ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τεχνικές Ενισχυτικής Μάθησης σε Πολυπρακτορικά Συστήματα

Γεώργιος Ι. Μπουλούγαρης

Επιβλέποντες: Ευστάθιος Χατζηευθυμιάδης, Επίκουρος Καθηγητής ΕΚΠΑ
Κωνσταντίνος Κολομβάτσος, Υποψήφιος Διδάκτωρ ΕΚΠΑ

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2008

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τεχνικές Ενισχυτικής Μάθησης σε Πολυπρακτορικά Συστήματα

Γεώργιος Ι. Μπουλούγαρης

A.M. : M705

ΕΠΙΒΛΕΠΟΝΤΕΣ:

Ευστάθιος Χατζηευθυμιάδης, Επίκουρος Καθηγητής ΕΚΠΑ
Κωνσταντίνος Κολομβάτσος, Υποψήφιος Διδάκτωρ ΕΚΠΑ

ΠΕΡΙΛΗΨΗ

Οι νοήμονες πράκτορες μπορούν να επιφέρουν πολλά οφέλη όταν χρησιμοποιούνται για να αναπαραστήσουν αγοραστές ή πωλητές σε Αγορές Πληροφορίας. Επιπρόσθετα μπορούν να αποδειχθούν ωφέλιμοι και στους δύο κατά τη διαδικασία συνδιαλλαγής όταν δρουν σαν ενδιάμεσες οντότητες. Σε αυτή την εργασία, εξετάζουμε την περίπτωση που οι αγοραστές χρησιμοποιούν τεχνικές ενισχυτικής μάθησης προκειμένου να επιλέξουν ποια ενδιάμεση οντότητα μπορούν να εμπιστευθούν κατά την αγορά προϊόντων πληροφορίας. Παρουσιάζουμε το μοντέλο μας και περιγράφουμε τη μεθοδολογία με την οποία κατασκευάζεται ο Q-πίνακας.

Οι προσομοιώσεις δείχνουν ότι αυτό το μοντέλο επιτυγχάνει μια σημαντική μείωση του χρόνου κατά τη διαδικασία αγοράς προϊόντων σε συνδυασμό με το γεγονός οι τεχνικές Q-μάθησης επιφέρουν την καλύτερη λύση σε σχέση με τα χαρακτηριστικά των προϊόντων και των ενδιαμέσων.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ : Τεχνητή Νοημοσύνη

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ενισχυτική Μάθηση, Πράκτορας Λογισμικού, Νοήμων Πράκτορας, Πολυπρακτορικά Συστήματα, Αγορά Πληροφορίας

ABSTRACT

Intelligent agents can provide many advantages representing buyers and sellers in Information Markets. Moreover, they can facilitate both of them in the purchase process acting as intermediaries. In this thesis, we examine the case where buyers use reinforcement learning techniques in order to learn on which middle entity they can rely when buying information products. We present our model and describe the methodology for the Q-table creation.

Simulations show that this model indicates a significant time reduction in the purchase process in conjunction with the fact that Q-learning techniques result the best solution according to the characteristics of products and mediators.

SUBJECT AREA: Artificial Intelligence

KEYWORDS: Reinforcement Learning, Software Agent, Intelligent Agent, Multiagent Systems, Information Market

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ

Η γνώριμη οδός συγγραφής ενός προλόγου θα ήταν η αναφορά του πλαισίου μέσα στο οποίο ολοκληρώθηκε αυτή η διπλωματική εργασία και το γνωστικό αντικείμενο το οποίο μελετά. Δεν θα την ακολουθήσουμε…

Ο λόγος…αυτή η εργασία δεν θα είχε ολοκληρωθεί χωρίς την αμέριστη συμπαράσταση των κυρίων Κολομβάτσου Κ. και Χατζηευθυμιάδη Ε. Χωρίς την υπομονή που επέδειξαν στο πρόσωπό μου αυτοί οι κύριοι, δεν θα διαβάζατε τώρα αυτές τις λέξεις.

Σας ευχαριστώ από τα βάθη της καρδιάς μου!!!

ΚΕΦΑΛΑΙΟ 1
ΕΙΣΑΓΩΓΗ

1.1 Διαδίκτυο

Διαδίκτυο (Internet), μια έννοια στην οποία γίνεται καθημερινή αναφορά στις συνομιλίες μας. Αποτελεί ένα αυτόνομο πλέγμα από αλληλοσυνδεδεμένα δίκτυα υπολογιστών που ανταλλάσσουν πληροφορίες. Το μέγεθος του, διογκώνεται μέρα με την ημέρα, ταυτόχρονα όμως, με τη χρησιμότητά του και την αναγκαιότητά του. Η άνευ χρονικών και τοπικών περιορισμών πρόσβαση σε έναν αστείρευτο πλούτο πληροφοριών το έχουν καταστήσει αναπόσπαστο κομμάτι οποιασδήποτε δραστηριότητας του ανθρώπου σήμερα.

Η αναζήτηση πληροφοριών όμως στο διαδίκτυο δεν είναι μια πολύ εύκολη υπόθεση. Κάθε χρήστης έχει αντιμετωπίσει το πρόβλημα της υπερπληροφόρησης. Ο τεράστιος διαθέσιμος όγκος πληροφοριών μπορεί να προκαλέσει μια σύγχυση στον χρήστη κατά τη διάρκεια αναζήτησης μιας πολύτιμης γι' αυτόν πληροφορίας. Είναι απαραίτητο πλέον να 'μάθει' να ψάχνει, να επιλέγει τις πληροφορίες που πραγματικά χρειάζεται και να αξιολογεί όσο υλικό συγκεντρώνει. Μια διαδικασία αρκετά χρονοβόρα και επίπονη όμως.

Μια λύση στο προαναφερθέν πρόβλημα θα μπορούσε να ήταν ο συνδυασμός των τεχνολογιών των Νοημόνων Πρακτόρων (Intelligent Agents) και των Αγορών Πληροφορίας (Information Markets).

1.2 Νοήμονες Πράκτορες

Πράκτορας, μια έννοια για την οποία δυστυχώς δεν υπάρχει ένα κοινά αποδεκτός ορισμός. Κάτω από την ομπρέλα της έννοιας 'πράκτορας' έχει αναπτυχθεί ένα ετερογενές πεδίο έρευνας. Θα μπορούσαμε να πούμε ότι ένας πράκτορας (agent) είναι μια οντότητα που αντιλαμβάνεται το περιβάλλον (environment) μέσα στο οποίο βρίσκεται με τη βοήθεια αισθητήρων (sensors), είναι μέρος του περιβάλλοντος αυτού, κάνει συλλογισμούς για το περιβάλλον και δρα πάνω σε αυτό με τη βοήθεια μηχανισμών δράσης (effectors), για την επίτευξη κάποιων στόχων (goals) (Russell & Norvig, 1995). Πράκτορας εναλλακτικά είναι ένα κομμάτι λογισμικού και/ή υλικού το οποίο είναι ικανό να δρα προκειμένου να εκτελέσει κάποιες εργασίες εκ μέρους των χρηστών του (Nwana, 1996).

Ένας νοήμων πράκτορας, τώρα, σύμφωνα με τον Wooldridge, έχει την ικανότητα εκτέλεσης μιας ευέλικτης αυτόνομης ενέργειας προκειμένου να επιτύχει τους στόχους

σχεδιασμού του. Με τον όρο ευέλικτη ενέργεια υπονοούνται τρία χαρακτηριστικά: αντιδραστικότητα, προνοητικότητα και κοινωνικότητα. Οι ικανότητες δηλαδή του πράκτορα να παρακολουθεί το περιβάλλον του και να αντιδρά σε κάποια χρονικά περιθώρια στις αλλαγές που παρατηρούνται σε αυτό, να εμφανίζει συμπεριφορά που κατευθύνεται από τους στόχους και να αλληλεπιδρά με τους άλλους πράκτορες προκειμένου να ικανοποιήσουν τους στόχους σχεδιασμού τους.

1.3 Αγορές Πληροφορίας

Οι αγορές πληροφορίας θα μπορούσαν να οριστούν σαν τα μέρη που οι συμμετέχοντες προσπαθούν να αγοράσουν ή να πουλήσουν προϊόντα πληροφορίας (π.χ. μουσική, εικόνες, βίντεο, ηλεκτρονικές εφημερίδες, κώδικα λογισμικού). Χαρακτηρίζονται για τη δυναμική τους φύση καθώς συνεχώς αλλάζει ο αριθμός και η συμπεριφορά των συμμετεχόντων. Καθώς λειτουργεί η αγορά παράγεται οικονομικό όφελος για τους πωλητές, τους αγοραστές και για τους ενδιαμέσους.

- πωλητές – έχουν στην κατοχή τους προϊόντα τα οποία θέλουν να πουλήσουν στους ενδιαφερόμενους με την πιο επικερδή τιμή
- αγοραστές – αναζητούν προϊόντα που ικανοποιούν τις απαιτήσεις τους και είναι πρόθυμοι να πληρώσουν γι’ αυτά
- ενδιάμεσοι – βοηθούν στο ‘ζευγάρωμα’ των πωλητών με τους αγοραστές, παρέχουν πληροφορίες των προϊόντων στους αγοραστές, παρέχουν πληροφορίες marketing στους πωλητές, συνδυάζουν προϊόντα πληροφορίας, μεταχειρίζονται τις μεταφορές και πληρωμές, παρέχουν το απαραίτητο επίπεδο εμπιστοσύνης και διασφαλίζουν την ακεραιότητα της αγοράς. Αυτοί μπορεί να είναι matchmakers, οι οποίοι μεταχειρίζονται διαφημίσεις προϊόντων, blackboard agents, οι οποίοι συλλέγουν εκδηλώσεις ενδιαφέροντος και brokers οι οποίοι κάνουν και τα δύο (Decker κ.ά., 1997).

Ο συνδυασμός των τεχνολογιών των νοημόνων πρακτόρων και των εικονικών αγορών φάνταζε σαν μια λύση η οποία θα μπορούσε αποτελεσματικά να αντιμετωπίσει το πρόβλημα της αναζήτησης και απόκτησης πληροφοριών. Είναι δυνατό να αναθέσουμε την εργασία αναζήτησης πληροφοριών σε κάποιον νοήμονα πράκτορα ο οποίος θα μετέχει σε κάποια αγορά πληροφορίας. Στον πράκτορα μπορούμε να μεταδώσουμε τις προτιμήσεις και τις ανάγκες μας και αυτός να λειτουργήσει αυτόνομα προς όφελός μας.

Στην εργασία μας θα επικεντρωθούμε στην αλληλεπίδραση των αγοραστών και των ενδιάμεσων οντοτήτων που λειτουργούν ως broker και προτείνουν προϊόντα

πληροφορίας. Οι πράκτορες-αγοραστές θα επιλέξουν τον broker που θα προτείνει το προϊόν που ανταποκρίνεται καλύτερα στις προτιμήσεις του χρήστη. Η αποτύπωση των προτιμήσεων καθώς και η επιλογή του broker με τον οποίο θα συνδιαλλαγεί ένας αγοραστής για την αγορά ενός προϊόντος πληροφορίας γίνεται με τη χρήση των τεχνικών της ενισχυτικής μάθησης και πιο συγκεκριμένα του αλγορίθμου της Q-μάθησης.

1.4 Ενισχυτική Μάθηση

Ως Ενισχυτική Μάθηση (reinforcement learning) αναφερόμαστε στο πρόβλημα που αντιμετωπίζει ένας πράκτορας ο οποίος διαμορφώνει τη συμπεριφορά του εντός ενός δυναμικού περιβάλλοντος διαμέσου αλληλεπιδράσεων της μορφής προσπάθεια–και–λάθος (trial and error) (Kaelbling κ.ά. 1996, Sutton & Barto, 1998).

Το μοντέλο αλληλεπίδρασης που μελετάται έχει ως εξής: σε κάθε διακριτό σημείο στο χρόνο ο πράκτορας μελετά την τρέχουσα κατάσταση (state) του περιβάλλοντος (environment) και επιλέγει την πραγματοποίηση μιας ενέργειας (action). Σαν αποτέλεσμα, το περιβάλλον μεταβαίνει σε μια νέα κατάσταση και ο πράκτορας λαμβάνει μια τιμή ανταμοιβής (reward). Αυτό το σήμα αποτελεί ένα μέτρο της ποιότητας των ενεργειών του πράκτορα, όπως καθορίζονται από το περιβάλλον.

Οι τεχνικές μάθησης που βασίζονται σε ένα τέτοιο μοντέλο είναι πολύ ελκυστικές. Κι αυτό διότι αν οι πράκτορες υπόκεινται σε ενισχυτική μάθηση, οι σχεδιαστές καλούνται μόνο να παράγουν την τιμή της ανταμοιβής.

Μία τεχνική ενισχυτικής μάθησης αποτελεί ο αλγόριθμος Q-μάθηση. Ο αλγόριθμος αυτός δεν απαιτεί ένα ακριβές μοντέλο του κόσμου, στοιχείο πολύ σημαντικό, ιδίως αν σκεφτούμε το μέγεθος και το δυναμικό χαρακτήρα μιας αγοράς πληροφοριών.

1.5 Στόχοι Εργασίας

Συνοπτικά οι στόχοι της εργασίας ήταν:

1. Η διερεύνηση και μελέτη του επιστημονικού πεδίου όσον αφορά στις τεχνικές ενισχυτικής μάθησης (reinforcement learning) στα πολυπρακτορικά συστήματα.
2. Η καταγραφή των χαρακτηριστικών του κάθε μοντέλου και ποιοτική σύγκρισή τους.
3. Η υλοποίηση έξυπνων πρακτόρων οι οποίοι χρησιμοποιούν τις πιο κατάλληλες από τις παραπάνω μεθόδους με βάση ένα συγκεκριμένο σενάριο. Το σενάριο αυτό περιλαμβάνει πράκτορες οι οποίοι ενεργούν σε ένα σύστημα κάτω από τις

ιδιότητες του αγοραστή, του πωλητή ή του ενδιάμεσου σε μια διαδικασία ανταλλαγής προϊόντων πληροφορίας.

4. Η έρευνα για την υλοποίηση αυτόματων μεθόδων εξάσκησης (training) των μηχανισμών μάθησης με βάση την μεγιστοποίηση του μελλοντικού αναμενόμενου οφέλους από το σύστημα.
5. Η έρευνα για την μοντελοποίηση των υπολοίπων οντοτήτων που ενεργούν στο περιβάλλον ενός πράκτορα σε σχέση με τα ιδιαίτερα χαρακτηριστικά τους και με την διαχείριση της αβεβαιότητας για αυτούς.

1.6 Οργάνωση Εργασίας

Η εργασία έχει οργανωθεί σε 6 κεφάλαια: Το Κεφάλαιο 2 αναφέρεται στις έννοιες Πράκτορας Λογισμικού, Πολυπρακτορικά Συστήματα και Αγορές Πληροφορίας. Πιο συγκεκριμένα, παρουσιάζονται τα χαρακτηριστικά των πρακτόρων λογισμικού και αναφέρεται κάποιο τυπολόγιο με τις διάφορες κατηγορίες αυτών. Στα πολυπρακτορικά συστήματα, αφού γίνει αναφορά στα χαρακτηριστικά τους, δίνεται ιδιαίτερη έμφαση στα προβλήματα της επικοινωνίας και διαπραγμάτευσης που συναντώνται μεταξύ των πρακτόρων που συμμετέχουν σε αυτά. Τέλος γίνεται μια παρουσίαση των χαρακτηριστικών, των μετεχόντων και των λειτουργιών μιας αγοράς πληροφορίας.

Το Κεφάλαιο 3 περιγράφει λεπτομερώς τις τεχνικές της Ενισχυτικής Μάθησης και τα χαρακτηριστικά αυτών. Γίνεται αναφορά σε αλγορίθμους που αφορούν έναν πράκτορα καθώς και σε αλγορίθμους που εφαρμόζονται σε πολυπρακτορικά συστήματα.

Το Κεφάλαιο 4 αποτελεί παρουσίαση του μοντέλου της αγοράς πληροφορίας που κατασκευάσαμε βασιζόμενοι στις ιδέες της ενισχυτικής μάθησης και των νοημόνων πρακτόρων.

Στο Κεφάλαιο 5, στη συνέχεια, γίνεται παρουσίαση των αποτελεσμάτων της σύγκρισης του μοντέλου μας με κάποιο μοντέλο πρότυπο με βάση τον χρόνο με τον οποίο ολοκληρώνεται η διαδικασία απόκτησης ενός προϊόντος πληροφορίας.

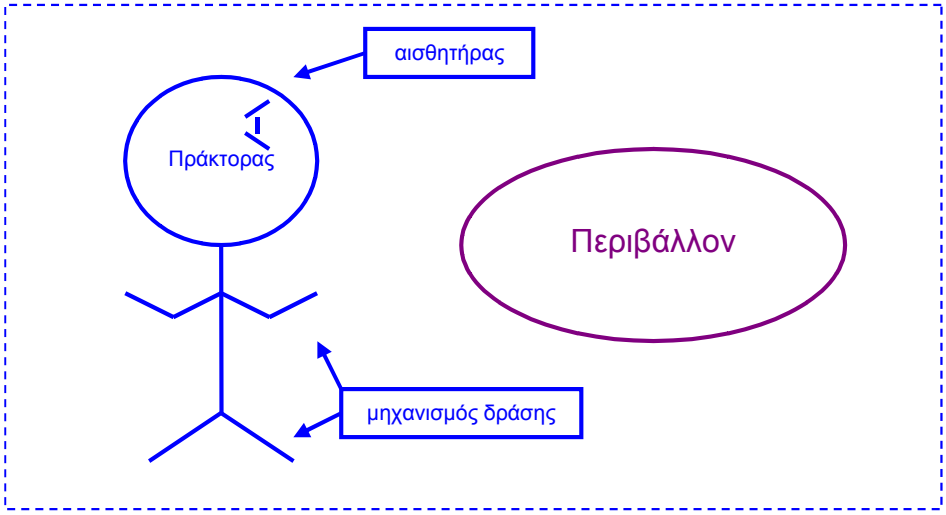
Στο Κεφάλαιο 6, τέλος, καταγράφονται κάποια ενδιαφέροντα συμπεράσματα για την εργασία μας, ενώ δίδονται και κάποιες μελλοντικές προεκτάσεις.

