each scenario throughout iterations in Context 3.

Contract	Scenario				
	1	2	3	4	5
2	0.84	0.62	1	0.33	0.65
4	0.77	0.61	1	0.30	0.62
6	0.82	0.55	1	0.34	0.68
8	0.83	0.53	1	0.30	0.66
10	0.98	0.68	1	0.42	0.82

context (small number of iterations), it is necessary a LR with a high value, like happened in context 1. The Table 8 presents a summary of normalized Q values throughout iterations for each scenario corresponding to the specifications of Fig. 6.

It can be observed in Fig. 9 that the learning process for the Context 4 is also in accordance with the one predicted in the priori-analyses of the simulation. The Scenario 1 presents the highest Q value, followed by Scenario 4. This case is similar to Context 3 in terms of number of established contracts, so the same learning parameters are chosen. Table 9 shows the Q normalized values throughout iterations for inputs and parameters described for Fig. 9.

The Table 10 presents the last learned Q value for each scenario, for each context of each case study (with context awareness).

When comparing the summary results tables of different study cases (Table 4 without context analysis and Table 10 with context analysis), it is possible to verify that the proposed model does not learn likewise. For instance, in the case of Player 1 in Contact Log 1, the most probable scenario is Scenario 1, when considering the same negotiation environment. On the other hand, when considering different contexts, the most probable scenario is not always Scenario 1, as it is only suitable in some situations of negotiation (namely in contexts 2 and 4, as observed

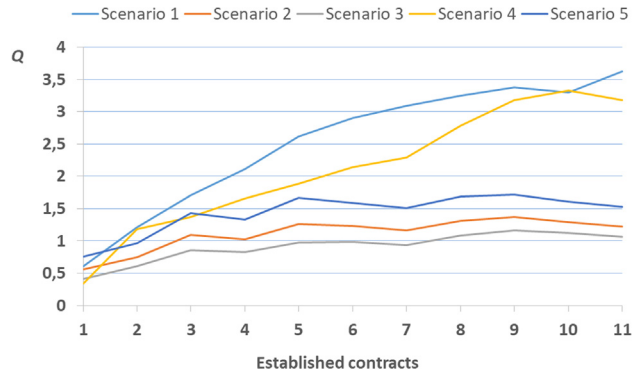

Fig. 9. Q-Learning algorithm evolution throughout established contract in Context 4. Parameters: LR = 0.9; DF = 0.8.

Table 9
Normalized Q values of each scenario throughout iterations of in Context 4.

Contract	Scenario				
	1	2	3	4	5
1	0.81	0.74	0.54	0.45	1
3	1	0.64	0.50	0.80	0.84
6	1	0.42	0.34	0.74	0.55
9	1	0.41	0.34	0.94	0.51
11	1	0.34	0.29	0.88	0.42

Table 10
Q last values of each scenario (with context awareness).

Context	Scenario				
	1	2	3	4	5
1	1.00	1.98	2.63	0.54	1.32
2	4.59	1.37	4.44	0.89	3.16
3	3.60	2.51	3.67	1.53	3.00
4	3.62	1.22	1.06	3.18	1.53

in Table 5). The same can be verified in the remaining cases, which demonstrates the importance of context analysis to obtain a contract price estimation that is more adapted to reality.

When learning on the same exact data, without considering the contextual dimension, the traditional Q-Learning algorithm is able to learn the overall best scenario that best fits the general data. However, when adding a contextual dimension, the proposed model is able to learn the best specific action in each specific scenario, thus being able to achieve a higher quality of results, adapted to the context.

This way, by comparing the expected results, presented in the pre-analysis of the input data and in Table 5, with the actual results, for the different test cases, it is possible to validate the proposed approach. The proposed algorithm is able to learn which of the potential scenarios is the closest approximation of the negotiation environment that the supported player will face.

4.2.3. Sensitivity analysis

This section provides an overview of the sensitivity analysis performed to find the best parametrization of the proposed method in the different performed tests. Fig. 10 presents several heat maps showing the quality of the results (overall prediction errors) achieved by the method when applied to the cases with no contextual learning, and to each of the contexts independently. The heat maps include the combinations between the values of the LR and DF. The dark green zones in the graphs represent the combinations of LR and DF that present the best performance in each test, while the dark red zones represent the worst combinations between the parameters.