ΚΕΦΑΛΑΙΟ 2
ΠΡΑΚΤΟΡΕΣ ΛΟΓΙΣΜΙΚΟΥ – ΠΟΛΥΠΡΑΚΤΟΡΙΚΑ ΣΥΣΤΗΜΑΤΑ – ΑΓΟΡΕΣ ΠΛΗΡΟΦΟΡΙΑΣ

2.1 Πράκτορες Λογισμικού

Ο πιο προφανής τρόπος να ξεκινήσουμε αυτό το κεφάλαιο είναι να παραβάλλουμε τον ορισμό της έννοιας ‘πράκτορας’. Όσο απλό κι αν φαίνεται, δεν είναι. Ο λόγος είναι ότι δεν υπάρχει ένας κοινά αποδεκτός ορισμός της έννοιας ‘πράκτορας’. Κάτω από την ομπρέλα της έννοιας ‘πράκτορας’ έχει αναπτυχθεί ένα ετερογενές πεδίο έρευνας. Πολλά από τα χαρακτηριστικά που έχουν αποδοθεί στους πράκτορες έχουν διαφορετικό συντελεστή βαρύτητας για τους διάφορους επιστημονικούς τομείς.

Θα μπορούσαμε να πούμε ότι ένας *πράκτορας (agent)* είναι μια οντότητα που αντιλαμβάνεται το *περιβάλλον (environment)* μέσα στο οποίο βρίσκεται με τη βοήθεια *αισθητήρων (sensors)*, είναι μέρος του περιβάλλοντος αυτού, κάνει συλλογισμούς για το περιβάλλον και δρα πάνω σε αυτό με τη βοήθεια *μηχανισμών δράσης* (effectors), για την επίτευξη κάποιων *στόχων (goals)* (Russell & Norvig, 1995).



Εικόνα 1. Αναπαράσταση Πράκτορα

Αλλιώς, πράκτορας είναι ένα κομμάτι λογισμικού και/ή υλικού το οποίο είναι ικανό να δρα προκειμένου να εκτελέσει κάποιες εργασίες εκ μέρους των χρηστών του (Nwana, 1996).

Η τεχνολογία των πρακτόρων αποκτά συνεχώς μεγαλύτερο ενδιαφέρον για τους ερευνητές καθώς εκτιμάται ότι θα αλλάξει τον τρόπο που διασυνδέονται οι χρήστες με το λογισμικό. Ο χρήστης πλέον δεν θα επικοινωνεί απευθείας με κάποια εφαρμογή

αλλά θα τον αντικαθιστά κάποιος πράκτορας ο οποίος θα τον διευκολύνει σε κάποιες χρονοβόρες και επίπονες διαδικασίες, διαδικασίες ρουτίνας ή διαδικασίες που χρειάζονται κάποια ικανότητα που ο χρήστης δεν έχει ακόμα αποκτήσει. Η αδυναμία διατύπωσης ενός ορισμού κοινής αποδοχής σε καμία περίπτωση δεν αποτελεί εμπόδιο για τη διάδοση και χρήση της τεχνολογίας των πρακτόρων σε πολλές εφαρμογές, καθώς επίσης και της έρευνας που διεξάγεται ολοένα και από μεγαλύτερο αριθμό επιστήμονων.

Χαρακτηριστικά Πρακτόρων

Σε τι όμως διαφέρει ένας πράκτορας από τα συμβατικά προγράμματα; Και σε αυτή την περίπτωση η απάντηση δεν είναι εύκολη, καθώς τα όρια είναι δυσδιάκριτα και ο όρος πράκτορας χαρακτηρίζει πολλά συστήματα με διαφορές τόσο στην πολυπλοκότητα όσο και στα επιμέρους χαρακτηριστικά.

Κάποια χαρακτηριστικά που διαφοροποιούν τους πράκτορες σε σχέση με τα συμβατικά προγράμματα παρατίθενται από τους Wooldridge και Jennings (1995). Συγκεκριμένα, αυτά είναι:

- ✓ *αυτονομία (autonomy)*: η ικανότητα των πρακτόρων να δρουν χωρίς την επέμβαση των χρηστών η άλλων συστημάτων. Έχουν πλήρη έλεγχο της εσωτερικής τους κατάστασης και της συμπεριφοράς.
- ✓ *αντιδραστικότητα (reactiveness):* η ικανότητα των πρακτόρων να αντιλαμβάνονται το περιβάλλον τους και να αντιδρούν μέσα σε συγκεκριμένα χρονικά πλαίσια στις αλλαγές που επέρχονται σε αυτό
- ✓ *προνοητικότητα (pro-activeness)*: η ικανότητα των πρακτόρων να επιδεικνύουν συμπεριφορά που κατευθύνεται από τους στόχους, λαμβάνοντας ουσιαστικά κάποια πρωτοβουλία ανάλογα με τις συνθήκες οι οποίες εμφανίζονται στο περιβάλλον
- ✓ *κοινωνικότητα (social ability)*: η ικανότητα των πρακτόρων να αλληλεπιδρούν με άλλους πράκτορες (και πιθανόν με χρήστες) για την επίτευξη των στόχων

Επιπρόσθετα στα παραπάνω, μερικές φορές προσδίδονται και κάποια δευτερεύοντα χαρακτηριστικά, όπως:

- ✓ *κινητικότητα (mobility)*: η ικανότητα των πρακτόρων να μην παραμένουν στατικοί, αλλά να κινούνται σε ένα υπολογιστικό περιβάλλον

- ✓ *προσαρμοστικότητα (adaptivity)*: η ικανότητα των πρακτόρων να προσαρμόζονται διαρκώς στο περιβάλλον τους ή τις απαιτήσεις του χρήστη, έχουν δηλαδή ικανότητα για μάθηση
- ✓ *ειλικρίνεια (veracity)*: οι πράκτορες δεν δίνουν εσκεμμένα λάθος πληροφορίες
- ✓ *αγαθή προαίρεση (benevolence)*: η προσπάθεια των πρακτόρων να επιτύχουν πάντα τους στόχους που τους έχουν ανατεθεί
- ✓ *λογικότητα (rationality)*: η ικανότητα των πρακτόρων να δρουν για την επίτευξη των στόχων που τους έχουν ανατεθεί.

Χρησιμοποιώντας αυτά τα χαρακτηριστικά θα προσπαθήσουμε στη συνέχεια με λίγα λόγια να αναδείξουμε τις διαφορές των πρακτόρων με τα συμβατικά προγράμματα και πιο συγκεκριμένα των πρακτόρων με τα *αντικείμενα* και τα *έμπειρα συστημάτων.*

Η παραδοσιακή θεώρηση των αντικειμένων σε σχέση με αυτή των πρακτόρων διαφοροποιούνται κυρίως στο ότι:

- ✓ οι πράκτορες χαρακτηρίζονται για τον πιο μεγάλο βαθμό αυτονομίας από ότι τα αντικείμενα. Αποφασίζουν οι ίδιοι εάν θα ζητήσουν κάτι από έναν άλλο πράκτορα.
- ✓ οι πράκτορες είναι ικανοί να επιδεικνύουν πιο εύκαμπτη (αντιδραστική, προνοητική, κοινωνική) συμπεριφορά, κάτι που δεν ισχύει στην περίπτωση των αντικειμένων.
- ✓ ένα πολυπρακτορικό σύστημα είναι εκ φύσεως πολυνηματικό, καθώς απαιτείται ένα νήμα ελέγχου για κάθε έναν πράκτορα.

Τα έμπειρα συστήματα έχουν κάποιες σημαντικές διαφορές σε σχέση με τους πράκτορες. Ειδικότερα:

- ✓ τα έμπειρα συστήματα δεν αλληλεπιδρούν απευθείας με το περιβάλλον. Δεν προσλαμβάνουν την απαραίτητη πληροφορία μέσω αισθητήρων αλλά μέσω κάποιων χρηστών που τα τροφοδοτούν κατάλληλα.
- ✓ τα έμπειρα συστήματα δεν δρουν σε κάποιο περιβάλλον, απλώς παρέχουν κάποια ανατροφοδότηση ή συμβουλή σε κάποια τρίτη οντότητα.
- ✓ τα έμπειρα συστήματα δεν απαιτείται να έχουν την ικανότητα της συνεργασίας με άλλους πράκτορες.

Περιβάλλοντα

Σε κάθε περίπτωση οι πράκτορες εκτελούν κάποιες ενέργειες εντός του περιβάλλοντος, το οποίο στη συνέχεια παρέχει κάποια ερεθίσματα στον πράκτορα. Σε αυτή την ενότητα θα περιγράψουμε τους διάφορους τύπους των περιβαλλόντων. Μπορούν να κατηγοριοποιηθούν ως (Russell & Norvig, 1995):

Προσβάσιμα – Μη προσβάσιμα (Accessible - Inaccessible)

Εάν ο πράκτορας μέσω των αισθητήρων έχει διαθέσιμη πλήρη, ακριβή και ανανεωμένη πληροφορία της κατάστασης του περιβάλλοντος, λέμε ότι ο περιβάλλον είναι προσβάσιμο στον πράκτορα. Εάν ο πράκτορας τοποθετείται σε ένα προσβάσιμο περιβάλλον δεν απαιτείται να διατηρεί καμιά εσωτερική κατάσταση.

Αιτιοκρατικά – Μη αιτιοκρατικά (Deterministic – Nondeterministic)

Εάν η επόμενη κατάσταση του περιβάλλοντος καθορίζεται αποκλειστικά από την τρέχουσα κατάσταση και τις ενέργειες που εκτελεί ο πράκτορας λέμε ότι το περιβάλλον είναι αιτιοκρατικό. Μία συγκεκριμένη ενέργεια επιφέρει συγκεκριμένα αποτελέσματα.

Επεισοδιακά – Μη επεισοδιακά (Episodic - Nonepisodic)

Σε ένα επεισοδιακό περιβάλλον, η εμπειρία του πράκτορα χωρίζεται σε διακριτά και ανεξάρτητα επεισόδια. Κάθε επεισόδιο αποτελείται από τα ερεθίσματα που λαμβάνει ο πράκτορας και την ενέργεια που αυτοί εκτελούν. Τα επεισοδιακά περιβάλλοντα είναι απλούστερα διότι ο πράκτορας δεν είναι απαραίτητο να σκέπτεται μακροπρόθεσμα.

Δυναμικά – Στατικά (Dynamic – Static)

Εάν το περιβάλλον επιδέχεται αλλαγές καθ’ όσο ο πράκτορας μελετά την ενέργεια που θα εκτελέσει, λέμε ότι το περιβάλλον είναι δυναμικό για τον πράκτορα, διαφορετικά είναι στατικό. Τα στατικά περιβάλλοντα είναι πιο απλά διότι ο πράκτορας δεν απαιτείται να παρατηρεί το περιβάλλον μέχρι να αποφασίσει ποια ενέργεια θα εκτελέσει.

Διακριτά – Συνεχή (Discrete - Continuous)

Εάν υπάρχει ένας πεπερασμένος αριθμός ενεργειών και ερεθισμάτων στο μηχανισμό αντίληψης του πράκτορα λέμε ότι το περιβάλλον είναι διακριτό.

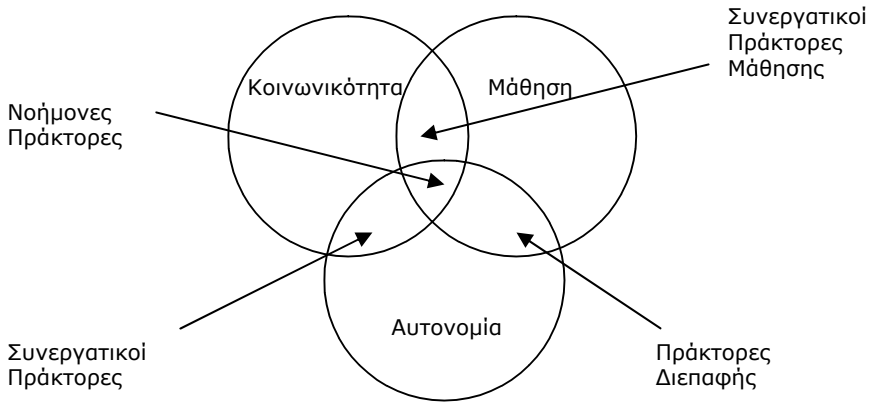
Διαφορετικοί τύποι περιβαλλόντων απαιτούν διαφορετικά προγράμματα για τους πράκτορες προκειμένου να λειτουργούν αποδοτικά και αποτελεσματικά σε αυτά. Προκύπτει εύκολα ότι η πιο δύσκολη περίπτωση είναι ένα μη προσβάσιμο, μη

επεισοδιακό, δυναμικό και συνεχές περιβάλλον και ότι οι περισσότερες πραγματικές καταστάσεις είναι μη αιτιοκρατικά.

Τυπολόγιο Πρακτόρων

Στην ενότητα αυτή θα προσπαθήσουμε να κατηγοριοποιήσουμε τους υπάρχοντες πράκτορες σε κάποιες συγκεκριμένες κλάσεις. Θα ακολουθήσουμε το τυπολόγιο που πρότεινε η Nwana το 1996. Υπάρχουν διάφορες παράμετροι τις οποίες θα λάβουμε υπόψη προκειμένου να κατηγοριοποιήσουμε τους υπάρχοντες πράκτορες: Αρχικά οι πράκτορες μπορούν να κατηγοριοποιηθούν σχετικά με το αν μπορούν να κινούνται εντός ενός δικτύου. Δημιουργούνται έτσι, οι κλάσεις των *στατικών* και των *κινητών* πρακτόρων. Στη συνέχεια μπορούμε να τους χαρακτηρίσουμε είτε σαν *αντιδραστικούς* είτε σαν *συλλογιστικούς* (deliberative). Οι συλλογιστικοί πράκτορες κατέχουν ένα εσωτερικό συμβολικό μοντέλο και αποσκοπούν στον προγραμματισμό και στη διαπραγμάτευση με άλλους πράκτορες για την επίτευξη των στόχων τους. Οι αντιδραστικοί πράκτορες, αντίθετα, δεν κατέχουν κανένα εσωτερικό, συμβολικό μοντέλο του περιβάλλοντός τους και δρουν βάσει ενός μοντέλου ερεθίσματος – απάντησης εξαρτώμενοι από την τρέχουσα κατάσταση του περιβάλλοντος στο οποίο έχουν τοποθετηθεί.

Επιπλέον, μπορούμε να τους εντάξουμε σε κατηγορίες με το αν επιδεικνύουν ή όχι κάποια χαρακτηριστικά που αναφέραμε παραπάνω. Αυτά είναι η αυτονομία, η κοινωνικότητα και η μάθηση. Στηριζόμενοι σε αυτά τα χαρακτηριστικά μπορούμε να εξάγουμε 4 διαφορετικούς τύπους πρακτόρων: τους *συνεργατικούς πράκτορες* (collaborative agents), τους *συνεργατικούς πράκτορες μάθησης* (collaborative learning agents), τους *πράκτορες διεπαφής* (interface agents) και τους *νοήμονες πράκτορες* (smart – intelligent agents).



Εικόνα 2. Διάκριση πρακτόρων (Nwana, 1996)

Στο (Nwana, 1996), τονίζεται ότι αυτός ο διαχωρισμός δεν πρέπει να θεωρείται απόλυτος κι ο λόγος είναι για παράδειγμα ότι ένας συνεργατικός πράκτορας μπορεί να επιδεικνύει κυρίως τα χαρακτηριστικά της αυτονομίας και της κοινωνικότητας, αλλά αυτό δεν σημαίνει ότι δεν μαθαίνει ποτέ.

Ένας άλλος τρόπος να κατηγοριοποιήσουμε τους πράκτορες είναι ο ρόλος τον οποίο καλούνται να παίξουν. Ένα χαρακτηριστικό παράδειγμα αποτελούν οι *πράκτορες πληροφορίας (information agents)*. Αυτοί οι πράκτορες επιβαρύνονται με την διαχείριση των υπέρογκων διαθέσιμων πληροφοριών τις οποίες έχει διαθέσιμες ένας χρήστης σε δίκτυα ευρείας περιοχής όπως το διαδίκτυο.

Τέλος μπορούμε να συμπεριλάβουμε στο τυπολόγιο και τη κατηγορία των *υβριδικών πρακτόρων (hybrid agents)* οι οποίοι συνδυάζουν δύο ή περισσότερες φιλοσοφίες πρακτόρων σε έναν απλό πράκτορα.

Προκύπτει από τα παραπάνω ότι η κατηγοριοποίηση των πρακτόρων απαιτεί πολλές διαστάσεις αλλά μια τέτοια απόπειρα δεν θα ήταν απολύτως ακριβής Γι’ αυτό το λόγο προτείνεται τελικά μια λίστα με επτά τύπους πρακτόρων. Αυτοί είναι οι:

- συνεργατικοί πράκτορες
- πράκτορες διεπαφής
- κινητοί πράκτορες
- πράκτορες πληροφορίας
- αντιδραστικοί πράκτορες
- υβριδικοί πράκτορες
- νοήμονες πράκτορες

Σε κάποιες εφαρμογές συνδυάζουμε πράκτορες από δύο ή περισσότερες κατηγορίες, και αναφερόμαστε σε αυτές ως *συστήματα ετερογενών πρακτόρων (heterogeneous agent systems).*

Για να είναι πιο πλήρες το τυπολόγιο που προτείνεται, είναι σημαντικό να αναφέρουμε ότι οι πράκτορες μπορούν επιπλέον να κατηγοριοποιηθούν σύμφωνα και με τα άλλα δευτερεύοντα χαρακτηριστικά που αναφέρθηκαν παραπάνω. Οι πράκτορες για παράδειγμα δεν είναι απαραίτητο να χαρακτηρίζονται για τις αγαθές προαιρέσεις μεταξύ τους. Οι πράκτορες είναι πιθανό να συναγωνίζονται μεταξύ τους για να επιτύχουν τους στόχους τους και σε κάποιες περιπτώσεις να είναι αρκετά ανταγωνιστικοί. Στη συνέχεια

θα προσπαθήσουμε να περιγράψουμε με συντομία τα χαρακτηριστικά καθεμιάς κατηγορίας πρακτόρων.

Συνεργατικοί Πράκτορες

Οι συνεργατικοί πράκτορες χαρακτηρίζονται κυρίως για την αυτονομία και την κοινωνικότητα με άλλους πράκτορες ώστε να εκτελέσουν κάποιες εργασίες εκ μέρους των ιδιοκτητών τους. Είναι πιθανό να μαθαίνουν αλλά αυτό δεν είναι το κύριο χαρακτηριστικό τους. Για να έχουμε μια καλά συντονισμένη και οργανωμένη κοινότητα πρακτόρων είναι απαραίτητο να διαπραγματεύονται μεταξύ τους προκειμένου να επιτύχουν κάποιες κοινά αποδεκτές συμφωνίες σε κάποια ζητήματα.

Κάποιοι ερευνητές αποδίδουν σε αυτούς τους πράκτορες κάποια επιπρόσθετα χαρακτηριστικά (π.χ. πεποιθήσεις (beliefs), επιθυμίες (desires) και προθέσεις (intentions) τα οποία τους εξειδικεύουν σε συνεργατικούς πράκτορες τύπου BDI.