From Fig. 10 it can be seen that the best combination of parameters when considering no contextual learning is LR = 0.3 and DF = 0.8. The best combination when applying the proposed method to Context 1 is LR = 0.9 and DF = 0.8. The higher LR in this case is a result from the low number of contracts under this context, which mean that a very fast learning process is required. In Context 2, the best parametrization is LR = 0.6 and DF = 0.8. Note that this is the context for which there is the larger number of contracts, which enables decreasing the LR and perform a more sustainable learning process throughout the time, although not as low as in the case with no contextual learning (much larger number of considered contracts). For contexts 3 and 4 the best parametrization is LR = 0.9 and DF = 0.8. These two contexts also have a small number of contracts, but still much more than in Context 1. It can be seen that, although the best combination of parameters is found for high values of LR in order to cope with the small number of contracts, the green zone in these two graphs extends further

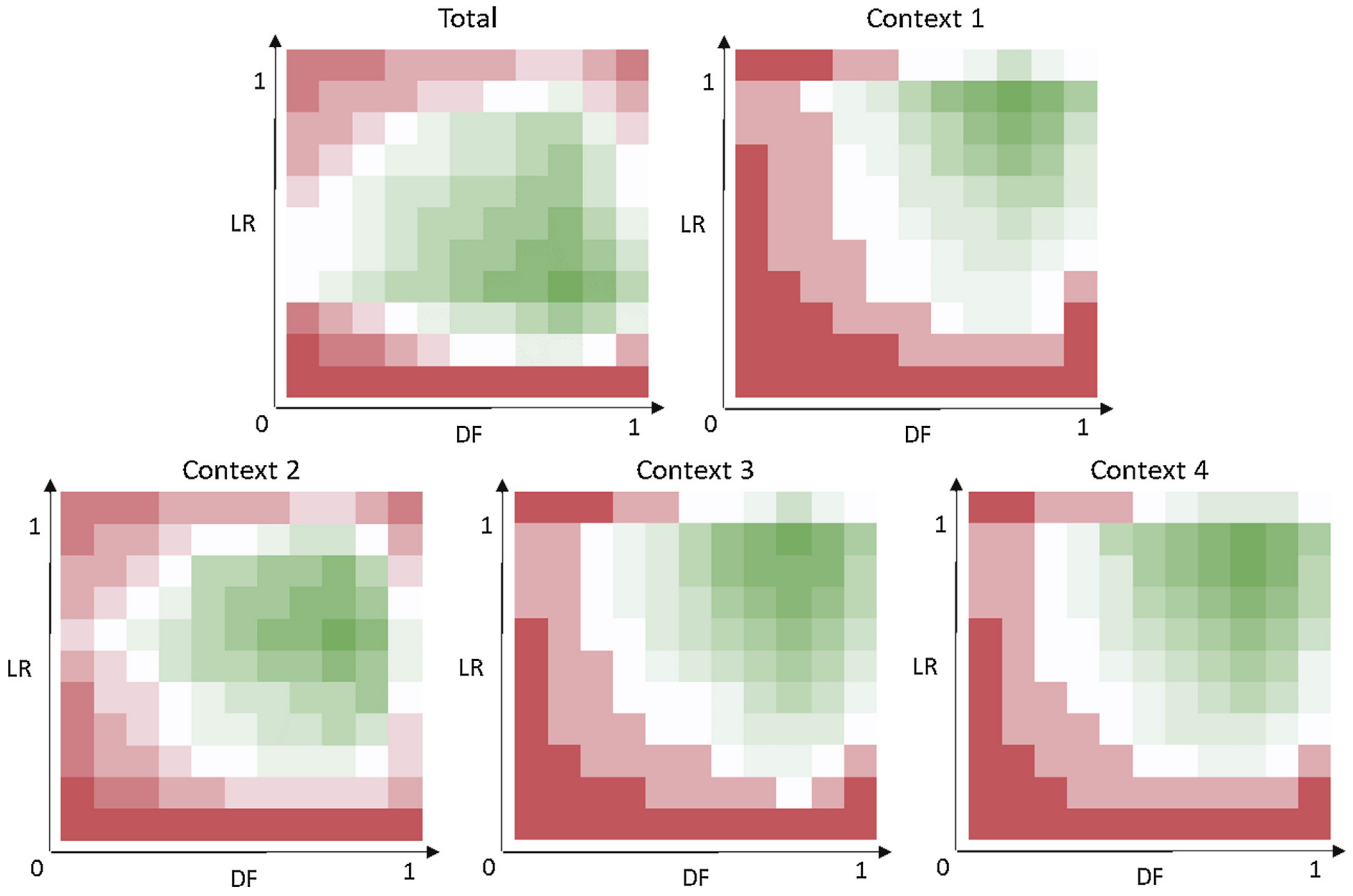

Fig. 10. Sensitivity analysis results for the combination between LR and DF in the different performed tests.

downwards when compared to the graph of Context 1, which means that the slightly larger number of contracts starts to enable a small decrease in the LR with still good results. These identified combinations of parameters are those used in the performed tests.

4.3. Illustrative example

This subsection includes an illustrative scenario considering only two contexts: Weekends and Business days, in which the proposed methodology learns over a total of 2500 contracts for each context. The proposed methodology is run with a LR of 0.3, allowing a slow learning, and DF of 0.8, favouring future rewards, considering the available amount of data. Five alternative forecasting methods are used, namely 3 alternative models based on ANN, one SVM one approach that simply calculates the average of past values. Figs. 11 and 12 present the normalized Q value of each scenario under Weekends and Business days

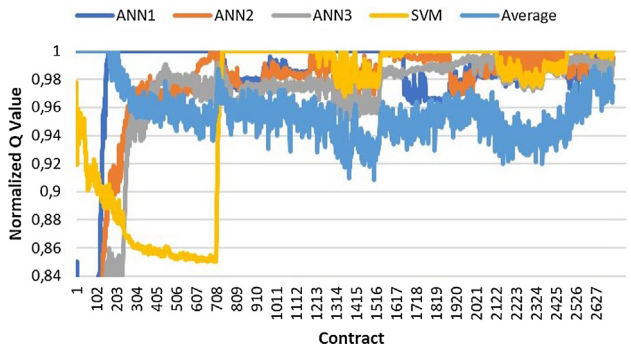

Fig. 11. The learning process for the Business days context.

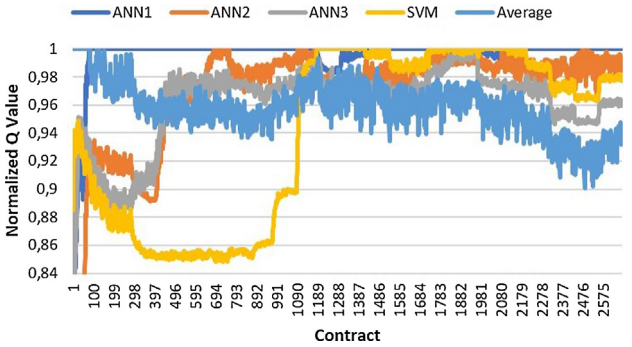

Fig. 12. The learning process for the Weekend context.

contexts, respectively, over all the analysed contracts.

The Fig. 11 proves that SVM scenario is really dominant, being the scenario with the maximum *Q* value during more contracts. However, it is not always the most probable scenario. In fact, the SVM scenario were very far from reality in the first 708 contracts, the period in which ANN1 dominated [102, 707], after the initial success of the Average scenario [1, 101]. Then, the SVM scenario is only surpassed by ANN1 [1314, 1516], and ANN2 [2122, 2425]. The success of each ANN method is measured by the amount of contracts considered. The fewer the number of contracts, the better the results. The Average scenario only had success in the beginning of the learning process, as it is a simple average, which does not requires much learning to know its potential, contrary to the other scenarios. Having seen the learning process of the Business days context, it may be interesting to see how it compares to a different context, which in this case is the Weekend (Fig. 12).

It is visible in Fig. 12 that SVM scenario does not have as much

Table 11
Comparative results between the proposed model and state of the art reinforcement learning algorithms.

Algorithm	Context	Scenario				
		1	2	3	4	5
Proposed Model	1	0.38	0.75	**1.00**	0.21	0.50
	2	**1.00**	0.30	0.97	0.19	0.69
	3	0.98	0.68	**1.00**	0.42	0.82
	4	**1.00**	0.34	0.29	0.88	0.42
Standard Q-Learning	1	**1.00**	0.41	0.88	0.52	0.65
	2	**1.00**	0.41	0.88	0.52	0.65
	3	**1.00**	0.41	0.88	0.52	0.65
	4	**1.00**	0.41	0.88	0.52	0.65
Roth-Erev	1	**1.00**	0.28	0.92	0.41	0.56
	2	**1.00**	0.28	0.92	0.41	0.56
	3	**1.00**	0.28	0.92	0.41	0.56
	4	**1.00**	0.28	0.92	0.41	0.56
UCB1	1	**1.00**	0.23	0.76	0.53	0.72
	2	**1.00**	0.23	0.76	0.53	0.72
	3	**1.00**	0.23	0.76	0.53	0.72
	4	**1.00**	0.23	0.76	0.53	0.72
EXP3	1	**1.00**	0.32	0.63	0.45	0.42
	2	**1.00**	0.32	0.63	0.45	0.42
	3	**1.00**	0.32	0.63	0.45	0.42
	4	**1.00**	0.32	0.63	0.45	0.42