Ο λόγος που καταφεύγουμε σε αυτόν τον τύπο των πρακτόρων είναι η επίλυση προβλημάτων τα οποία είναι πέρα των δυνατοτήτων του ενός πράκτορα.

Ένα χαρακτηριστικό παράδειγμα αυτής της κατηγορίας πρακτόρων αποτελεί η εφαρμογή Pleiades που σχεδιάστηκε από τον Tom Mitchell και την Katia Sycara (URL1). Οι Pleiades είναι μια κατανεμημένη αρχιτεκτονική που βασίζεται σε πράκτορες η οποία έχει δύο επίπεδα αφαίρεσης. Το πρώτο επίπεδο περιλαμβάνει συνεργατικούς πράκτορες οι οποίοι είναι συγκεκριμένοι για κάποια εργασία και το δεύτερο επίπεδο περιλαμβάνει συνεργατικούς πράκτορες οι οποίοι έχουν συγκεκριμένη πληροφορία. Οι πράκτορες του πρώτου επιπέδου συντονίζονται και κατασκευάζουν πλάνα σχετικά με την εξειδίκευσή τους. Συνεργάζονται μεταξύ τους προκειμένου να αποφευχθούν οι συγκρούσεις και να συνδυάσουν την διαθέσιμη πληροφορία. Οι πράκτορες του δευτέρου επιπέδου συνεργάζονται μεταξύ τους για να τροφοδοτήσουν με την απαραίτητη πληροφορία τους πράκτορες του πρώτου επιπέδου.

Πράκτορες Διεπαφής

Οι πράκτορες διεπαφής έχουν σαν κύρια χαρακτηριστικά την αυτονομία και την μάθηση για να εκτελέσουν κάποιες εργασίες για τους ιδιοκτήτες τους. Ένα στοιχείο κλειδί το οποίο παρουσιάστηκε από την Pattie Maes για τους πράκτορες αυτής της κατηγορίας είναι ότι αποτελούν προσωπικούς βοηθούς οι οποίοι συνεργάζονται με τον χρήστη στο ίδιο περιβάλλον εργασίας. Είναι αξιοσημείωτη η διαφορά της συνεργασίας τους με τον χρήστη σε σχέση με τη συνεργασία τους με άλλους πράκτορες όπως στην περίπτωση των συνεργατικών πρακτόρων. Η συνεργασία με ένα χρήστη δεν απαιτεί την ύπαρξη

μιας κοινά αποδεκτής γλώσσας επικοινωνίας μεταξύ των πρακτόρων, που είναι απαραίτητη στη συνεργασία με άλλους πράκτορες.

Οι πράκτορες διεπαφής παρέχουν βοήθεια στον χρήστη μαθαίνοντας να κάνει χρήση μια συγκεκριμένη εφαρμογή όπως τα λογιστικά φύλλα. Όσον αφορά τη μάθηση, οι πράκτορες διεπαφής βελτιώνουν τη βοήθεια που παρέχουν στο χρήστη με τέσσερις τρόπους όπως τους παρουσιάζει η Maes:

- ✓ με το να παρατηρεί και να μιμείται το χρήστη
- ✓ με το να λαμβάνει θετική και αρνητική ανατροφοδότηση από το χρήστη
- ✓ με το να δέχεται ρητές οδηγίες από το χρήστη
- ✓ με το να ρωτά άλλους πράκτορες για συμβουλές

Η συνεργασία με τους άλλους πράκτορες παραμένει στο επίπεδο της παροχής συμβουλών κι όχι στην διαδικασία συμφωνιών με διαπραγμάτευση όπως στους συνεργατικούς πράκτορες. Οι πράκτορες διεπαφής θα μπορούσαν να χρησιμοποιηθούν επομένως είτε για την εκπαίδευση άπειρων χρηστών πάνω σε μια συγκεκριμένη εφαρμογή, είτε για την πλήρη εκτέλεση κάποιων εργασιών που τους αναθέτουν οι χρήστες τους.

Τα οφέλη κατασκευής πρακτόρων διεπαφής είναι πολλαπλά. Αρχικά, απαιτούν πολύ λιγότερο κόπο και χρόνο από τον σχεδιαστή της εφαρμογής σε σχέση με άλλες εφαρμογές πρακτόρων. Επίσης, ο πράκτορας μπορεί με την πάροδο του χρόνου να προσαρμόζεται στις προτιμήσεις και τις συνήθειες των χρηστών τους και τέλος είναι πιθανό να διαμοιραστεί σε μια κοινότητα χρηστών το know – how της εφαρμογής.

Παραδείγματα εφαρμογών πρακτόρων διεπαφής έχουν αναπτυχθεί με μεγάλη ποικιλία ρόλων. Έχουμε συναντήσει βοηθό ο οποίο αναλαμβάνει τον προγραμματισμό των συναντήσεων ενός χρήστη (Calendar Agent – Kozierok & Maes, 1993), οδηγό ο οποίος βοηθά την αναζήτηση στο διαδίκτυο (Letizia – Liebemann, 1995), βοηθούς σε θέματα μνήμης όπως ο Remembrance Agent (Rhodes & Starner, 1996) ο οποίος κατά τη συγγραφή ενός e-mail μας παρουσίαζε 5 e-mails μετά από αναζήτηση με βάσει μια λέξη κλειδί που είχαν σχετικό περιεχόμενο και τέλος βοηθούς που είτε φιλτράρουν και επιλέγουν άρθρα (NewT – Sheth & Maes, 1993), είτε εκτελούν αγοραπωλησίες εκ μέρους του χρήστη τους (Kasbah – Chavez & Maes, 1996), είτε αναζητούν γι' αυτούς χρήστες που έχουν τις ίδιες καλλιτεχνικές προτιμήσεις με σκοπό να τους προτείνουν κάποιες ενδιαφέροντες λύσεις (Ringo/HOMR – Shardanand & Maes, 1995).

Κινητοί Πράκτορες

Οι κινητοί πράκτορες είναι διεργασίες οι οποίες έχουν την ικανότητα να μετακινούνται σε δίκτυα ευρείας περιοχής. Οι πράκτορες αυτοί αλληλεπιδρούν με απομακρυσμένους hosts, μεταφέρουν πληροφορίες εκ μέρους του χρήστη τους και επιστρέφουν έχοντας εκτελέσει τις εργασίες που τους έχουν αναθέσει. Είναι σημαντικό να αναφέρουμε ότι η κινητικότητα δεν αποτελεί ικανή και αναγκαία συνθήκη για να χαρακτηρισθεί κάποιος πράκτορας. Οι κινητοί πράκτορες είναι πράκτορες διότι έχουν την απαραίτητη αυτονομία και μπορούν να συνεργαστούν με άλλους συνεργατικούς πράκτορες.

Ο λόγος για τον οποίο καταφεύγουμε στην κατασκευή κινητών πρακτόρων είναι ότι παρουσιάζουν κάποια πρακτικά οφέλη τα οποία δεν τα έχουμε στην περίπτωση των στατικών. Πιο συγκεκριμένα, με τους κινητούς πράκτορες έχουμε μειωμένα υπολογιστικά κόστη επικοινωνίας, λιγότερους πόρους τοπικά, ευκολότερο συντονισμό, μεγαλύτερη αξιοπιστία και η δυνατότητα ασύγχρονης εκτέλεσης του κινητού πράκτορα και των υπολοίπων εφαρμογών του χρήστη. Όμως μια σημαντική παράμετρος που πρέπει να απασχολήσει τους ερευνητές στο μέλλον είναι αυτή της ασφάλειας τόσο του ίδιου του πράκτορα όσο και του συστήματος που θα τον φιλοξενήσει. Θα πρέπει να εξασφαλιστεί ότι ο κώδικας του πράκτορα δεν θα αλλοιωθεί τόσο κατά την παραμονή του σε κάποιο ξένο σύστημα όσο και κατά την μεταφορά του στο δίκτυο. Επίσης το σύστημα το οποίο δέχεται τον κινητό πράκτορα θα πρέπει να έχει εγγυήσεις ότι δεν θα δεχθεί κακόβουλες επιθέσεις από αυτόν.

Πράκτορες Πληροφορίας

Οι πράκτορες πληροφορίας γεννήθηκαν από την ανάγκη μας να υπάρξει κάποιος βοηθός μας κατά τη συλλογή, μεταχείριση και συσχέτιση πληροφοριών από πολλές κατανεμημένες πηγές. Για να μην υπάρξει σύγχυση για την ‘ορθότητα’ ύπαρξης αυτής της κατηγορίας πρακτόρων καθόσον οι συνεργατικοί πράκτορες ή οι πράκτορες διεπαφής ασχολούνται με παρεμφερείς περιοχές, η συγγραφέας Nwana τονίζει ότι αυτή η διαφοροποίηση οφείλεται σε διαφορετικά κριτήρια. Οι πράκτορες πληροφορίας ορίστηκαν από το τι αυτοί κάνουν ενώ οι άλλες κατηγορίες από το τι αυτοί είναι.

Οι πράκτορες πληροφορίας έχουν διάφορα χαρακτηριστικά, μπορεί να είναι στατικοί ή κινητοί, μη-συνεργατικοί ή κοινωνικοί, και μπορεί να έχουν ή όχι την ικανότητα μάθησης.

Οι Etzioni & Weld (1994) περιέγραψαν έναν πρωτότυπο πράκτορα πληροφορίας τον internet softbot. Ο χρήστης έχει τη δυνατότητα να υποβάλει ένα υψηλού επιπέδου ερώτημα και στη συνέχεια ο πράκτορας μπορεί να χρησιμοποιήσει τη γνώση

αναζήτησης και συμπερασμού που διαθέτει προκειμένου να ικανοποιήσει αυτό το ερώτημα στο διαδίκτυο.

Μια άλλη προσπάθεια πράκτορα πληροφορίας αποτελεί ο πράκτορας Jasper (Joint Access to Stored Pages with Easy Retrieval) που αναπτύχθηκε από τους Davies & Weeks (1995). Οι πράκτορες Jasper λειτουργούν εκ μέρους ενός χρήστη ή μιας κοινότητας χρηστών και έχουν την δυνατότητα να αποθηκεύουν, ανακτούν, συγκεντρώνουν και να ενημερώνουν άλλους πράκτορες με πληροφορίες χρήσιμες γι' αυτούς με πηγή το διαδίκτυο. Καθώς ένας χρήστης εργάζεται με τον πράκτορά του, αυτός κατασκευάζει ένα δυναμικό προφίλ των προτιμήσεών του, βασιζόμενος σε λέξεις κλειδιά. Στη συνέχεια χρησιμοποιεί αυτό το προφίλ προκειμένου να προτείνει στο χρήστη κάποιες ενδιαφέρουσες τοποθεσίες στο διαδίκτυο.

Αντιδραστικοί Πράκτορες

Οι αντιδραστικοί πράκτορες αποτελούν μια ειδική κατηγορία πρακτόρων οι οποίοι έχουν σαν κυρίαρχο χαρακτηριστικό την απουσία μιας εσωτερικής αναπαράστασης του περιβάλλοντος στο οποίο βρίσκονται. Αντιθέτως βασίζουν τη συμπεριφορά τους μόνο σε αντιδράσεις στα ερεθίσματα που λαμβάνουν από την τρέχουσα κατάσταση του περιβάλλοντος στο οποίο τοποθετούνται. Ένα σημαντικό στοιχείο που πρέπει να προσθέσουμε εδώ είναι ότι αυτοί οι πράκτορες είναι σχετικά απλοί και αλληλεπιδρούν με άλλους πράκτορες με κάποιους βασικούς τρόπους επικοινωνίας. Δεν σχεδιάζουν πλάνα ενεργειών, ενώ οι ενέργειές τους εξαρτώνται αποκλειστικά από το τι συμβαίνει την παρούσα στιγμή.

Τα οφέλη τα οποία παρακινούν την ανάπτυξη τέτοιων πρακτόρων είναι η ελπίδα ότι αυτοί θα είναι πιο ακμαίοι και περισσότερο ανεκτικοί στα λάθη από τα υπόλοιπα συστήματα πρακτόρων (ένας πράκτορας μπορεί να χάσει τον προσανατολισμό του, αλλά αυτό δεν θα έχει καταστροφικές συνέπειες), η προσαρμοστικότητά τους και ο γρήγοροι χρόνοι αντίδρασης και τέλος η προσδοκία ότι αυτός ο τομέας έρευνας θα αντιμετωπίσει το πρόβλημα του πλαισίου που οι παραδοσιακές τεχνικές της ΤΝ δεν έχουν κατορθώσει ακόμα.

Ένα αντιπροσωπευτικό παράδειγμα αρχιτεκτονικής τέτοιων συστημάτων αποτελεί η *αρχιτεκτονική υπαγωγής (subsumption architecture)*, η οποία σχεδιάστηκε από τον Brooks (1986) και εφαρμόστηκε κυρίως σε ρομπότ. Σύμφωνα λοιπόν με τον ερευνητή το σύστημα αποτελείται από *επαυξημένες μηχανές πεπερασμένων καταστάσεων (Augmented Finite State Machines - AFSM)*, κάθε μια από τις οποίες αναλαμβάνει μια

ενέργεια/συμπεριφορά. Κάθε AFSM ενεργοποιείται βάσει των τιμών των αισθητήρων, μέσω των οποίων ο πράκτορας αντιλαμβάνεται το περιβάλλον. Οι AFSM είναι τοποθετημένες σε επίπεδα με τέτοιο τρόπο έτσι ώστε εκείνες που ανήκουν σε ανώτερα επίπεδα να μπορούν να αναστείλουν τη λειτουργία των κατώτερων. Από την αλληλεπίδραση των μηχανών αυτών, το σύστημα είναι ικανό να αντιδράσει γρήγορα στα ερεθίσματα που λαμβάνει κάθε φορά από το περιβάλλον και να επιδείξει έξυπνη συμπεριφορά.

Υβριδικοί Πράκτορες

Με τον όρο υβριδικός πράκτορας αναφερόμαστε σε έναν πράκτορα ο οποίος αποτελεί συνδυασμό δύο ή περισσοτέρων φιλοσοφιών που είδαμε προηγουμένως. Ο λόγος για τον οποίο καταφεύγουμε σε έναν τέτοιο συνδυασμό πολλών φιλοσοφιών σε έναν πράκτορα είναι η πεποίθηση ότι για κάποιες εφαρμογές το κέρδος που αποκομίζουμε είναι μεγαλύτερο από την περίπτωση που ο πράκτορας σχεδιάζεται ακολουθώντας αποκλειστικά μία φιλοσοφία. Τα ιδανικά οφέλη θα είναι η ένωση των οφελών των ξεχωριστών φιλοσοφιών που συμμετέχουν στον υβριδικό πράκτορα. Σε μια τέτοια αρχιτεκτονική υπάρχουν συνήθως τουλάχιστον δύο επίπεδα: α) ένα υπεύθυνο για την αντιδραστική συμπεριφορά του πράκτορα και β) ένα για εκείνη της συμπεριφοράς με εσωτερική κατάσταση. Ο έλεγχος ροής μπορεί να είναι είτε οριζόντιος, δηλαδή όλα τα επίπεδα να είναι συνδεδεμένα στους αισθητήρες εισόδου και στους μηχανισμούς δράσης, ή κάθετος έχοντας ένα μόνο επίπεδο συνδεδεμένο στους αισθητήρες και ένα στους μηχανισμούς δράσης.

Ένα παράδειγμα αρχιτεκτονικής αυτής της κατηγορίας είναι η αρχιτεκτονική των πρακτόρων Touring Machine (Ferguson, 1992) που προτάθηκε για την καθοδήγηση αυτόνομων οχημάτων μεταξύ σημείων σε ένα περιβάλλον όπου υπάρχουν δρόμοι στους οποίους εμφανίζονται παρόμοιοι πράκτορες. Περιέχει τρία επίπεδα: το αντιδραστικό επίπεδο, το επίπεδο σχεδιασμού και το επίπεδο μοντελοποίησης. Αυτά δέχονται ταυτόχρονα δεδομένα εισόδου από τους αισθητήρες και προτείνουν ενέργειες που πρέπει να κάνει ο πράκτορας.

Κάθετη ροή ελέγχου χρησιμοποιείται στο σύστημα InterRRaP (Muller 1994). Και σε αυτό το σύστημα υπάρχουν τρία επίπεδα: το επίπεδο καθορισμού συμπεριφοράς, το επίπεδο σχεδιασμού και το επίπεδο συνεργασίας. Σε κάθε επίπεδο είναι προσαρτημένη μια βάση γνώσης η οποία εμπεριέχει μια αναπαράσταση των στοιχείων του κόσμου που ενδιαφέρουν κάθε επίπεδο. Τέλος, υπάρχει και ένα κατώτατο επίπεδο το οποίο

διαχειρίζεται τόσο την είσοδο όσο και την έξοδο του πράκτορα με το περιβάλλον (δεδομένα από τους αισθητήρες, ενέργειες, επικοινωνία, κλπ).

Νοήμονες Πράκτορες

Θα κλείσουμε την περιγραφή των πρακτόρων που εμπεριέχονται στο τυπολόγιο, με την αναφορά μας στους νοήμονες πράκτορες. Τι είναι ένας νοήμων πράκτορας; Η απάντηση σε αυτό το ερώτημα είναι τόσο εύκολη όσο και η απάντηση στο ερώτημα τι είναι νοημοσύνη. Σύμφωνα με τον Wooldridge, ένας νοήμων πράκτορας έχει την ικανότητα εκτέλεσης μιας ευέλικτης αυτόνομης ενέργειας προκειμένου να επιτύχει τους στόχους σχεδιασμού του, όπου με τον όρο ευέλικτη υπονοούνται τρία χαρακτηριστικά: αντιδραστικότητα, προνοητικότητα και κοινωνικότητα. Οι ικανότητες δηλαδή του πράκτορα να παρακολουθεί το περιβάλλον του και να αντιδρά σε κάποια χρονικά περιθώρια στις αλλαγές που παρατηρούνται σε αυτό, να εμφανίζει συμπεριφορά που κατευθύνεται από τους στόχους και να αλληλεπιδρά με τους άλλους πράκτορες προκειμένου να ικανοποιήσουν τους στόχους σχεδιασμού τους.

2.2 Πολυπρακτορικά Συστήματα

Ένα σύστημα που σχεδιάστηκε και υλοποιήθηκε ως ένα σύνολο πρακτόρων που αλληλεπιδρούν αποτελεί ένα *πολυπρακτορικό σύστημα* (*Multi-Agent System - MAS)*. Τα πολυπρακτορικά συστήματα μαζί με την κατανεμημένη επίλυση προβλημάτων (distributed problem solving) θεωρούνται βασικός τομέας της Κατανεμημένης Τεχνητής Νοημοσύνης (distributed artificial intelligence).