success as the one presented in the Business days context.

4.4. Benchmark study – statistical analysis

This subsection shows a comparison of the results achieved by the proposed model, as described in Section 4.3, the results of the standard Q-Learning (Section 4.2) and also the results under the same simulation settings, of using other reference state of the art reinforcement learning algorithms, namely Roth-Erev [40], UCB1 [41] and EXP3 [42]. Table 11 shows the global results, *i.e.* normalized confidence values (or Q values) in each of the 5 considered scenarios, in each of the 4 considered contexts. Table 12 shows the comparison of the average prediction errors resulting from the scenarios chosen in each iteration by the different algorithms in each context. This enables assessing the overall quality of the learning methods in each context. Note that it is not expected that the achieved error values match those achieved by the best scenarios themselves, as presented in Table 5, because due to the required exploration phase of the reinforcement learning algorithms several different scenarios, even if bad, must be tried, which results in an overall trial and error procedure. However, these average errors enable assessing the algorithms quality in terms of exploration vs exploitation balance, and their capability of converging to the best scenario, as shown by the confidence values in each scenario, as shown by Table 11. The bold values in Table 11 highlight the maximum value (1) in each line, which refers to the scenario that is identified as the best one for each context.

Table 11 shows that, as discussed in Section 4.3, the proposed model is able to learn and identify the best scenario for each of the four considered contexts, namely scenario 3 for contexts 1 and 3, and scenario 1 in contexts 2 and 4. On the other hand, all the other state of the art reinforcement learning algorithms are able to effectively learn the best global scenario (scenario 1), but, by not including a contextual dimension, they are not able to identify the best scenario for the specific contexts. In summary, the current algorithms are able to learn the best overall approaches, but lack the adaptation capabilities to be able to identify different performances under different contexts.

From Table 12 it can be seen that the proposed method is the algorithm that achieves the lowest prediction errors in all four contexts, as result from this method's context aware learning capability. However, some other methods reach very close results in the contexts in which the prediction is from Scenario 1 (identified by all methods as the best one, as seen from Table 11), namely in contexts 2 and 4. Nevertheless, the results from the proposed method are still better in these contexts because it is able to converge faster to the best scenario, by considering the different contexts as independent, while the other methods need for exploration (and more trial and error) to reach the best overall scenario. Tables 13 and 14 present the average prediction errors for each context for two additional data sets. These are identical to the original one used in the tests presented in the previous sections, but refer to: data set 1: January 2014 to December 2015; data set 2: January 2016 to December 2017.

From Tables 13 and 14 it is visible that the proposed method is still able to reach the lowest prediction errors from all considered benchmark algorithms in nearly all contexts in both data sets. The only

Table 12
Comparison of average prediction errors of the different algorithms in each context.

Context	Algorithm	MAE	MAPE (%)	STD
1	Proposed Model	7.45	9.89	8.98
	Standard Q-Learning	11.24	16.28	14.04
	Roth-Erev	10.49	14.87	13.41
	UCB1	15.36	21.04	18.93
	EXP3	18.56	24.90	21.39
2	Proposed Model	4.28	6.46	5.89
	Standard Q-Learning	4.88	7.23	6.68
	Roth-Erev	4.46	6.83	6.03
	UCB1	5.89	9.31	8.72
	EXP3	6.53	10.85	9.38
3	Proposed Model	5.37	8.21	6.98
	Standard Q-Learning	9.16	13.28	9.37
	Roth-Erev	8.43	12.73	9.14
	UCB1	12.54	18.02	12.71
	EXP3	15.11	22.02	17.37
4	Proposed Model	6.22	9.35	8.31
	Standard Q-Learning	6.81	9.97	9.02
	Roth-Erev	7.47	10.28	11.07
	UCB1	6.74	9.63	8.86
	EXP3	7.26	10.15	10.62

Table 13
Comparison of data set 2 average prediction errors of the different algorithms in each context.

Context	Algorithm	MAE	MAPE (%)	STD
1	Proposed Model	5.78	10.20	8.89
	Standard Q-Learning	7.38	10.92	9.21
	Roth-Erev	7.19	11.20	10.20
	UCB1	7.40	9.07	9.07
	EXP3	7.42	10.49	10.77
2	Proposed Model	5.75	8.77	7.62
	Standard Q-Learning	9.68	14.33	10.26
	Roth-Erev	8.74	12.07	9.06
	UCB1	11.34	19.25	12.40
	EXP3	15.11	23.57	18.26
3	Proposed Model	4.39	6.66	6.43
	Standard Q-Learning	5.22	6.62	6.82
	Roth-Erev	4.53	7.17	5.55
	UCB1	6.44	9.68	9.38
	EXP3	6.27	11.06	10.13
4	Proposed Model	7.51	9.02	9.44
	Standard Q-Learning	10.61	17.33	14.39
	Roth-Erev	9.65	15.13	12.20
	UCB1	15.43	20.01	18.38
	EXP3	20.09	23.63	20.29

Table 14
Comparison of data set 3 average prediction errors of the different algorithms in each context.

Context	Algorithm	MAE	MAPE (%)	STD
1	Proposed Model	5.90	8.48	6.92
	Standard Q-Learning	9.04	12.57	10.07
	Roth-Erev	7.98	12.44	9.27
	UCB1	11.34	16.53	11.51
	EXP3	15.00	22.41	18.64
2	Proposed Model	7.03	10.07	8.66
	Standard Q-Learning	11.64	17.01	13.57
	Roth-Erev	10.84	15.64	12.52
	UCB1	16.33	19.77	19.93
	EXP3	18.72	25.96	22.54
3	Proposed Model	5.74	9.66	8.64
	Standard Q-Learning	6.45	10.37	9.32
	Roth-Erev	7.18	9.57	10.55
	UCB1	7.18	9.58	8.30
	EXP3	7.70	10.71	11.30
4	Proposed Model	4.61	6.73	5.66
	Standard Q-Learning	4.62	7.35	6.56
	Roth-Erev	4.46	6.23	6.14
	UCB1	5.63	9.45	8.94
	EXP3	5.98	10.33	9.46

exception in context 4 of data set 3, in which Roth-Erev is able to achieve slightly lower average prediction errors.

The Kruscal-Wallis test is the nonparametric test used to compare three or more independent samples. Indicates if there is a difference between at least two of them. This is used to test the null hypothesis that all populations have equal distribution functions against the alternative hypothesis that at least two of the populations have different distribution functions. In this way it is assumed that equality of averages when equality of equal distributions exists [43].

By the test Kruscal-Wallis it is possible to obtain the value of $p = 0$ that gives us indication of rejection of the null hypothesis that all data samples come from the same distribution at a 1% significance level. Given the result of the test that gives the indication of the null hypothesis, the comparison between the pairs of groups is made in order to verify which of the samples differ from each other.

The Bonferroni procedure is performed in order to make the comparison in pairs. Fig. 13 represents the 95% confidence interval for all sample groups (5 methods, in which group 1 is the proposed method), in the total of all executions using the three data sets. In this way, it is possible to see which groups differ in the value of the average, using the Bonferroni procedure.

By analyzing the graph of Fig. 13, it is possible to observe that all methods have significantly different mean values. Table 15 shows the results of this analysis.

Since the p-value is equal to 1 in all these group tests, the null hypothesis where the groups are considered to have similar means with an error of 5% is accepted.

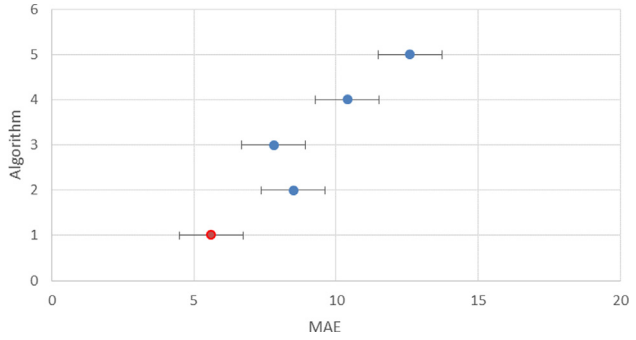

Fig. 13. Bonferroni confidence interval by 95%.