Ο λόγος που καταφεύγουμε στη δημιουργία ενός δικτύου από πράκτορες που αλληλεπιδρούν είναι η επίλυση προβλημάτων που ξεπερνούν τις δυνατοτήτες και τις γνώσεις ενός μόνο πράκτορα. Επιπρόσθετοι στόχοι της δημιουργίας ενός πολυπρακτορικού συστήματος θα μπορούσαν να είναι: α) η δυνατότητα επίλυσης πολύπλοκων προβλημάτων στα οποία δεν είναι δυνατή η εξεύρεση αποδοτικής λύσης από ένα μόνο πράκτορα, β) η επίλυση προβλημάτων που από τη φύσης τους είναι κατανεμημένα και γ) η διασύνδεση και λειτουργία ήδη υπαρχόντων συστημάτων ώστε να διευκολύνεται η εκμετάλλευσή τους χωρίς σημαντικές τροποποιήσεις. Στα πολυπρακτορικά συστήματα παρατηρούμε δύο τρόπους λειτουργίας των απλών πρακτόρων, α) είτε εργάζονται αυτόνομα και ανταλλάσσουν πληροφορίες με σκοπό την επίτευξη των δικών τους ανεξάρτητων στόχων είτε β) συνεργάζονται επιλύοντας υποπροβλήματα έτσι ώστε ο συνδυασμός των επιμέρους λύσεων που θα προκύψουν να αποτελέσει την τελική λύση.

Τα κύρια χαρακτηριστικά ενός δικτύου συνεργαζόμενων πρακτόρων είναι:

- η δυνατότητα συνεργασίας/διαπραγμάτευσης μέσω κάποιας γλώσσας επικοινωνίας.
- κανένας πράκτορας δεν έχει πλήρη πληροφορία.
- δεν υπάρχει κεντρικός έλεγχος στο σύστημα.
- τα δεδομένα είναι κατανεμημένα.
- οι υπολογισμοί γίνονται με ασύγχρονο τρόπο.

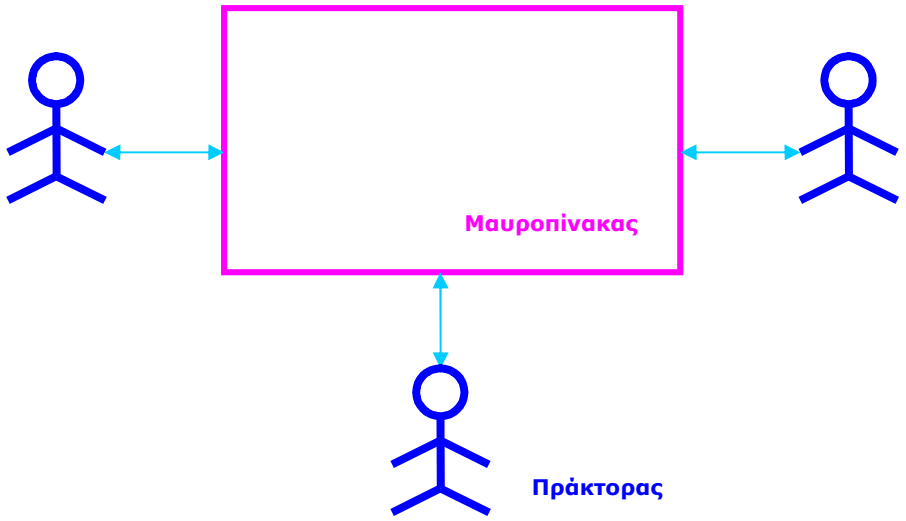
Όμως, η άποψη ότι πολλοί πράκτορες μπορούν ολοκληρώσουν καλύτερα μια εργασία που τους έχει ανατεθεί από ότι ένας μόνο πράκτορας, δεν ισχύει πάντα δυστυχώς. Τα προβλήματα στα οποία είναι απαραίτητο να δοθεί ιδιαίτερη προσοχή ώστε να βρεθεί ο καταλληλότερος τρόπος αντιμετώπισής τους έχουν να κάνουν κυρίως με την επικοινωνία των πρακτόρων και τον τρόπο συνεργασίας τους. Τα προβλήματα επικοινωνίας έχουν να κάνουν κυρίως με το πότε αυτοί θα επιχειρήσουν να επικοινωνήσουν και με το τι πληροφορία ανταλλάσσουν κατά την επικοινωνία τους. Επίσης σημαντικό πρόβλημα που προκύπτει, αποτελεί το ποιες γλώσσες και πρωτόκολλα θα χρησιμοποιηθούν. Τα προβλήματα επικοινωνίας αφορούν το πώς να τυποποιηθεί, να περιγραφεί, να μοιραστεί το πρόβλημα στους διάφορους νοήμονες πράκτορες που μετέχουν στην ομάδα και να συντεθούν οι λύσεις του. Επιπλέον θα πρέπει να αντιμετωπισθεί το πρόβλημα του πώς θα γίνει ο συμβιβασμός διαφορετικών απόψεων από διαφορετικούς πράκτορες, πώς θα αντιμετωπισθούν ενδεχόμενες συγκρουόμενες προθέσεις και επιθυμίες τους και πώς θα γίνει η διαχείριση περιορισμένων πόρων του συστήματος. Τέλος πρέπει να επιλυθεί το πρόβλημα του πώς κάθε πράκτορας θα αναπαραστήσει και θα συλλογιστεί για τις ενέργειες, τα σχέδια δράσης και τη γνώση άλλων συνεργαζόμενων πρακτόρων μέσα στην ομάδα.

Η επικοινωνία πρακτόρων απαιτεί την συμφωνία σε τρία διαφορετικά επίπεδα:

- κατώτερο επίπεδο – τρόπος διασύνδεσης
- μεσαίο επίπεδο – σύνταξη και μορφή μηνυμάτων
- ανώτερο επίπεδο – σημασιολογία

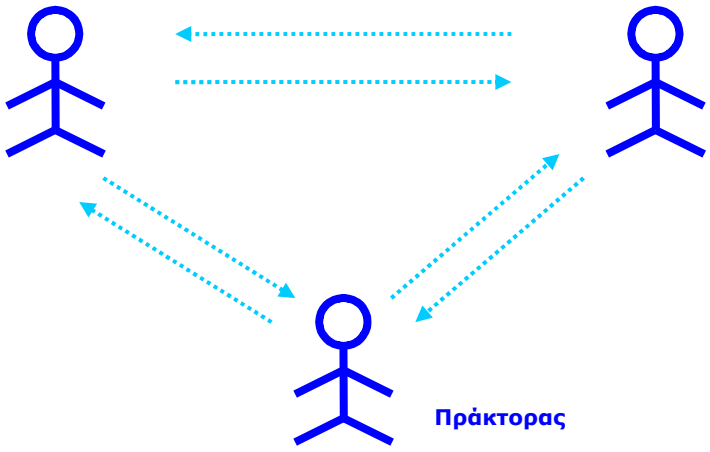
Τα διάφορα μοντέλα διασύνδεσης τα οποία έχουν προταθεί από τους ερευνητές για την επίλυση του προβλήματος της συνεργασίας των πρακτόρων μπορούν να ομαδοποιηθούν σε δύο κύριες κατηγορίες: στα *συστήματα μαυροπίνακα (blackboard systems)* και τα *συστήματα ανταλλαγής μηνυμάτων (message passing systems)*.

Στα συστήματα μαυροπίνακα έχουμε έναν κοινό χώρο εργασίας για όλους τους πράκτορες του συστήματος. Μέσα σε αυτό τον χώρο οι πράκτορες είτε ανταλλάσουν αποτελέσματα είτε μοιράζονται εργασίες. Από τη στιγμή που κάτι τοποθετείται στον κοινό αυτό χώρο είναι αυτόματα προσπελάσιμο από όλους τους πράκτορες που συμμετέχουν στο σύστημα.



Εικόνα 3. Συστήματα Μαυροπίνακα

Στα συστήματα ανταλλαγής μηνυμάτων οι πράκτορες ανταλλάσουν πληροφορία και συνεργάζονται μέσω μηνυμάτων τα οποία αποστέλλουν ο ένας στον άλλο βάσει συγκεκριμένων γλωσσών υψηλού επιπέδου. Αυτά τα συστήματα προσφέρουν μεγαλύτερη ευελιξία στην ανταλλαγή πληροφοριών από ότι τα συστήματα μαυροπίνακα.



Εικόνα 4. Συστήματα Ανταλλαγής Μηνυμάτων

Μέριμνα πρέπει να ληφθεί και για τον τύπο της επικοινωνίας. Αυτός μπορεί να είναι είτε *σύγχρονος*, είτε *ασύγχρονος*. Κατά τη σύγχρονη επικοινωνία ο πράκτορας που θέτει μια

ερώτηση είναι απαραίτητο να διακόψει τη λειτουργία του μέχρι να πάρει μια απάντηση ενώ κατά την ασύγχρονη η απάντηση μπορεί να έλθει οποιαδήποτε στιγμή μετά το χρόνο υποβολής της ερώτησης, χωρίς η λειτουργία του πράκτορα να διακόπτεται.

Τέλος ένα άλλο θέμα που αφορά στην επικοινωνία των πρακτόρων σε ένα σύστημα είναι ο βαθμός της επικοινωνίας. Σαν βαθμό ορίζουμε τον αριθμό των αποστολέων και των παραληπτών κατά την ανταλλαγή πληροφορίας. Ο βαθμός μπορεί να είναι 1 προς 1 όταν έχουμε μόνο έναν αποστολέα και έναν μόνο παραλήπτη, 1 προς Ν όταν ένας αποστολέας στέλνει σε πολλούς παραλήπτες και Ν προς Ν όταν όλοι είναι ταυτόχρονα αποστολείς και παραλήπτες.

2.3 Αγορές Πληροφορίας

Οι αγορές κατέχουν έναν ζωτικό ρόλο στην οικονομική ζωή, καθώς είναι τα μέρη στα οποία οι οντότητες που συμμετέχουν, διαπραγματεύονται την ανταλλαγή αγαθών, υπηρεσιών, πληροφοριών και αμοιβών. Εμείς θα επικεντρώσουμε το ενδιαφέρον μας στις αγορές πληροφορίας. Οι αγορές πληροφορίας με τη σειρά θα μπορούσαν να οριστούν σαν τα μέρη που οι συμμετέχοντες προσπαθούν να αγοράσουν ή να πουλήσουν προϊόντα πληροφορίας (π.χ. μουσική, εικόνες, βίντεο, ηλεκτρονικές εφημερίδες, κώδικα λογισμικού). Χαρακτηρίζονται για τη δυναμική τους φύση καθώς συνεχώς αλλάζει ο αριθμός και η συμπεριφορά των συμμετεχόντων

Μετέχοντες σε μια Αγορά

Καθώς λειτουργεί η αγορά παράγεται οικονομικό όφελος για τους πωλητές, τους αγοραστές και για τους ενδιαμέσους:

- πωλητές – έχουν στην κατοχή τους προϊόντα τα οποία θέλουν να πουλήσουν στους ενδιαφερόμενους με την πιο επικερδή τιμή. Οι πωλητές έχουν ένα συγκεκριμένο κόστος για να αναζητήσουν και να ανακτήσουν την πληροφορία και δεν επιθυμούν να πουλήσουν το κάθε προϊόν σε τιμή κάτω από το κόστος αυτό.
- αγοραστές – αναζητούν προϊόντα που ικανοποιούν τις απαιτήσεις τους και είναι πρόθυμοι να πληρώσουν γι’ αυτά. Έχουν μια συγκεκριμένη εκτίμηση τιμής για κάθε προϊόν και δεν είναι διατεθειμένοι να το αγοράσουν σε τιμή μεγαλύτερη από αυτή.
- ενδιάμεσοι – βοηθούν στο ‘ζευγάρωμα’ των πωλητών με τους αγοραστές. Παρέχουν πληροφορίες των διαθέσιμων προϊόντων στους αγοραστές, παρέχουν υπηρεσίες marketing στους πωλητές, συνδυάζουν προϊόντα πληροφορίας, επιβαρύνονται κάποιες φορές με τις μεταφορές και πληρωμές, παρέχουν το

απαραίτητο επίπεδο εμπιστοσύνης μεταξύ των πωλητών και των αγοραστών και διασφαλίζουν την ακεραιότητα της αγοράς (Bakos, 1998). Διαφοροποιούνται ως προς τη λειτουργία και το ρόλο τους στην εικονική αγορά σε τρεις κατηγορίες σύμφωνα με τους Decker κ.ά. (1997). Κατηγοριοποιούνται σε : α) τους πράκτορες οι οποίοι παίζουν τον ρόλο του μαυροπίνακα (blackboard agents) στην εικονική αγορά. Στους πράκτορες αυτούς καταγράφονται οι διάφορες εκδηλώσεις ενδιαφέροντος για την απόκτηση προϊόντων από τους αγοραστές. Οι πωλητές με τη σειρά τους υποβάλλουν ερωτήματα σε αυτούς τους πράκτορες εάν υπάρχουν περιπτώσεις τις οποίες θα μπορούσαν να ικανοποιήσουν. β) τους brokers, οι οποίοι διασφαλίζουν την ανωνυμία και των υποψηφίων αγοραστών και των υποψηφίων πωλητών. Οι πράκτορες αυτοί καταγράφουν τις επιθυμίες των αγοραστών και τις προσφορές των πωλητών και προσπαθούν να τις συνταιριάξουν. Αξιοσημείωτο όμως, είναι το γεγονός ότι ούτε οι αγοραστές, ούτε οι πωλητές γνωρίζουν τις οντότητες με τις οποίες συνδιαλέγονται για την ολοκλήρωση μιας αγοραστικής διαδικασίας. γ) τους matchmakers, οι οποίοι συγκεντρώνουν διαφημίσεις προσφορών από τους πωλητές τις οποίες μπορούν να αναζητήσουν οι αγοραστές προκειμένου να βρουν αυτή που ικανοποιεί καλύτερα τις προτιμήσεις και ανάγκες τους. Η διαφορά έγκειται ότι σε αυτή την περίπτωση ο αγοραστής συνδιαλέγεται άμεσα με τον πωλητή για την ολοκλήρωση της αγοράς, χωρίς την συμμετοχή της ενδιάμεσης οντότητας.

Οι αγορές έχουν τρεις κυρίως λειτουργίες: το ταίριασμα των αγοραστών και των πωλητών, την διευκόλυνση της ανταλλαγής των πληροφοριών, προϊόντων, υπηρεσιών και πληρωμών που σχετίζονται με τις συναλλαγές της αγοράς και την παροχή της κατάλληλης υποδομής (θεσμικό πλαίσιο) η οποία επιτρέπει την αποτελεσματική λειτουργία της αγοράς (Bakos, 1998). Οι δύο πρώτες λειτουργίες εκτελούνται από τους ενδιάμεσες οντότητες, ενώ για την τρίτη μεριμνούν οι κυβερνήσεις. Στις ηλεκτρονικές αγορές (όπως η αγορά πληροφορίας) κάνουμε χρήση των επιτευγμάτων της επιστήμης της Πληροφορικής προκειμένου να εκτελέσουμε αυτές τις λειτουργίες με αυξημένη αποτελεσματικότητα και με μειωμένα κόστη συναλλαγών.

Ταίριασμα Αγοραστών – Πωλητών

Όταν λέμε ταίριασμα αγοραστών – πωλητών αναφερόμαστε ουσιαστικά στη διαδικασία ζευγαρώματος της προσφοράς των αγαθών και της αντίστοιχης ζήτησής τους. Αυτή η διαδικασία εμπεριέχει τρία κομμάτια: *τον καθορισμό των προσφορών*, *την έρευνα* και *τον προσδιορισμός της τιμής*. Οι πωλητές παρακολουθούν διαρκώς την αγορά και

αποκτούν χρήσιμες πληροφορίες (ζήτηση) προκειμένου να αναπτύξουν προϊόντα τα οποία ικανοποιούν τις απαιτήσεις των αγοραστών. Προσπαθούν να προγραμματίσουν την προσφορά των αγαθών τους στην αγορά με τέτοιο τρόπο ώστε να μεγιστοποιούν τα οφέλη τους. Από την άλλη, οι αγοραστές προβαίνουν στην αγορά προϊόντων αφού λάβουν υπόψη τους την τιμή και τα χαρακτηριστικά τους. Η απόκτηση όμως αυτών των πληροφοριών επιφέρει κάποιο κόστος (π.χ. χρόνος - περιοδικά). Κόστος όμως έχουμε και για τους πωλητές οι οποίοι πρέπει να μεριμνήσουν και οι ίδιοι να υπάρξει η απαραίτητη πληροφόρηση στους υποψήφιους αγοραστές των προϊόντων τους (π.χ. διαφήμιση). Το τρίτο κομμάτι όπως είπαμε είναι ο προσδιορισμός της τιμής, η διαδικασία καθορισμού της τιμής στην οποία η ζήτηση συναντά την προσφορά και πραγματοποιείται η εμπορική συναλλαγή.

Εφόσον έχει επέλθει συμφωνία μεταξύ αγοραστών και πωλητών πρέπει να διασφαλιστεί η μεταφορά του προϊόντος στον αγοραστή και της πληρωμής στον πωλητή. Επιπρόσθετα είναι απαραίτητο να θεσμοθετηθεί ένα κατάλληλο επίπεδο ασφαλείας το οποίο θα προφυλάξει τους αγοραστές, πωλητές και ενδιαμέσους από την εμπλοκή άλλων κακόβουλων συμμετεχόντων. Τέλος οι αγορές πρέπει να παρέχουν την απαραίτητη υλική υποδομή η οποία επιτρέπει την διενέργεια εμπορικών συναλλαγών, όπως Η/Υ, δίκτυα υπολογιστών και συστήματα μεταφοράς.

Η θεσμική υποδομή προσδιορίζει τους νόμους και τους κανόνες οι οποίοι επικρατούν κατά τις εμπορικές συναλλαγές. Αυτοί οι κανόνες πρέπει να λαμβάνουν υπόψη τους τον δυναμικό χαρακτήρα των ηλεκτρονικών αγορών τον οποίο εκμεταλλεύονται κάποιοι με κακόβουλες διαθέσεις.

<u>Η επίδραση του Internet στις Αγορές</u>

Ο ηλεκτρονικός χαρακτήρας κάποιων αγορών, κυρίως των διαδικτυακών αγορών, επιδρά σημαντικά στις λειτουργίες που αναφέραμε προηγουμένως.