Table 15
Bonferroni procedure.

Group pairs		p-value
1	1	1
1	3	1
1	4	1
1	5	1

Taking into account this analysis, it is concluded that the applied benchmark methods achieve significantly different results, thus supporting the relevance of the proposed approach.

5. Conclusions

The EM restructuring and the growth in penetration of distributed energy resources, introduced the need of a better preparation, by the part of the participating players in this dynamic environment, which trade constantly in different situations. Currently, automated negotiations are an active area of research within the field of computing, particularly with the development of artificial intelligence. However, in EM field there is not significant works to support automated negotiation decisions, as previously mentioned in the introductory section, especially those regarding the analysis of previous information from competitor players, and in particular regarding the pre-negotiation stage of negotiations.

This paper proposes a model, integrated into the DECON system, to provide decision support for the pre-negotiation step of bilateral contracts in the electricity market. In summary, the pre-negotiation is a stage that has great importance because it performs all the preparation and planning of actual negotiation. This process aims to identify the ideal negotiators that supported player could trade with to obtain the greatest possible benefit.

A common behaviour of the players, when performing bilateral negotiations, is the strategic definition of prices for different energy amounts of each competitor player, to have an adequate forecasting about possible contract price, before the negotiation. The proposed methodology is focused in the bilateral contract price estimation approach, based on the application of an adaptation of the Q-Learning reinforcement learning algorithm. This way, the implemented model can learn which of the potential scenarios is the most probable to represent a reliable approximation of the actual negotiation between the supported player and the competitor in matter, depending on the negotiation context. The potential negotiation scenarios are determined by using an explicitly modelling of gathered information about past actions of opposing negotiators. Additionally, is performed an analysis and definition of different contexts of bilateral contract negotiation in EM. This way, it is possible to properly represent the behaviour of negotiating players, as they usually adapt their actions to the distinct circumstances they are encountering in each moment.

From the analysis of the implemented learning process, it can be concluded that a balance of the learning parameters is very important for the quality of the results of the Q-Learning algorithm. Therefore, the LR should be selected according to the number of expected observations. In relation to the DF parameter, it is possible to conclude that, for this study, higher values are more desirable. Regarding the context study, the results of the case study show that the context awareness provides the Q-Learning algorithm with a more realistic learning. The context definition process considers some influential conditions that affect the agreement of contract prices. Therefore, it can be concluded that suiting actions to different contexts allows to adapt the behaviour of negotiating entities to the different circumstances, improving their decision making-process.

Finally, it is noteworthy that it is also demonstrated that the simulated process is in accordance with the previous analysis of the input

data. The scenario which should be chosen as the most reliable was effectively the scenario that has obtained the largest Q values, at the end of learning process.

As future work, other learning techniques can be experimented, such as adapting Roth-Erev algorithm, models based on the Bayes theory of conditional probability. Moreover, promising emerging approaches such as adaptive probabilistic behavioural learning [44], bulk negotiation behavioural learning [45], and probabilistic decision making [46], will also be considered as alternative approaches, in specific for the actual negotiation process. This way it is possible to compare their results with the proposed model to facilitate the choice of the most appropriate learning method for each type of problem.

Acknowledgments

This work has been developed in the scope of the European Union's Horizon 2020 Research and Innovation Programme under the Marie Sklodowska-Curie grant agreement No 703689 (project ADAPT) and grant agreement No 641794 (project DREAM-GO); and has also been supported by the CONTEST project – SAICT-POL/23575/2016.

References

[1] Shahidehpour M, Yamin H, Li Z. Market operations in electric power systems: forecasting, scheduling, and risk management. Institute of Electrical and Electronics Engineers, Wiley-Interscience; 2002. https://doi.org/10.1002/047122412X. arXiv:arXiv:1011.1669v3.
[2] Meeus L, Purchala K, Belmans R. Development of the internal electricity market in Europe. Electricity J 2005;18(6):25–35. https://doi.org/10.1016/j.tej.2005.06.008.
[3] Klessmann C, Held A, Rathmann M, Ragwitz M. Status and perspectives of renewable energy policy and deployment in the European Union-What is needed to reach the 2020 targets? Energy Policy 2011;39(12):7637–57. https://doi.org/10.1016/j.enpol.2011.08.038.
[4] OMIE. Omie; 2018. URL:http://www.omie.es [accessed March-2018].
[5] MISO. Miso Energy; 2018. URL:https://www.misoenergy.org [accessed March-2018].
[6] Algarvio H, Lopes F, Santana J. Bilateral contracting in multi-agent energy markets: forward contracts and risk management. Communications in Computer and Information Science Cham: Springer; 2015. p. 260–9. https://doi.org/10.1007/978-3-319-19033-422. vol. 524.
[7] Pinto T, Vale Z, Sousa TM, Praça I. Negotiation context analysis in electricity markets. Energy 2015;85:78–93. https://doi.org/10.1016/j.energy.2015.03.017.
[8] Pinto T, Vale Z, Sousa TM, Praça I, Santos G, Morais H. Adaptive learning in agents behaviour: a framework for electricity markets simulation. Integr Computer-Aided Eng 2014;21(4):399–415. https://doi.org/10.3233/ICA-140477.
[9] Koritarov VS. Real-world market representation with agents. IEEE Power Energy Mag 2004;2(4):39–46. https://doi.org/10.1109/MPAE.2004.1310872.
[10] Li H, Tesfatsion L. Development of open source software for power market research: the AMES test bed. J Energy Markets 2009;2(2):111–28.
[11] Cincotti S, Gallo G. The genoa artificial power-exchange. Communications in computer and information science Berlin, Heidelberg: Springer; 2013. p. 348–63. https://doi.org/10.1007/978-3-642-36907-023. vol. 358.
[12] Vale Z, Pinto T, Praça I, Morais H. MASCEM: electricity markets simulation with strategic agents. IEEE Intell Syst 2011;26(2):9–17. https://doi.org/10.1109/MIS.2011.3.
[13] Lopes F, Wooldridge M, Novais AQ. Negotiation among autonomous computational agents: principles, analysis and challenges. Artif Intell Rev 2008;29(1):1–44. https://doi.org/10.1007/s10462-009-9107-8.
[14] Pinto T, Vale Z, Praça I, Pires EJ, Lopes F. Decision support for energy contracts negotiation with game theory and adaptive learning. Energies 2015;8(9):9817–42. https://doi.org/10.3390/en8099817.
[15] Golberg D. Genetic algorithms in search, optimization, and machine learning. Addison-Wesley Longman Publishing Co., Inc; 1989. vol. 1989.
[16] Gaines DA, Pakath R. An examination of evolved behavior in two reinforcement learning systems. Decis Support Syst 2013;55(1):194–205. https://doi.org/10.1016/j.dss.2013.01.019.
[17] Veselka T, Boyd G, Conzelmann G, Koritarov V, Macal C, North M, et al. Thimmapuram, simulating the behavior of electricity markets with an agent-based methodology: the electric market complex adaptive systems (emcas) model; 2002.
[18] Lin R, Kraus S, Baarslag T, Tykhonov D, Hindriks K, Jonker CM. GENIUS: an integrated environment for supporting the design of generic automated negotiators. Comput Intell 2014;30(1):48–70. https://doi.org/10.1111/j.1467-8640.2012.00463.x.
[19] Lopes F, Rodrigues T, Sousa J. Negotiating bilateral contracts in a multi-agent electricity market: a case study. In: 2012 23rd International Workshop on Database and Expert Systems Applications; 2012. pp. 326–330. doi:10.1109/DEXA.2012.77.
[20] Jain AK. Data clustering: 50 years beyond K-means. Pattern Recogn Lett 2010;31(8):651–66. https://doi.org/10.1016/j.patrec.2009.09.011. arXiv:0402594v3.
[21] Pinto T, Sousa TM, Vale Z. Dynamic artificial neural network for electricity market prices forecast. Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on IEEE; 2012. p. 311–6. https://doi.org/10.1109/INES.2012.6249850.
[22] Rigatos GG. Adaptive fuzzy control for differentially flat MIMO nonlinear dynamical systems. Integr Computer-Aided Eng 2013;20(2):111–26. https://doi.org/10.3233/ICA-130421.
[23] Sheblé GB. Computational auction mechanisms for restructured power industry operation. Springer Science & Business Media; 1999. vol. 500.