→ Προσφορές προϊόντων

Δύο χαρακτηριστικά που διαφοροποιούν τα προϊόντα των ηλεκτρονικών αγορών είναι η αυξημένη εξατομίκευση και ο συνδυασμός – διαμελισμός των προϊόντων πληροφορίας. Στις ηλεκτρονικές αγορές κυρίαρχος στόχος είναι η ανάπτυξη προϊόντων που ανταποκρίνονται στις ιδιαίτερες προτιμήσεις των πελατών, είτε αυτές εκφράζονται, είτε συμπεραίνονται. Αυτό έχει σαν απόρροια την βελτίωση της αποτελεσματικότητας των πωλήσεων. Αυτό επιτυγχάνεται, επιπλέον, από την ικανότητα του εντοπισμού της χρονικής στιγμής κατά την οποία ο πελάτης είναι πιθανότερο να προβεί στην αγορά

ενός προϊόντος πιο γρήγορα από τους ανταγωνιστικούς πωλητές. Επιπρόσθετα όμως κατά την κατάθεση των προσφορών των προϊόντων οι πωλητές πρέπει να προσδιορίσουν πως θα συνδυάσουν ή αποσυνθέσουν τα διάφορα κομμάτια των διαθέσιμων προϊόντων έτσι ώστε να μεγιστοποιήσουν τα οφέλη τους αλλά και να μεγιστοποιήσουν την ικανοποίηση των υποψήφιων πελατών τους.

→ Έρευνα

Στις ηλεκτρονικές αγορές μειώνεται σημαντικά το κόστος των αγοραστών για την αναζήτηση πληροφοριών σχετικές με την τιμή και τα χαρακτηριστικά των προϊόντων που προσφέρονται από τους πωλητές, καθώς επίσης και το κόστος των πωλητών για την προώθηση και επικοινωνία των διαθέσιμων προϊόντων. Η μείωση του κόστους έρευνας των αγοραστών επιφέρει βελτίωση της οικονομικής αποτελεσματικότητας. Οι αγοραστές δεν καρπώνονται μόνο χαμηλότερες τιμές καθώς έχουν να επιλέξουν ανάμεσα σε πληθώρα προσφορών, αλλά επίσης από τη δυνατότητα επιλογής προϊόντων που ανταποκρίνονται καλύτερα στις ανάγκες τους και τις προτιμήσεις τους.

→ Προσδιορισμός τιμής

Η παραγωγή εξατομικευμένων προϊόντων σε συνδυασμό με την πρόσβαση σε πληροφορίες που περιγράφουν και προσδιορίζουν τους υποψήφιους αγοραστές καθιστά ικανούς τους πωλητές στο να προτείνουν διαφορετικές τιμές στους διάφορους αγοραστές. Αυτή η διαφοροποίηση της τιμολογιακής πολιτικής των πωλητών αυξάνει από τη μια τα κέρδη και από την άλλη τους επιτρέπει να εξυπηρετήσουν αγοραστές οι οποίοι θα είχαν λάβει διαφορετική τιμή εκτός της αγοράς.

→ Διευκόλυνση συναλλαγών

Οι ηλεκτρονικές αγορές όπως προαναφέραμε διευκολύνουν τη μεταφορά πληροφοριών μεταξύ αγοραστών και πωλητών, έχοντας σαν αποτέλεσμα χαμηλότερα κόστη και ταχύτερη μεταφορά προϊόντων καθώς και περιορισμό των απαραίτητων επενδύσεων.

2.4 Επίλογος

Μια αγορά επιτρέπει την ανταλλαγή πληροφοριών, αγαθών, υπηρεσιών και χρημάτων. Όταν λειτουργεί, επιφέρει όφελος οικονομικό σε όλες τις οντότητες που συμμετέχουν σε αυτή, είτε πρόκειται για πωλητές, είτε για αγοραστές, είτε για ενδιαμέσους. Στην περίπτωση που τα αγαθά που διατίθενται είναι προϊόντα πληροφορίας (π.χ. μουσική, εικόνα και βίντεο) αναφερόμαστε σε μια αγορά πληροφορίας. Η χρήση νοημόνων πρακτόρων σε τέτοιες αγορές θεωρείται μια αποτελεσματική λύση. Θα μπορούσαμε να αναθέσουμε σε αυτούς την αναζήτηση και απόκτηση των επιθυμητών προϊόντων

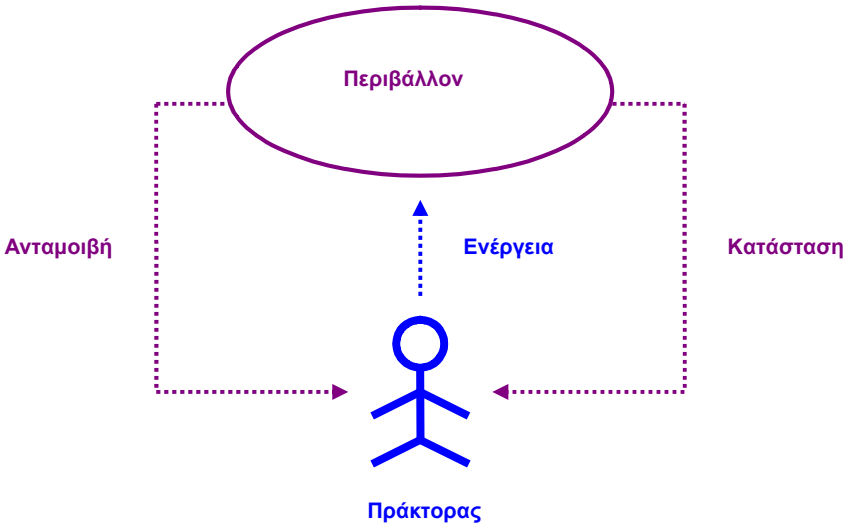
δίδοντάς τους μόνο τα επιθυμητά χαρακτηριστικά που θα πρέπει αυτά να έχουν αν λειτουργούσαν σαν αγοραστές. Θα μπορούσαν επίσης να ήταν οι εκπρόσωποί μας για την προώθηση των προϊόντων μας εάν είχαν το ρόλο του πωλητή. Ενώ θα μπορούσαν επίσης να αναλάβουν και το ταίριασμα της ζήτησης με την προσφορά αν δρούσαν σαν ενδιάμεσοι.

ΚΕΦΑΛΑΙΟ 3
ΠΟΛΥΠΡΑΚΤΟΡΙΚΗ ΕΝΙΣΧΥΤΙΚΗ ΜΑΘΗΣΗ

3.1 Εισαγωγή

Ως Ενισχυτική Μάθηση (reinforcement learning) αναφερόμαστε στο πρόβλημα που αντιμετωπίζει ένας πράκτορας ο οποίος διαμορφώνει τη συμπεριφορά του εντός ενός δυναμικού περιβάλλοντος διαμέσου αλληλεπιδράσεων της μορφής προσπάθεια–και–λάθος (trial and error). (Kaelbling κ.ά. 1996, Sutton & Barto, 1998). Προέκυψε σαν πεδίο έρευνας από τον συνδυασμό κυρίως της μάθησης τύπου προσπάθεια-και-λάθος που εφαρμόζεται στην ψυχολογία των ζώων και του δυναμικού προγραμματισμού.

Το μοντέλο αλληλεπίδρασης που μελετάται έχει ως εξής: σε κάθε διακριτό σημείο στο χρόνο ο πράκτορας μελετά την τρέχουσα κατάσταση (state) του περιβάλλοντος (environment) και επιλέγει την πραγματοποίηση μιας ενέργειας (action). Σαν αποτέλεσμα, το περιβάλλον μεταβαίνει σε μια νέα κατάσταση και ο πράκτορας λαμβάνει μια τιμή ανταμοιβής (reward) (δες Εικόνα 5). Αυτό το σήμα αποτελεί ένα μέτρο της ποιότητας των ενεργειών του πράκτορα, όπως καθορίζονται από το περιβάλλον.



Εικόνα 5. Αλληλεπίδραση πράκτορα – περιβάλλοντος

Ένα παράδειγμα το οποίο ξεδιαλύνει αυτά που αναφέραμε προηγουμένως είναι η ‘περιπέτεια’ ενός σκύλου σε έναν λαβύρινθο. Ως ενέργειες του σκύλου θεωρούμε τις κινήσεις του και σαν κατάσταση αναφερόμαστε στη θέση του, η οποία αλλάζει συνεχώς σαν συνέπεια των κινήσεων του. Ο σκύλος λαμβάνει μηδενική ανταμοιβή καθ’ όσο βρίσκεται στον λαβύρινθο και θετική , ένα μεγάλο κόκαλο, όταν βγαίνει από αυτόν.

Σε αυτό το απλό παράδειγμα μπορούμε να διακρίνουμε τα δυο βασικά χαρακτηριστικά της ενισχυτικής μάθησης: της αναζήτησης με δοκιμή-και-λάθος και της καθυστερημένης ενίσχυσης (Sutton & Barto, 1998). Στο σκύλο δεν παρέχεται καμία οδηγία ως προς το ποια θα ήταν η καλύτερη επιλογή σε κάποια διασταύρωση στον λαβύρινθο, όπως συμβαίνει στην περίπτωση της μάθησης με επίβλεψη (supervised learning). Κάθε φορά ενημερώνεται για την ποιότητα της επιλογής του μέσω μιας τιμής ανταμοιβής. Μόνος του καλείται να επιλέξει ποιες κινήσεις είναι οι καλύτερες (αναζήτηση με δοκιμή-και-λάθος). Ας υποθέσουμε τώρα ότι ο σκύλος παίρνει μια 'σημαντική' απόφαση για την κατεύθυνση που θα ακολουθήσει σε μια διασταύρωση του λαβύρινθου η οποία βρίσκεται σχετικά κοντά στην είσοδό του και η οποία βοήθησε το σκύλο να φτάσει στην έξοδο μετά από χρόνο t. Πρέπει να επισημάνουμε ότι η θετική ανταμοιβή, αποτέλεσμα αυτής της επιλογής θα δοθεί, στο σκύλο όταν παρέλθει αυτός ο χρόνος t. Ο σκύλος – πράκτορας θα κληθεί τότε να ξεδιαλύνει ποιες αποφάσεις του κατά την περιπλάνησή του στο λαβύρινθο και σε ποιον βαθμό τον βοήθησαν στο να βρει την έξοδο και να λάβει αυτήν την ανταμοιβή, καθυστερημένη ενίσχυση (delayed reinforcement).

Είναι προφανές ότι οι τεχνικές μάθησης που βασίζονται σε ένα τέτοιο μοντέλο είναι πολύ ελκυστικές. Κι αυτό διότι αν οι πράκτορες υπόκεινται σε ενισχυτική μάθηση, οι σχεδιαστές καλούνται μόνο να παράγουν την τιμή της ανταμοιβής. Η τιμή της ανταμοιβής είναι απαραίτητο να ανταποκρίνεται με ακρίβεια στο σύνολο των στόχων του πολυπρακτορικού συστήματος. Πράγμα που αποδεικνύεται ότι δεν είναι και πολύ εύκολο τελικά.

3.2 Ενισχυτική μάθηση ενός πράκτορα

Επειδή οι περισσότεροι αλγόριθμοι πολυπρακτορικής ενισχυτικής μάθησης βασίζονται σε μεθόδους μάθησης ενός πράκτορα θα ξεκινήσουμε με την παρουσίαση της θεωρία και των τεχνικών που εφαρμόζονται από αυτές τις μεθόδους.

Τυπικό μοντέλο

Στην περίπτωση της μάθησης ενός μόνο πράκτορα, η διαδικασία της ενισχυτικής μάθησης τυπικά αντιστοιχίζεται σε μια διαδικασία απόφασης Markov (Markov Decision Process - MDP).

Ορισμός: Μια κλειστή (finite) διαδικασία απόφασης Markov είναι μια πλειάδα (S, A, p_r, p_s) όπου το S αντιστοιχεί στο διακριτό (discrete) και κλειστό διάστημα των καταστάσεων, το A στο διακριτό και κλειστό διάστημα των ενεργειών, το p_r: S **x** A **x** S **x**

R → [0,1] είναι η κατανομή της πιθανότητας της ανταμοιβής και p_s: S **x** A **x** S → [0,1] η κατανομή της πιθανότητας της μετάβασης των καταστάσεων.

Είναι σημαντικό να αναφερθεί ότι υπάρχει κι ένας γενικός τύπος MDP με ανοικτά (infinite) και συνεχή (continuous) διαστήματα καταστάσεων κι ενεργειών και μέθοδοι που μεταχειρίζονται προβλήματα τέτοιου τύπου. Εμείς όμως θα αναφερθούμε μόνο στο κλειστά MDP.

Η συμπεριφορά του πράκτορα περιγράφεται από μια στοχαστική πολιτική η οποία αντιστοιχεί καταστάσεις σε ενέργειες h: S **x** A → [0,1]. Αυτή η πολιτική αλλάζει με την πάροδο του χρόνου από τον αλγόριθμο της ενισχυτικής μάθησης. Χρησιμοποιώντας το τυπικό μοντέλο, μπορούμε να περιγράψουμε την ενισχυτική μάθηση ως εξής: ο πράκτορας παρατηρεί την τρέχουσα κατάσταση s_k του περιβάλλοντος και επιλέγει μια ενέργεια a_k . Σαν αποτέλεσμα αυτής, το περιβάλλον μεταβαίνει σε μια νέα κατάσταση s_{k+1} με πιθανότητα p_s (s_k , a_k , s_{k+1}) και ο πράκτορας ανταμείβεται με r_{k+1} με πιθανότητα p_r (s_k , a_k , s_{k+1} , r_{k+1}).

Η πολιτική ενισχυτικής μάθησης μεταβάλλεται με το χρόνο όπως επιτάσσει ο αλγόριθμος μάθησης. Στο δικό μας μοντέλο η πολιτική δεν μεταβάλλεται με το χρόνο αλλά αποτελεί μια συνάρτηση των μεταβαλλόμενων με το χρόνο εσωτερικών καταστάσεων.

Στόχος μάθησης

Ένας πράκτορας ενισχυτικής μάθησης είναι ένας λογικός (rational) πράκτορας ο οποίος χρησιμοποιεί ένα μέτρο βελτιστότητας που εκφράζεται συναρτήσει των ανταμοιβών του. Ο πιο κοινός τύπος είναι ο ατελής ορίζοντας (infinite horizon) του προσδοκώμενου φθίνοντος αποτελέσματος (expected discount return):

$$R_k = E\left\{\sum_{l=0}^{\infty} \gamma^l r_{k+l+1}\right\} \quad (1)$$

ο οποίος είναι κατάλληλος και στις επεισοδιακές και συνεχείς περιπτώσεις. Η μεταβλητή γ που παίρνει τιμές στο [0,1] είναι ο παράγοντας μείωσης (Kaelbling κ.ά., 1996).

Η ιδιότητα Markov

Μια πολύ σημαντική υπόθεση πάνω στην οποία βασίζονται πολλά θεωρητικά αποτελέσματα είναι η ιδιότητα Markov. Αυτή η ιδιότητα αναφέρεται στο πως περιγράφεται η τρέχουσα κατάσταση. Ικανοποιείται εάν σε κάθε βήμα, περιλαμβάνεται όλη η πληροφορία που είναι απαραίτητη για τη διαδικασία λήψης απόφασης από τον

πράκτορα. Με απλά λόγια, μας είναι αρκετή η γνώση της τρέχουσας κατάστασης και της επιλεχθείσας ενέργειας ώστε να καθοριστεί η επόμενη κατάσταση και ανταμοιβή. Μπορούμε να την περιγράψουμε με την ακόλουθη ταυτότητα (Sutton & Barto, 1998).

$$P(s_{k+1}=s,r_{k+1}=r|s_k,a_k,s_{k-1},a_{k-1},\ldots,s_0,a_0) = P(s_{k+1}=s,r_{k+1}=r|s_k,a_k) \quad (2)$$

Συναρτήσεις αξίας και ισότητες Bellman

Σχεδόν όλοι οι αλγόριθμοι ενισχυτικής μάθησης βασίζουν τη λειτουργία τους στην εκτίμηση του πόσο καλό είναι για τον πράκτορα να βρίσκεται σε μια δεδομένη κατάσταση ή να πραγματοποιεί μια δεδομένη ενέργεια σε μια δεδομένη κατάσταση. Αυτές οι εκτιμήσεις εμπεριέχονται στις συναρτήσεις αξίας κατάστασης (state value function) και ενέργειας (action value function) αντίστοιχα (Kaelbling κ.ά. 1996, Sutton & Barto, 1998).

Ορισμός: Η αξία μιας κατάστασης s υπό την πολιτική h είναι το προσδοκώμενο φθίνων αποτέλεσμα όταν ξεκινούμε από την s και ακολουθούμε στη συνέχεια την h.

$$V^h(s) = E_h\left\{\sum_{l=0}^{\infty}\gamma^l r_{k+l+1} \mid s_k = s\right\} \quad (3)$$

Ορισμός: Η αξία πραγματοποίησης της ενέργειας a στην κατάσταση s υπό την πολιτική h είναι το προσδοκώμενο φθίνων αποτέλεσμα όταν ξεκινούμε από την s, πραγματοποιούμε την a και ακολουθούμε την h στη συνέχεια.

$$Q^h(s,a) = E_h\left\{\sum_{l=0}^{\infty}\gamma^l r_{k+l+1} \mid s_k = s, a_k = a\right\} \quad (4)$$

Οι αξίες ενέργειας αναφέρονται και ως ‘Q – αξίες’ (Q-values).

Μια πολύ σημαντική ιδιότητα των συναρτήσεων αξίας είναι ότι ικανοποιούν τις ακόλουθες αναδρομικές ισότητες για κάθε πολιτική h.

$$V^h(s) = \sum_{a\in A} h(s,a)\sum_{s'\in S} p_s(s,a,s')[p_r(s,a,s') + \gamma V^h(s')] \forall s \in S \quad (5)$$

$$Q^h(s,a) = \sum_{s'\in S} p_s(s,a,s')\left[p_r(s,a,s') + \gamma\sum_{a'\in A} h(s',a')Q^h(s',a')\right] \forall s \in S, a \in A \quad (6)$$

Ορισμός (Βελτιστότητα στα MDPs)

Η βέλτιστη συνάρτηση αξίας κατάστασης είναι η συνάρτηση που συσχετίζει κάθε κατάσταση με την μέγιστη αξία που μπορεί να αποκτηθεί από αυτήν την κατάσταση.

$$V^*(s) = \max_h V^h(s), \forall s \in S \quad (7)$$

Η βέλτιστη συνάρτηση αξίας ενέργειας είναι η συνάρτηση που συσχετίζει κάθε ζευγάρι κατάστασης – ενέργειας με την μέγιστη αξία που μπορεί να αποκτηθεί μετά την πραγματοποίηση αυτής της ενέργειας σε αυτήν την κατάσταση.

$$Q^*(s,a) = \max_h Q^h(s,a), \forall s \in S, a \in A \quad (8)$$

Μια πολιτική h^* χαρακτηρίζεται βέλτιστη όταν επιτυγχάνει τη χρήση της βέλτιστης συνάρτησης αξίας κατάστασης ή της βέλτιστης συνάρτησης αξίας ενέργειας.