[24] Algarvio H, Lopes F. Risk management and bilateral contracts in multi-agent electricity markets. Cham: Springer International Publishing; 2014. p. 297–308.
[25] Kirschen DS, Strbac G. Fundamentals of power system economics. John Wiley & Sons; 2004.
[26] MIBEL. Mibel; 2018. URL:http://www.mibel.com [accessed March-2018].
[27] Fujita K, Ito T, Zhang M, Robu V. Next frontier in agent-based complex automated negotiation, vol. 596 of studies in computational intelligence. Japan, Tokyo: Springer; 2015. https://doi.org/10.1007/978-4-431-55525-4.
[28] Von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton University Press; 2007.
[29] Aristizabal JA, Ramos-Álvarez MM, Callejas-Aguilera JE, Rosas JM. Testing a cue outside the training context increases attention to the contexts and impairs performance in human predictive learning. Behav Processes 2017;145:31–6.
[30] Liu Y, Liu B, Shan L, Wang X. Modelling context with neural networks for recommending idioms in essay writing. Neurocomputing 2018;275:2287–93. https://doi.org/10.1016/j.neucom.2017.11.005.
[31] Pal R. Context-sensitive probabilistic boolean networks: Steady-state properties, reduction, and steady-state approximation. IEEE Trans Signal Process 2010;58(2):879–90. https://doi.org/10.1109/TSP.2009.2030832.
[32] Sama M, Elbaum S, Raimondi F, Rosenblum DS, Wang Z. Context-aware adaptive applications: fault patterns and their automated identification. IEEE Trans Software Eng 2010;36(5):644–61. https://doi.org/10.1109/TSE.2010.35.
[33] Padovitz A, Loke SW, Zaslavsky A. Multiple-agent perspectives in reasoning about situations for context-aware pervasive computing systems. IEEE Trans Syst Man Cybern Part A: Syst Humans 2008;38(4):729–42. https://doi.org/10.1109/TSMCA.2008.918589.
[34] Jain AK. Data clustering: 50 years beyond k-means. Pattern Recogn Lett 2010;31(8):651–66. https://doi.org/10.1016/j.patrec.2009.09.011. award winning papers from the 19th International Conference on Pattern Recognition (ICPR).
[35] Nowotarski J, Weron R. Recent advances in electricity price forecasting: a review of probabilistic forecasting. Renew Sustain Energy Rev 2018;81:1548–68. https://doi.org/10.1016/j.rser.2017.05.234.
[36] Rahimi-Kian A, Sadeghi B, Thomas R. Q-learning based supplier-agents for electricity markets. IEEE power engineering society general meeting IEEE; 2005. p. 2116–23. https://doi.org/10.1109/PES.2005.1489514.
[37] Sutton R, Barto A. Reinforcement learning: an introduction. MIT Press, Cambridge, Massachusetts 9 (5); 1998: 1054–1054. arXiv:1603.02199. doi:10.1109/TNN.1998.712192.
[38] OMIE. ejecucioncbfom; 2018. http://www.omie.es/aplicaciones/datosftp/datosftp.jsp?path=/ejecucioncbfom/ [accessed March-2018].
[39] Rajavel R, Thangarathinam M. A negotiation framework for the cloud management system using similarity and gale shapely stable matching approach. KSII Transactions on Internet & Information Systems 9(6).
[40] Erev I, Roth AE. Multi-agent learning and the descriptive value of simple models. Artif Intell 2007;171(7):423–8. https://doi.org/10.1016/j.artint.2007.01.001. foundations of Multi-Agent Learning.
[41] Burtini G, Loeppky J, Lawrence R. A survey of online experiment design with the stochastic multi-armed bandit. arXiv preprint arXiv:1510.00757.
[42] Bouneffouf D, Feraud R. Multi-armed bandit problem with known trend. Neurocomputing 2016;205:16–21.
[43] Theodorsson-Norheim E. Kruskal-wallis test: basic computer program to perform nonparametric one-way analysis of variance and multiple comparisons on ranks of several independent samples. Comput Methods Programs Biomed 1986;23(1):57–62. https://doi.org/10.1016/0169-2607(86)90081-7.
[44] Rajavel R, Thangarathanam M. Adaptive probabilistic behavioural learning system for the effective behavioural decision in cloud trading negotiation market. Future Gener Comput Syst 2016;58:29–41. https://doi.org/10.1016/j.future.2015.12.007.
[45] Rajavel R, Thangarathinam M. Adslanf: a negotiation framework for cloud management systems using a bulk negotiation behavioral learning approach. Turk J Electr Eng Comput Sci 2017;25(1):563–90.
[46] Rajavel R, Thangarathinam M. Optimizing negotiation conflict in the cloud service negotiation framework using probabilistic decision making model. Sci World J 2015.