Μια MDP έχει πάντα μια ντετερμινιστική βέλτιστη πολιτική που δίδεται από την

$$h^*(s) = \arg\max_{a \in A} Q^*(s,a) \quad (9)$$

Είναι πιθανό να υπάρχουν περισσότερες της μιας βέλτιστες πολιτικές, αλλά όλες επιτυγχάνουν τη χρήση των ίδιων βέλτιστων συναρτήσεων αξίας κατάστασης και ενέργειας. Οι βέλτιστες συναρτήσεις αξίας πρέπει να ικανοποιούν τις αναδρομικές σχέσεις (5) και (6). Οι εκφράσεις που προκύπτουν αποτελούν τις γνωστές ισότητες βελτιστότητας Bellman.

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} p_s(s,a,s')[p_r(s,a,s') + \gamma V^*(s')], \forall s \in S \quad (10)$$

$$Q^*(s,a) = \sum_{s' \in S} p_s(s,a,s')\left[p_r(s,a,s') + \gamma \max_{a' \in A} Q^*(s',a')\right], \forall s \in S, a \in A \quad (11)$$

Αυτές οι ισότητες αποτελούν τη ραχοκοκαλιά των περισσότερων τεχνικών Ενισχυτικής Μάθησης.

Εξερεύνηση

Αν δούμε την ισότητα (9) θα παρατηρήσουμε ότι ο πράκτορας επιλέγει σε κάθε βήμα την ενέργεια που αποφέρει τη μέγιστη αξία σύμφωνα με τη συνάρτηση αξίας ενέργειας. Αυτή η επιλογή ονομάζεται άπληστη (greedy). Εάν κάθε φορά ο πράκτορας επιλέγει μόνο την άπληστη ενέργεια, λέμε ότι αυτός ο πράκτορας αξιοποιεί τη γνώση. Η άπληστη επιλογή θα ήταν αναμφίβολα η ορθότερη ιδέα στην περίπτωση που ο πράκτορας γνώριζε τη βέλτιστη συνάρτηση αξίας. Είναι γεγονός όμως, ότι κατά τη διάρκεια της μάθησης ο πράκτορας έχει μόνο μια εκτίμηση γι' αυτή τη συνάρτηση αξίας και όχι την βέλτιστη. Είναι πιθανό η επιλογή κάποιας ενέργειας που τη δεδομένη στιγμή έχει μεγαλύτερη αξία σύμφωνα με τη συνάρτηση αξίας, να επιφέρει 'χειρότερα' αποτελέσματα από κάποια που υποτιμήθηκε. Σε κάθε βήμα, ο πράκτορας είναι

απαραίτητο να σταθμίσει το άμεσο όφελος της επιλογής της άπληστης ενέργειας με το πιθανό όφελος της επιλογής μιας άλλης ενέργειας, η πραγματοποίηση της οποίας μπορεί να μας αποκαλύψει τη λανθασμένη εκτίμηση της. Αυτός ο τρόπος επιλογής της ενέργειας ονομάζεται *εξερεύνηση (exploration)*.

Ας γυρίσουμε στο παράδειγμα του λαβυρίνθου πάλι και ας υποθέσουμε ότι σε μια διασταύρωση Τ ο σκύλος επιλέγει να κατευθυνθεί αριστερά. Η έξοδος είναι πιο κοντά εάν ακολουθήσει το δεξιό μονοπάτι, αλλά ο σκύλος έχει τη δυνατότητα να βγει από το λαβύρινθο ακολουθώντας και το αριστερό. Όταν ο σκύλος φτάσει στην έξοδο, η πιθανότητα επιλογής της αριστερής κατεύθυνσης θα γίνει μεγαλύτερη από τη δεξιά, καθώς αυτή η επιλογή τον οδήγησε στην επίτευξη του στόχου. Όταν ξαναφτάσει, λοιπόν, στη συγκεκριμένη διασταύρωση είναι πιο πιθανό να επιλέξει να κινηθεί αριστερά (άπληστη επιλογή). Μόνο στην περίπτωση που ο σκύλος είχε επιλέξει αντίθετα να κινηθεί δεξιά, θα διαπίστωνε ότι θα έφτανε στην έξοδο πιο γρήγορα.

Όταν λοιπόν η ενίσχυση καθυστερεί, χρησιμοποιούμε διάφορα ευριστικά δεδομένα προκειμένου να επιλέξουμε μεταξύ εξερεύνησης και αξιοποίησης. Ο Thrun (1992) ταξινόμησε τις στρατηγικές εξερεύνησης σε δυο κατηγορίες, σε *κατευθυνόμενες (directed)* και *μη κατευθυνόμενες (undirected)*. Η εξερεύνηση είναι κατευθυνόμενη όταν λαμβάνεται υπόψη κάποια μετρική για το προσδοκώμενο όφελος απόκτησης πληροφορίας. Στόχος η μεγιστοποίησή της. Στις περιπτώσεις που δεν χρησιμοποιούμε μια τέτοια μετρική έχουμε να κάνουμε με μη κατευθυνόμενες στρατηγικές.

Τεχνικές επίλυσης

→ *Τεχνικές με μοντέλο περιβάλλοντος*

Οι τεχνικές *με μοντέλο περιβάλλοντος (model-based techniques)* υποθέτουν την ύπαρξη ενός μοντέλου του περιβάλλοντος (συναρτήσεις ανταμοιβής και μετάβασης) και δίνουν σαν αποτέλεσμα τη βέλτιστη αλληλουχία ενεργειών χρησιμοποιώντας αυτό το μοντέλο. Αυτό επιτυγχάνεται εύκολα εάν χρησιμοποιήσουμε στην αρχή μια συνάρτηση βέλτιστης αξίας και στη συνέχεια ακολουθήσουμε άπληστη επιλογή ενεργειών. Η μέθοδος αυτή ονομάζεται *επανάληψη αξίας (Value Iteration)* και παρουσιάζεται στη συνέχεια.

Αλγόριθμος *Value Iteration*
Απαιτήσεις: p_s, το μοντέλο μετάβασης, p_r, το μοντέλο ανταμοιβής
Είσοδος: κατώφλι θ >=0, παράγοντας μείωσης γ, 0<=γ<1
Έξοδος: μια βέλτιστη πολιτική h*

1. $Q(s,a) \leftarrow 0, \forall s \in S, a \in A$
2. **επανάλαβε**

3. $\delta \leftarrow 0$
4. **για όλα** $s \in S, a \in A$
5. $q \leftarrow Q(s,a)$
6. $Q(s,a) \leftarrow \sum_{s' \in S} p_s(s,a,s')\left[p_r(s,a,s') + \gamma \max_{a' \in A} Q(s',a')\right]$
7. $\delta \leftarrow \max\{\delta, |Q(s,a) - q|\}$
8. **τέλος_για**
9. **μέχρις_ότου** $\delta \le \theta$
10. $h^*(s) \leftarrow \arg\max_{a \in A} Q(s,a), \forall s \in S$ → οι ισοπαλίες αντιμετωπίζονται με τυχαίο τρόπο

Ο αλγόριθμος αυτός μετατρέπει την ισότητα Bellman σε ανάθεση. Αποδεικνύεται ότι υπό συγκεκριμένες συνθήκες ο αλγόριθμος συγκλίνει στη βέλτιστη τιμή. Η *επανάληψη πολιτικής (Policy Iteration)* επενεργεί απ' ευθείας στην πολιτική του πράκτορα. Σε αυτόν τον αλγόριθμο διενεργούνται επαναληπτικά δυο αλληλοεξαρτώμενες διαδικασίες: α) η αξιολόγηση της πολιτικής, όπου υπολογίζεται μέσω της συνάρτησης αξίας η ποιότητα της τρέχουσας πολιτικής και β) η βελτίωση της πολιτικής, όπου χρησιμοποιείται αυτή η συνάρτηση αξίας προκειμένου να υπολογιστεί μια νέα καλύτερη πολιτική. Αποδεικνύεται ότι η αξιολόγηση της πολιτικής συγκλίνει στην πραγματική αξία της πολιτικής και ότι η βελτίωση της πολιτικής επιφέρει μια καλύτερη πολιτική, εκτός από την περίπτωση που η πολιτική είναι βέλτιστη.

Αλγόριθμος *Policy Iteration*
Απαιτήσεις: p_s, το μοντέλο μετάβασης, p_r, το μοντέλο ανταμοιβής
Είσοδος: κατώφλι θ >=0, παράγοντας μείωσης γ, 0<=γ<1
Έξοδος: μια βέλτιστη πολιτική h*

1. $h(s,a) \leftarrow$ *random action* $a \in A, \forall s \in S$
2. **επανάλαβε**
3. $Q(s,a) \leftarrow 0, \forall s \in S, a \in A$
4. **επανάλαβε** →αξιολόγηση πολιτικής
5. $\delta \leftarrow 0$
6. **για όλα** $s \in S, a \in A$
7. $q \leftarrow Q(s,a)$
8. $Q(s,a) \leftarrow \sum_{s' \in S} p_s(s,a,s')\left[p_r(s,a,s') + \gamma Q(s', h(s'))\right]$
9. $\delta \leftarrow \max\{\delta, |Q(s,a) - q|\}$
10. **τέλος_για**
11. **μέχρις_ότου** $\delta \le \theta$
12. $h_{stable} \leftarrow true$ →βελτίωση πολιτικής
13. **για όλα** $s \in S$
14. $c \leftarrow h(s)$
15. $h(s) \leftarrow \arg\max_{a \in A} Q(s,a)$ → οι ισοπαλίες αντιμετωπίζονται με τυχαίο τρόπο
16. **αν** $h(s) \neq c$ **τότε**
17. $h_{stable} \leftarrow false$
18. **τέλος_αν**

19. **τέλος_για**
20. **μέχρις_ότου** h_{stable}
21. $h^*(s) \leftarrow h$

→ *Τεχνικές χωρίς μοντέλο*

Στις περισσότερες των περιπτώσεων, όμως, δεν είναι εύκολη η απόκτηση ακριβών μοντέλων του κόσμου, όπως απαιτείται από τις τεχνικές που προαναφέραμε. Επιπρόσθετα, επειδή το διάστημα των καταστάσεων μερικών προβλημάτων είναι πολύ μεγάλο, είναι ανέφικτη η επεξεργασία όλων των καταστάσεων. Αυτά τα προβλήματα επιλύονται από τις τεχνικές χωρίς μοντέλο (model-free techniques).

Η κλάση τεχνικών χωρίς μοντέλο *Monte Carlo* αξιολογεί μια πολιτική με έναν πιο απλό τρόπο. Ακολουθούμε αυτή την πολιτική για ένα μεγάλο αριθμό επεισοδίων και υπολογίζουμε το μέσο των αποτελεσμάτων που προκύπτουν (Sutton & Barto, 1998). Σε αυτή την περίπτωση δεν μας χρειάζεται πλέον ένα μοντέλο καθώς τα αποτελέσματα προκύπτουν από την άμεση εμπειρία μας με το περιβάλλον. Επιπλέον, μας δίνεται η δυνατότητα να επικεντρώσουμε τον υπολογισμό στις περιοχές του διαστήματος καταστάσεων που παρουσιάζουν κάποιο ενδιαφέρον. Στο παράδειγμά μας, εάν ο σκύλος ξεκινά από μια θέση κοντά στην έξοδο ενός πολύ μεγάλου λαβυρίνθου, δεν μας ενδιαφέρει ο ακριβής υπολογισμός των αξιών που βρίσκονται πολύ κοντά στην αφετηρία.

Εκτός από τα προφανή πλεονεκτήματα των μεθόδων Monte Carlo υπάρχουν και κάποια μειονεκτήματα. Μεταξύ αυτών, είναι ο μεγάλος αριθμός των επεισοδίων που απαιτείται, ο οποίος είναι δαπανηρός στη συλλογή του και το γεγονός ότι οι συνεχείς εργασίες δεν έχουν ένα καλώς ορισμένο τέλος. Οι μέθοδοι, τώρα, *Χρονικής Διαφοράς* (*Temporal Difference Methods*) οι οποίες συνδυάζουν τις ιδέες του δυναμικού προγραμματισμού και των μεθόδων Monte Carlo, είναι ικανές να επιφέρουν τις απαραίτητες αλλαγές σε κάθε βήμα του κόσμου. Όπως και στις μεθόδους Monte Carlo χρησιμοποιείται η εμπειρία αλληλεπίδρασης με τον κόσμο προκειμένου να υπολογισθεί η ανταμοιβή, με τη διαφορά όμως, ότι η χρονική διαφορά βελτιώνει την πολιτική σε κάθε βήμα, κι όχι μετά από μεγάλες αλληλουχίες εμπειριών. Ο κανόνας ανανέωσης κατευθύνει την τρέχουσα εκτίμηση της αξίας προς κάποιον στόχο, απλά εδώ χρησιμοποιείται η πραγματική εμπειρία αντί του μοντέλου για να προσδιορισθεί η επόμενη κατάσταση και ανταμοιβή:

$$Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha\left[r_{k+1} + \gamma Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k)\right] \quad (12)$$

Εδώ ο πράκτορας επιλέγει την ενέργεια a_k στην κατάσταση s_k στο βήμα k και σαν αποτέλεσμα ο κόσμος μεταβαίνει στην κατάσταση s_{k+1}, λαμβάνοντας ανταμοιβή r_{k+1} . Η ενέργεια a_{k+1} είναι αυτή που επιλέγεται στο βήμα k+1. Κάποια ενδιαφέροντα στοιχεία:

- ✓ Αντίθετα με την περίπτωση της επανάληψης αξίας εδώ η εκτίμηση $Q(s_k, a_k)$ δεν μετακινείται αμέσως προς τον στόχο $\left[r_{k+1} + \gamma Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k)\right]$, αλλά μόνο κατά ένα κλάσμα α της απόστασης αυτής. Αυτό το κλάσμα αποτελεί το ρυθμό μάθησης (learning rate).
- ✓ Η εκτίμηση $Q(s_k, a_k)$ ανανεώνεται χρησιμοποιώντας κάποια άλλη εκτίμηση $\left[Q(s_{k+1}, a_{k+1})\right]$.
- ✓ Επειδή ο πράκτορας χρησιμοποιεί την επόμενη ενέργεια a_{k+1} για να προσδιορίσει την αξία της επόμενης κατάστασης οι εκτιμήσεις θα συγκλίνουν στην αξία της πολιτικής που χρησιμοποιείται πραγματικά. Ο κανόνας ανανέωσης γι' αυτό το λόγο λέγεται ότι είναι 'on – policy'.

Ο 'on – policy' αλγόριθμος που προκύπτει από το συνδυασμό της (12) με μια πολιτική που προκύπτει από τις τρέχουσες εκτιμήσεις αξίας ενέργειας ονομάζεται *SARSA*. Κάτω από συγκεκριμένες συνθήκες αποδεικνύεται ότι ο SARSA συγκλίνει στην βέλτιστη πολιτική. Ένα πρόβλημα που αντιμετωπίζουμε στην περίπτωση του SARSA είναι ότι ο πράκτορας δεν μπορεί να γνωρίζει τις πραγματικές βέλτιστες αξίες ενέργειας μέχρις ότου η πολιτική έχει συγκλίνει στην βέλτιστη πολιτική. Επιπρόσθετα οι συνθήκες σύγκλισης απαιτούν να ακολουθείται εξερευνητική πολιτική και επομένως όχι βέλτιστη. Αποδεικνύεται ότι οι μέθοδοι χρονικής διαφοράς μπορούν να εκτιμήσουν τη βέλτιστη συνάρτηση αξίας ενώ ακολουθούν μια μη βέλτιστη πολιτική. Αυτές οι μέθοδοι χαρακτηρίζονται 'off-policy'. Η τροποποίηση που απαιτείται στον κανόνα ανανέωσης δίδεται παρακάτω:

$$Q(s_k, a_k) \leftarrow Q(s_k, a_k) + \alpha\left[r_{k+1} + \gamma \max_{a' \in A} Q(s_{k+1}, a') - Q(s_k, a_k)\right] \quad (13)$$

Ο αλγόριθμος που προκύπτει από το συνδυασμό της (13) με μια πολιτική που παράγεται από τις τρέχουσες εκτιμήσεις αξίας ενέργειας ονομάζεται *Q-μάθηση* (Q-learning) και παρουσιάζεται στη συνέχεια (Watkins & Dayan, 1992).

Αλγόριθμος *Q-learning*
Είσοδος: ρυθμός μάθησης α, παράγοντας μείωσης γ

1. $Q(s, a) \leftarrow 0, \forall s \in S, a \in A$
2. παρατήρησε αρχική κατάσταση s

3. **βρόχος**
4. $a \leftarrow h(s)$ όπου h είναι μια πολιτική που παράγεται από το Q
5. εφάρμοσε a, παρατήρησε r και s'
6. $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a' \in A} Q(s',a') - Q(s,a)\right]$
7. $s \leftarrow s'$
8. **τέλος_βρόχου**

Αποδεικνύεται ότι ο Q-μάθηση συγκλίνει στη βέλτιστη πολιτική υπό πιο χαλαρές προϋποθέσεις απ' ότι ο SARSA (Watkins & Dayan, 1992).

Πρέπει να σημειώσουμε ότι ο κανόνας ανανέωσης της Q-μάθησης διαδίδει την πληροφορία μόνο για ένα βήμα κατά μήκος της τροχιάς του πράκτορα στο διάστημα των καταστάσεων. Για να γίνει αυτό αντιληπτό, θα μεταφερθούμε στο παράδειγμα του λαβυρίνθου. Αν υποθέσουμε ότι ο σκύλος φτάνει στο κόκαλο – έξοδο του λαβυρίνθου για πρώτη φορά, η πληροφορία της ανταμοιβής διαδίδεται μόνο στην πιο κοντινή στην έξοδο κατάσταση που επισκέφθηκε ο σκύλος, ας πούμε την s. Όταν λοιπόν τοποθετήσουμε ξανά το σκύλο στο λαβύρινθο σε κάποια άλλη θέση, αυτό δεν θα έχει ιδέα για το πώς θα φτάσει στο κόκαλο παρά μόνο όταν αυτό προσεγγίσει με κάποιο τρόπο κάποια κατάσταση κοντινή στην s. Από αυτήν θα μετακινηθεί στη συνέχεια στην s. Η πληροφορία στο σημείο αυτό μεταδίδεται άλλο ένα βήμα, κι ούτω καθ' εξής. Μέθοδος όμως που με απλά λόγια χαρακτηρίζεται ως αρκετά αναποτελεσματική.

Θα δούμε στη συνέχεια μια πιο αποτελεσματική μέθοδο. Σε αυτήν ο πράκτορας αποδίδει στην τροχιά που σχηματίζει με την κίνησή του στο λαβύρινθο ένα ίχνος εκλεξιμότητας (eligibility trace), όπως λέγεται Ε. Σε κάθε βήμα ανανεώνει όχι μόνο την τελευταία Q αξία αλλά ολόκληρο το σύνολο των Q αξιών που αντιστοιχούν στην τροχιά, αναλογικά στις τιμές της εκλεξιμότητά τους. Το ίχνος φθίνει εκθετικά με ένα παράγοντα λ ο οποίος ονομάζεται παράγοντας συχνότητας (recency factor).

Αλγόριθμος *Q(λ) (με προσαυξημένο ίχνος)*
Είσοδος: ρυθμός μάθησης α, παράγοντας μείωσης γ, παράγοντας συχνότητας λ

1. $Q(s,a) \leftarrow 0,\ E(s,a) \leftarrow 0,\ \forall s \in S, a \in A$
2. παρατήρησε αρχική κατάσταση s
3. **βρόχος**
4. $a \leftarrow \arg\max_{a' \in A} Q(s,a')$
5. τροποποίησε a σε εξερευν. κίνηση σύμφωνα με στρατηγική
6. **αν** $a \neq \arg\max_{a' \in A} Q(s,a')$ **τότε**
7. $E(s,a) \leftarrow 0,\ \forall s \in S, a \in A$ →αρχικοποίηση ίχνους
8. **τέλος_αν**
9. εφάρμοσε a, παρατήρησε r και s'

10. $\delta \leftarrow r + \gamma \max_{a' \in A} Q(s', a') - Q(s, a)$ →υπολόγισε χρονική διαφορά
11. $E(s, a) \leftarrow E(s, a) + 1$ →προσαύξησε ίχνος
12. **για όλα** $\tilde{s} \in S, \tilde{a} \in A$
13. $Q(\tilde{s}, \tilde{a}) \leftarrow Q(\tilde{s}, \tilde{a}) + \alpha \delta E(\tilde{s}, \tilde{a})$
14. $E(\tilde{s}, \tilde{a}) \leftarrow \lambda E(\tilde{s}, \tilde{a})$ →μείωσε ίχνος
15. **τέλος_για**
16. $s \leftarrow s'$
17. **τέλος_βρόχου**

Ο όρος προσαυξημένο (accumulating) αναφέρεται στο γεγονός ότι το ίχνος αυξάνεται κατά 1 στην γραμμή 11.

Μια άλλη μέθοδος χωρίς μοντέλο η οποία είναι ευριστική και πιο πολύπλοκη είναι το *Σύστημα Εκμάθησης Ταξινομητή* (Learning Classifier System). Ένα σύστημα εκμάθησης ταξινομητή είναι ένα “παράλληλο σύστημα, ανταλλαγής μηνυμάτων, βασιζόμενο σε κανόνες το οποίο σχεδιάστηκε ώστε να επιτρέπει πολύπλοκους μετασχηματισμούς και αναδιοργανώσεις της γνώσης του καθώς αυτό επιτελεί μια εργασία” (Booker, 1988). Για την αναπαράσταση της γνώσης του συστήματος χρησιμοποιείται ένα σύνολο κανόνων ταξινομητή. Κάθε ένας από αυτούς αποτελείται από το τμήμα της συνθήκης και το τμήμα του μηνύματος. Το τμήμα της συνθήκης είναι μια συμβολοσειρά από bit. Οι τιμές που μπορεί να πάρει κάθε bit είναι 0,1, ή #. Η δίεση # έχει τη σημασία του ‘δεν πειράζει’. Η συμπεριφορά του συστήματος καθορίζεται από 3 υποσυστήματα: 1) *εκτέλεσης (performance),* 2) *εκχώρησης πίστωσης (credit assignment)* και 3) *ανακάλυψης κανόνα (rule discovery)*. Η πληροφορία μεταφέρεται εντός του συστήματος μέσω μιας λίστας μηνυμάτων. Σε κάθε επανάληψη, μια διεπαφή εισόδου αναλαμβάνει τη μετατροπή της τρέχουσας κατάστασης του περιβάλλοντος σε ένα σύνολο δυαδικών μηνυμάτων τα οποία τοποθετούνται στη λίστα των μηνυμάτων. Τα μηνύματα της λίστας συγκρίνονται με το τμήμα συνθήκης των κανόνων. Το υποσύστημα της εκτέλεσης τότε, πυροδοτεί κάποιους από τους κανόνες βασιζόμενο στο βαθμό ταιριάσματος με τα μηνύματα και στη δύναμή τους. Η δύναμη περιγράφει την συνολική αξία του κανόνα στο σύστημα. Οι κανόνες οι οποίοι πυροδοτήθηκαν παράγουν αντίστοιχα μηνύματα τα οποία συνθέτουν μια καινούρια λίστα μηνυμάτων που αντικαθιστά την παλαιότερη. Η διεπαφή εξόδου τώρα, επεξεργάζεται αυτή τη λίστα για να παράγει τις ενέργειες που θα εκτελέσει το σύστημα. Είναι άξιο αναφοράς το γεγονός ότι κάποιοι κανόνες διατηρούνται στη λίστα, προσδίδοντας στο σύστημα τη δυνατότητα της μνήμης.

Πώς λειτουργεί τώρα το υποσύστημα εκχώρησης πίστωσης;. Κάθε κανόνας όταν πυροδοτείται ‘πληρώνει’ ένα τμήμα της δύναμής του. Με αυτό το τμήμα της δύναμής του ‘πληρώνει’ όλους τους κανόνες που ήταν ενεργοί στο προηγούμενο βήμα του χρόνου. Ταυτόχρονα, όμως πληρώνεται από όλους τους κανόνες που θα πυροδοτηθούν στο επόμενο βήμα. Η άμεση ανταμοιβή από το περιβάλλον κατανέμεται σε όλους τους κανόνες οι οποίοι ήταν ενεργοί πριν επικοινωνήσουμε μ’ αυτό. Το υποσύστημα εκχώρησης πίστωσης είναι παρόμοιο με τον κανόνα ανανέωσης της χρονικής διαφοράς. Οι δυνάμεις των ταξινομητών είναι παρόμοιες με τις αξίες ενέργειας και οι πληρωμές παίζουν το ρόλο της αύξησης της αξίας.

Το υποσύστημα ανακάλυψης κανόνων παράγει νέους κανόνες ταξινόμησης οι οποίοι θα βελτιώσουν την απόδοση του συστήματος χρησιμοποιώντας γενετικούς αλγορίθμους.

Η γενική έκδοση του συστήματος μάθησης ταξινομητών παρουσιάζει μερικά σημαντικά πλεονεκτήματα έναντι της μάθησης με χρονική διαφορά. α) Η παρουσία των συμβόλων ‘δεν πειράζει’ παρέχει τη δυνατότητα της γενίκευσης. β) Το σύστημα έχει τη δυνατότητα της μνήμης μέσω της άμεσης μεταφοράς μηνυμάτων από τη μια επανάληψη στην άλλη. Η δυνατότητα αυτή βοηθά στις περιπτώσεις όπου η διεργασία της μάθησης δεν ικανοποιεί την ιδιότητα Markov. γ) Τα συστήματα μάθησης ταξινομητών παρουσιάζουν και τη δυνατότητα της δομικής προσαρμογής μέσω του μηχανισμού ανακάλυψης κανόνων. Δυστυχώς, όμως, εξαιτίας της πολυπλοκότητάς του δεν έχει βρεθεί μαθηματικός τρόπος παρουσίασης συστημάτων μάθησης ταξινομητών ο οποίος να παρέχει θεωρητική σύγκλιση και εγγύηση βελτιστότητας.

Συνδυάζοντας μεθόδους με μοντέλο με μεθόδους χωρίς μοντέλο

Ο Sutton (1991) εισήγαγε την αρχιτεκτονική Dyna. Αυτή η αρχιτεκτονική που βασίζεται στο δυναμικό προγραμματισμό αξιοποιεί την εμπειρία με έναν πιο αποδοτικό τρόπο από τις μεθόδους χωρίς μοντέλο. Από τη μία πλευρά, αντί να χρησιμοποιεί τον κανόνα Q-μάθησης (13) ο Dyna κατασκευάζει ένα μοντέλο του περιβάλλοντος από την εμπειρία λειτουργίας σε αυτό και το χρησιμοποιεί προκειμένου να κάνει τις κατάλληλες ανανεώσεις. Από την άλλη όμως, ο Dyna συνδυάζει την Q-μάθηση με την ιδέα της επανάληψης αξίας, ώστε με λιγότερες εμπειρίες αλληλεπίδρασης με τον πραγματικό κόσμο να επιτυγχάνεται καλή συμπεριφορά. Η αλήθεια πάντως είναι ότι το κόστος των υπολογισμών που απαιτείται στην περίπτωση του Dyna είναι μεγαλύτερο από τις μεθόδους χωρίς μοντέλο.

Άμεση αναζήτηση πολιτικής

Μια άλλη σημαντική κλάση μεθόδων επίλυσης προβλημάτων ενισχυτικής μάθησης είναι η άμεση αναζήτηση στο διάστημα της πολιτικής. Αυτές οι μέθοδοι αποσκοπούν στον άμεσο εντοπισμό της κατάλληλης πολιτικής κάνοντας χρήση είτε μεθόδων κλίσης (gradient) είτε γενετικών αλγορίθμων, και δεν χρησιμοποιούν τη λογική του δυναμικού προγραμματισμού.

Το κύριο πλεονέκτημα των μεθόδων αυτών είναι η μεγαλύτερη γενίκευση. Από την άλλη ένα μειονέκτημα τους είναι ότι αυτή η πολιτική αναζήτησης απαιτεί περισσότερο χρόνο από τις κλασσικές μεθόδους. Αν χρησιμοποιήσουμε μεθόδους κλίσης έχουμε να αντιμετωπίσουμε τον κίνδυνο ότι παγιδεύονται σε τοπικά βέλτιστα, τα οποία δεν είναι βέβαιο ότι μας δίνουν καλές πολιτικές. Για να αντιμετωπισθεί το πρόβλημα της τοπικότητας μπορούμε να καταφύγουμε σε άλλες τεχνικές βελτιστοποίησης όπως οι γενετικοί αλγόριθμοι. Αν χρησιμοποιήσουμε όμως γενετικούς αλγορίθμους βρισκόμαστε μπροστά στο πρόβλημα της πολυπλοκότητα αυτών των μεθόδων που καθιστούν πολύ δύσκολη την εξαγωγή θεωρητικών αποτελεσμάτων και το υπολογιστικό κόστος των οποίων είναι μεγαλύτερο από τις μεθόδους κλίσης.

3.3 Η Πολυπρακτορική περίπτωση

Η Ενισχυτική Μάθηση ενός πράκτορα αποτελεί ένα ώριμο πεδίο με καλά θεωρητικά αποτελέσματα και αποδεδειγμένες πρακτικές εφαρμογές. Θα αποτελούσε αδιαμφισβήτητα μια ελκυστική λύση για το πρόβλημα της Πολυπρακτορικής μάθησης. Είναι όμως έτσι;

Η επέκταση από την ενισχυτική μάθηση ενός απλού πράκτορα στην Πολυπρακτορική Ενισχυτική Μάθηση, δυστυχώς δεν είναι εύκολη υπόθεση. Πρέπει να σκεφτούμε τι γίνεται με την ιδιότητα Markov. Η κύρια δυσκολία έγκειται στο ότι από την οπτική γωνία καθενός εκ των πρακτόρων στο σύστημα, το περιβάλλον παύει πλέον να έχει αυτή την ιδιότητα. Μέρος πλέον του περιβάλλοντος είναι και οι άλλοι πράκτορες. Όμως κάθε πράκτορας αποτελεί ένα δυναμικό σύστημα το οποίο αλλάζει τη συμπεριφορά του καθώς εκπαιδεύεται. Σε αυτή την περίπτωση τα αποτελέσματα μιας ενέργειας ενός πράκτορα δεν εξαρτώνται μόνο από την κατάσταση του περιβάλλοντος τη στιγμή που πραγματοποιήθηκε αλλά και από τις ενέργειες των άλλων πρακτόρων που εκτελέστηκαν την ίδια στιγμή. Αυτή η παραβίαση της υπόθεσης Markov δεν μας εξασφαλίζει πια τη θεωρητική σύγκλιση όπως στην περίπτωση της ενισχυτικής μάθησης ενός πράκτορα. Αυτή η παραβίαση όμως, δεν είναι δεν είναι το μόνο πρόβλημα που πρέπει να αντιμετωπίσουμε στην Πολυπρακτορική Ενισχυτική Μάθηση. Ένα άλλο

σημαντικό εμπόδιο είναι η εκθετική έκρηξη του διαστήματος καταστάσεων που συνεπάγεται η αύξηση του αριθμού των πρακτόρων.

Στο σημείο αυτό είναι σημαντικό να αναφέρουμε ότι υπάρχουν και κάποια σημαντικά οφέλη για τους πράκτορες κατά την εκπαίδευσή τους σε κάποιο πολυπρακτορικό σύστημα. Αυτά προέρχονται κυρίως από το διαμοιρασμό της γνώσης. Για παράδειγμα εάν διάφοροι πράκτορες μαθαίνουν να εκτελούν παρόμοιες εργασίες ο διαμοιρασμός των εμπειριών τους θα είχε σαν αποτέλεσμα την επίσπευση της διαδικασίας (Tan 1993). Αν πάλι κάποιος καινούριος πράκτορας έμπαινε στο σύστημα οι παλαιότεροι θα μπορούσαν να παίξουν το ρόλο του δασκάλου (Clouse 1995). Εάν δεν είναι επιθυμητή η μάθηση ο νεοεισερχόμενος θα μπορούσε να εκπαιδευτεί παρακολουθώντας και μιμούμενος κάποιους ειδικευμένους πράκτορες οι οποίοι εκτελούν τις εργασίες τους (Price & Boutilier, 2003). Η συμπεριφορά πολλών λογικών πρακτόρων που αλληλεπιδρούν έχει μελετηθεί κυρίως στη θεωρία παιγνίων. Σαν συνέπεια αυτού, μεγάλο μέρος της έρευνας της Πολυπρακτορικής Ενισχυτικής Μάθησης βασίζεται σε ιδέες αυτού του θεωρητικού πεδίου.

Τυπικό μοντέλο

Πολλές από τις ιδέες που θα μελετηθούν στη συνέχεια συναντώνται με διάφορα ονόματα στη βιβλιογραφία. Όπου κρίνεται απαραίτητο θα δίδονται εντός παρενθέσεων κάποιες εναλλακτικές ονομασίες.

Η ιδέα του παιγνίου Markov (Markov Game) αποτελεί τη βάση των περισσοτέρων πολυπρακτορικών μοντέλων ενισχυτικής μάθησης. Πριν ορίσουμε όμως ένα παίγνιο Markov, θα ορίσουμε την χωρίς-καταστάσεις έκδοσή του, το στρατηγικό παίγνιο (strategic game, matrix game).

Ορισμός Ένα στρατηγικό παίγνιο είναι μια πλειάδα $< N, \{A_i\}_{i \in N}, \{p_i\}_{i \in N} >$, όπου το N είναι το σύνολο των πρακτόρων, |N|=n είναι ο αριθμός τους, $\{A_i\}$ είναι τα διακριτά σύνολα των ενεργειών που είναι διαθέσιμες για κάθε πράκτορα, τα οποία επιφέρουν το σύνολο A που προκύπτει με συνδυασμό (joint) αυτών $\mathbf{A} = X_{i \in A} A_i$ και $\bar{p_i} : A \to R, i \in N$, είναι οι συναρτήσεις επιβράβευσης των πρακτόρων.

Οι πράκτορες εκτελούν κάποιες ενέργειες οι οποίες υποδεικνύονται από τις στρατηγικές $\sigma_i : A_i \rightarrow [0,1]$ και λαμβάνουν επιβραβεύσεις οι οποίες εξαρτώνται από την συνδυασμένη ενέργεια $a = [a_1,\ldots,a_n]^T : r_i = \bar{p_i}(a)$. Οι στρατηγικές μπορεί να είναι στοχαστικές (καλούνται επίσης τυχαιοποιημένες, randomized) ή ντετερμινιστικές

(καλούνται επίσης αγνές, pure). Αναφερόμαστε στη κοινή (joint) στρατηγική των πρακτόρων ως $\sigma = (\sigma_1,\ldots,\sigma_n)$, $\sigma \in \Pi(A)$, όπου το Π(.) δηλώνει το διάστημα των κατανομών πιθανότητας του συνόλου που δίδεται ως παράμετρος.

Κάθε συνάρτηση επιβράβευσης $\bar{p}_i$ μπορεί να γραφεί σαν ένας n-διάστατος πίνακας με δείκτες τις διακριτές ενέργειες και περιεχόμενο τις τιμές επιβράβευσης, γεγονός που δικαιολογεί το όνομα matrix game. Κάποιες φορές αναφερόμαστε στις επιβραβεύσεις στα στρατηγικά παίγνια με τον όρο *όφελος (payoff)* της θεωρίας παιγνίων.

Κάποια προβλήματα στην πολυπρακτορική μάθηση και συντονισμό μοντελοποιούνται ως επαναλαμβανόμενα στρατηγικά παίγνια. Είναι στρατηγικά παίγνια τα οποία παίζονται επαναληπτικά από τους ίδιους πράκτορες.

Ορισμός Ένα παίγνιο Markov (Markov Game ή αλλιώς στοχαστικό παίγνιο) είναι μια πλειάδα $< N, S, \{A_i\}_{i\in N}, p_s, \{p_i\}_{i\in N} >$, όπου:

- N είναι το σύνολο των πρακτόρων, |N|=n ο αριθμός τους
- S είναι το διακριτό σύνολο των καταστάσεων
- $\{A_i\}_{i\in N}$ είναι τα διακριτά σύνολα των ενεργειών που είναι διαθέσιμες σε κάθε πράκτορα, τα οποία αποδίδουν το σύνολο $\mathbf{A} = X_{i\in N} A_i$
- $p_s : S \times A \times S \rightarrow [0,1]$ είναι η κατανομή πιθανότητας της μετάβασης της κατάστασης
- $p_i: S \times A \times S \times R \rightarrow [0,1], i\in N$ είναι οι κατανομές πιθανότητας της επιβράβευσης των πρακτόρων

Σε κάθε βήμα στο χρόνο k, κάθε πράκτορας i παρατηρεί την κατάσταση s_k και εκτελεί κάποια ενέργεια $a_{i,k}$ όπως επιδεικνύεται από την πολιτική του $h_i : S \times A_i \rightarrow [0,1]$. Σαν αποτέλεσμα της συνδυαστικής ενέργειας $a_k = [a_{1,k},\ldots,a_{n,k}]^T$, ο κόσμος μεταβαίνει στην κατάσταση s_{k+1} με πιθανότητα $P(s_{k+1}|s_k, a_k) = p_s(s_k, a_k, s_{k+1})$ και κάθε πράκτορας i λαμβάνει μια επιβράβευση $r_{i,k+1}$ με πιθανότητα $P(r_{i,k+1}|s_k, a_k, s_{k+1}) = p_i(s_k, a_k, s_{k+1}, r_{i,k+1})$. Οι πολιτικές, με παρόμοιο τρόπο, μπορεί να είναι στοχαστικές ή ντετερμινιστικές. Μια σταθεροποιημένη (stationary) πολιτική είναι μία πολιτική η οποία δεν αλλάζει τις πιθανότητες επιλογής μιας ενέργειας καθώς περνά ο χρόνος. Δηλώνουμε τη συνδυαστική πολιτική των πρακτόρων ως $h = (h_1,\ldots,h_n)$.

Το παίγνιο Markov είναι μια επέκταση του στρατηγικού παιγνίου όπου έχουμε πολλαπλές καταστάσεις καθώς και στοχαστικές επιβραβεύσεις. Κάθε κατάσταση του

παιγνίου Markov είναι ένα διαφορετικό στρατηγικό παίγνιο με στοχαστικές επιβραβεύσεις τα οποία παίζονται από τους ίδιους πράκτορες. Πολλοί αλγόριθμοι Πολυπρακτορικής Ενισχυτικής Μάθησης για να αντιμετωπίσουν τα στοχαστικά παίγνια επιλύουν ξεχωριστά τα στρατηγικά παίγνια που προκύπτουν σε κάθε κατάσταση του στοχαστικού παιγνίου.

Παρόμοια, μια πολιτική είναι μια επέκταση της στρατηγικής σε πολλαπλές καταστάσεις και αντίθετα μια στρατηγική είναι μια απλούστευση της πολιτικής σε μια κατάσταση. Ένα παίγνιο Markov είναι η επέκταση μιας διαδικασίας λήψης απόφασης Markov και μια διαδικασία λήψης απόφασης Markov είναι ένα παίγνιο Markov με n=1. Μερικές φορές για να περιγράψουμε την εξέλιξη ενός πράκτορα σε ένα παίγνιο Markov δανειζόμαστε από τη θεωρία παιγνίων τον όρο παίξιμο (play). Έτσι στις περιπτώσεις που οι πράκτορες είναι ομογενείς, χρησιμοποιούν δηλαδή τον ίδιο αλγόριθμο μάθησης αναφερόμαστε στη διαδικασία μάθησης σαν αυτό-παίξιμο (self-play). Το αυτό-παίξιμο ορίζεται στη βιβλιογραφία στα πλαίσια των (επαναλαμβανόμενων) στρατηγικών παιγνίων, αλλά ο ορισμός που δώσαμε προηγουμένως αναφέρεται σε αυτό σαν ειδική περίπτωση.

Εμείς θα επικεντρωθούμε κυρίως σε δύο ειδικές περιπτώσεις παιγνίων Markov: α) όταν οι πράκτορες έχουν έναν κοινό στόχο (δηλ. την ίδια συνάρτηση επιβράβευσης) και β) όταν δύο πράκτορες δρουν ο ένας εναντίον του άλλου.

Ορισμός Μια πολυπρακτορική διαδικασία λήψης απόφασης Markov (multiagent Markov decision process, πλήρως συνεργατικό παίγνιο – fully cooperative game) είναι ένα παίγνιο Markov $< N, S, \{A_i\}_{i \in N}, p_s, \bar{p} >$ όπου όλοι οι πράκτορες μοιράζονται την ίδια συνάρτηση επιβράβευσης $\bar{p}$

Ορισμός Ένα πλήρως ανταγωνιστικό παίγνιο (fully competitive game) είναι ένα παίγνιο Markov δύο παικτών $< \{n_1, n_2\}, S, A_1, A_2, p_s, \bar{p_1}, \bar{p_2} >$ όπου

$\bar{p_1}(s, a_1, a_2) = -\bar{p_2}(s, a_1, a_2), \forall s \in S, a_1 \in A_1, a_2 \in A_2$

Το όνομα πολυπρακτορική διαδικασία λήψης απόφασης Markov δικαιολογείται ως εξής: Εάν όλοι οι πράκτορες θεωρηθούν σαν ένα άτομο που λαμβάνει τις αποφάσεις, τότε η πολυπρακτορική διαδικασία λήψης απόφασης Markov εκπίπτει σε μια διαδικασία λήψης απόφασης Markov με ένα διάστημα ενέργειας που δίδεται από το συνδυασμό των διαστημάτων ενέργειας της πολυπρακτορικής διαδικασίας. Ένα σημαντικό κομμάτι της

δουλειάς στην πολυπρακτορική μάθηση και συντονισμό βασίζεται σε αυτόν τον τύπο παιγνίου Markov.

Ο Kok (2005b) χρησιμοποίησε μια διαφορετική έκδοση της πολυπρακτορικής διαδικασίας λήψης απόφασης Markov.

Ορισμός Μια πολυπρακτορική διαδικασία λήψης απόφασης Markov είναι ένα στοχαστικό παίγνιο $< N, S, \{A_i\}_{i \in N}, p_s, \{\bar{p}_i\}_{i \in N}, \bar{p} >$, όπου $\bar{p}$ είναι η κοινή (global) συνάρτηση επιβράβευσης και δίδεται από τον τύπο

$$\bar{p}(s,a) = \sum_{i \in N} \bar{p}_i(s,a), \forall s \in S, a \in A \quad (14)$$

Ορισμός ισοδύναμος με τον παραπάνω διότι η πολυπρακτορική διαδικασία λήψης απόφασης προσπαθεί να μεγιστοποιήσει την κοινή επιβράβευση σε κάθε βήμα. Το πλήρως ανταγωνιστικό παίγνιο καλείται και μηδενικού αθροίσματος επειδή οι επιβραβεύσεις των δύο πρακτόρων πάντα έχουν σαν άθροισμα την τιμή 0.

Ιδέες επίλυσης

Οι απόψεις των διαφόρων ερευνητών που μελετούν τα πολυπρακτορικά συστήματα ενισχυτικής μάθησης δυστυχώς δεν ταυτίζονται ως προς τον στόχο μάθησης που θα έπρεπε να έχουν αυτά τα συστήματα. Ο θεμέλιος λίθος πάνω στον οποίο βασίζονται οι σκέψεις επίλυσης των πολυπρακτορικών συστημάτων ενισχυτικής μάθησης είναι η συνάρτηση αξίας όπως και στην περίπτωση της ενισχυτικής μάθησης ενός πράκτορα. Οι συναρτήσεις αξίας κατάστασης και ενέργειας ορίζονται παρόμοια με την περίπτωση του ενός πράκτορα, με τη διαφορά ότι λαμβάνονται υπόψη οι πολιτικές όλων των πρακτόρων. Οι ορισμοί δίδονται στη συνέχεια με τη χρήση του φθίνοντος προσδοκώμενου αποτελέσματος.

Ορισμός Η αξία μιας κατάστασης s του πράκτορα i υπό την συνδυαστική πολιτική h είναι το προσδωκόμενο αποτέλεσμα του πράκτορα i όταν τον περιβάλλον ξεκινά από την s και οι πράκτορες ακολουθούν στη συνέχεια την h

$$V_i^h(s) = E_h\left\{\sum_{l=0}^{\infty} \gamma^l r_{i,k+l+1} \mid s_k = s\right\} \quad (15)$$

Ορισμός Η αξία της συνδυαστικής ενέργειας a στην κατάσταση s για τον πράκτορα i υπό την συνδυαστική πολιτική h είναι το προσδοκώμενο αποτέλεσμα όταν το περιβάλλον ξεκινά από την s, οι πράκτορες εκτελούν την a και ακολουθούν στη συνέχεια την h

$$Q_i^{\ h}(s,a) = E_h\left\{\sum_{l=0}^{\infty}\gamma^l r_{i,k+l+1} \mid s_k = s, a_k = a\right\} \quad (16)$$

Όπως αναφέραμε σε προηγούμενη ενότητα, πολλοί αλγόριθμοι πολυπρακτορικής ενισχυτικής μάθησης καταφεύγουν στην επίλυση των στρατηγικών παιγνίων που προκύπτουν σε κάθε κατάσταση του παιγνίου Markov. Η επιλογή αυτή έχει σαν αποτέλεσμα οι αξίες κατάστασης $V_i^{\mathbf{h}}(s)$ υπό τις συνδυαστικές πολιτικές να εκπίπτουν σε αξίες $V_i^{\boldsymbol{\sigma}}$ υπό τις συνδυαστικές στρατηγικές, καθώς η ιδέα της κατάστασης χάνει την αξία της

$$V_i^{\ s} = E_s\{r_i\} \quad (17)$$

Η συνδυαστική στρατηγική σ των πρακτόρων καλείται επίσης και 'προφίλ στρατηγικής'. Δηλώνουμε την συνδυαστική στρατηγική, όλων, πλην του πράκτορα i, ως $\sigma_{-i} = (\sigma_1,\ldots, \sigma_{i-1}, \sigma_{i+1},\ldots, \sigma_n)$ (περιορισμένο προφίλ στρατηγικής – reduced strategy profile). Η σ_i^* είναι η καλύτερη απάντηση (best response) του πράκτορα i στο περιορισμένο προφίλ στρατηγικής σ_{-i}, εάν $V_i^{(s_i,s_{-i})} \le V_i^{(s_i^*,s_{-i})}, \forall \sigma_t \in \Pi(A_i)$. Η ιδέα της θεωρίας παιγνίων που συναντάται πιο συχνά στην πολυπρακτορική μάθηση είναι η Nash-ισορροπία (Nash equilibrium).

Ορισμός Η Nash-ισορροπία είναι μια συνδυαστική στρατηγική σ* τέτοια ώστε, για κάθε έναν πράκτορα i, η σ_i^* είναι η καλύτερη απάντηση στην σ_{-i}^*.

Στα πλήρως ανταγωνιστικά παίγνια, η Nash-ισορροπία αποκτά μια ειδική σημασία. Κάθε στρατηγική σ_1 αξιολογείται με βάση την αντίπαλη στρατηγική σ_2, η οποία σε συνδυασμό με την σ_1 επιφέρουν την ελάχιστη αξία. Ο πράκτορας λειτουργεί έτσι ώστε να μεγιστοποιεί το όφελός του στη χειρότερη περίπτωση:

$$\sigma_1^* = \arg\max_{\sigma_1\in\Pi(A_1)} \min_{\sigma_2\in\Pi(A_2)} V_1^{(\sigma_1,\sigma_2)} \quad (18)$$

Αυτή η αρχή καλείται minimax (Littman,2001b). Παραμένει άγνωστο εάν οι Nash-ισορροπίες μπορούν να υπολογιστούν σε πολυωνυμικό χρόνο. Για το λόγω αυτό χρησιμοποιούμε μια πιο γενική κλάση ισορροπιών που καλούνται συσχετισμένες ισορροπίες (correlated equilibria) οι οποίες υπολογίζονται σε πολυωνυμικό χρόνο με τον γραμμικό προγραμματισμό.

Στον γενικό τους ορισμό τα παίγνια Markov είναι συμμετρικά. Κανένας πράκτορας δεν αντιμετωπίζεται ευνοϊκά έναντι άλλων. Εάν όμως επιτραπεί σε κάποιους πράκτορες (ηγέτες – leaders) να κατέχουν την πληροφορία για το πώς οι άλλοι πράκτορες (οπαδοί

– followers) θα δράσουν, αναπηδά μια νέα τάξη ισορροπιών. Για λόγους απλότητας θα περιορίσουμε τον ορισμό στους δύο πράκτορες (Basar, 1985).

Ορισμός Ένα ζεύγος στρατηγικών (σ1*, σ2*) είναι μια Stackelberg ισορροπία (Stackelberg equilibrium) με διακριτές απαντήσεις του οπαδού των στρατηγικών παιγνίων δύο παικτών $<\{n_1,n_2\},A_1,A_2,\bar{p_1},\bar{p_2}>$ εάν υπάρχει μια μοναδική αντιστοιχία T: Π(A1)→Π(A2) η οποία ικανοποιεί τις συνθήκες:

$$V_2^{(\sigma_1,\mathrm{T}(\sigma_1))} \geq V_2^{(\sigma_1,\sigma_2)}, \forall \sigma_1 \in \Pi(A_1), \sigma_2 \in \Pi(A_2) \quad (19)$$

$$V_1^{(\sigma_1^*,\mathrm{T}(\sigma_1^*))} \geq V_1^{(\sigma_1,\mathrm{T}(\sigma_1))}, \forall \sigma_1 \in \Pi(A_1) \quad (20)$$

με $\sigma_2^*=\mathrm{T}(\sigma_1^*)$. Ο πράκτορας 1 είναι ο ηγέτης και ο πράκτορας 2 ο οπαδός. Η αντιστοιχία T περιγράφει πως ο οπαδός αντιδρά στις ενέργειες του ηγέτη. Η συνθήκη (19) διασφαλίζει ότι ο λογικός οπαδός θα υπακούσει στην αντιστοιχία T και η συνθήκη (20) αποτελεί τη συνθήκη της ισορροπίας. Ένας άλλος τρόπος επίλυσης είναι η ιδέα της μετάνοιας (regret). Η μετάνοια μετρά τη διαφορά μεταξύ της μέγιστης συνολικής επιβράβευσης η οποία θα μπορούσε να επιτευχθεί από κάθε σταθερή ντετερμινιστική πολιτική και της πραγματικής επιβράβευσης που δέχεται ο πράκτορας

$$R_{i,k} = \max_{\bar{h}_i} \sum_{l=0}^{k-1} [p_i(s_l, \bar{h}_i(s_l)) - r_{i,l+1}] \quad (21)$$

όπου θεωρούμε ότι λαμβάνουμε ντετερμινιστικές επιβραβεύσεις προκειμένου να απλοποιήσουμε την περιγραφή. Μια καλή απόδοση του πράκτορα συνδέεται με μικρή (ή αρνητική) μετάνοια.

Στόχος μάθησης

Πολλοί αλγόριθμοι Πολυπρακτορικής Ενισχυτικής Μάθησης θέτουν σαν στόχο μάθησης για τον πράκτορα τη σύγκλιση σε μια ισορροπία της θεωρίας παιγνίων, και πιο συχνά σε μια Nash-ισορροπία (Littman 2001b, Hu & Wellman 2003, Greenwald & Hall 2003). Αυτό φαίνεται φυσιολογικό καθώς η Nash-ισορροπία αποτελεί τη ‘φυσική’ μακροχρόνια συνέπεια των επιλογών ενός λογικού πράκτορα.

Οι Shoham κ.ά.(2003) από την άλλη μεριά, ασκούν έντονη κριτική για τη χρήση των ισορροπιών γενικά, και της Nash-ισορροπίας ειδικότερα. Συμφωνούν ότι η Nash-ισορροπία παρουσιάζει κάποια σημαντικά προβλήματα:

- ✓ Γενικά, ένα στρατηγικό παίγνιο μπορεί να έχει πολλές Nash-ισορροπίες. Αυτό οδηγεί σε δύσχρηστες εγγυήσεις σύγκλισης, διότι απαιτεί από τους πράκτορες να

συντονίσουν με κάποιο τρόπο την επιλογή τους μεταξύ των ισορροπιών, ίσως και με κάποιον εξωτερικό μηχανισμό.

- ✓ Η έννοια και η επιθυμία της Nash-ισορροπίας, ορισμένη σε όρους παιγνίων χωρίς καταστάσεις, είναι αμφίβολες στην περίπτωση του πλήρους παιγνίου Markov, όπου η καθυστερημένη επιβράβευση έχει έναν σημαντικό ρόλο.

Αυτή η αμφισβήτηση επεκτείνεται σε όλες τις προτάσεις που χρησιμοποιούν κάποια ισορροπία. Το πρόβλημα της επιλογής της ίδιας ισορροπίας μεταξύ πολλών εναλλακτικών σε μια κατάσταση του παιγνίου Markov είναι ένα τρέχον θέμα στην πολυπρακτορική ενισχυτική μάθηση και είναι γνωστό ως το πρόβλημα επιλογής ισορροπίας (equilibrium selection problem). Ο Shoham κ.ά. (2003) τονίζουν ότι αυτή η αδικαιολόγητη εστίαση στις Nash-ισορροπίες είναι αποτέλεσμα της έλλειψης μιας ξεκάθαρης περιγραφής του προβλήματος ενισχυτικής μάθησης στην περίπτωση ύπαρξης πολλών πρακτόρων.

Μια πρώτη προσπάθεια περιγραφής είχε γίνει από τους Bowling και Veloso (2002) στον ορισμό των κριτηρίων της *λογικότητας* (rationality) και της *σύγκλισης* (convergence). Ένας αλγόριθμος μάθησης είναι 'λογικός', εάν, δεδομένου ότι όλοι οι άλλοι πράκτορες έχουν συγκλίνει σε σταθεροποιημένες πολιτικές, ο μαθητευόμενος πράκτορας συγκλίνει σε μια πολιτική η οποία είναι η καλύτερη απάντηση σε αυτές τις σταθεροποιημένες πολιτικές. Ένας αλγόριθμος μάθησης είναι συγκλίνων εάν, δεδομένου ότι οι άλλοι πράκτορες χρησιμοποιούν αλγορίθμους μάθησης από ένα δεδομένο σύνολο, ο μαθητευόμενος πράκτορας θα συγκλίνει σε μια σταθεροποιημένη πολιτική. Είναι επιθυμητό οι αλγόριθμοι μάθησης να είναι και λογικοί και συγκλίνοντες.

Οι Powers και Shoham (2004) διατύπωσαν κάποιες επιφυλάξεις στο κατά πόσον αυτά τα κριτήρια θα μπορούσαν να επιτευχθούν. Τα επιχειρήματά τους:

- ✓ οι αλγόριθμοι απαιτούν αδικαιολόγητα και ο μαθητευόμενος και οι άλλοι πράκτορες να συγκλίνουν σε σταθεροποιημένες πολιτικές, ενώ και μη-σταθεροποιημένες πολιτικές θα μπορούσαν να ήταν ενδιαφέρουσες για τον σχεδιαστή του συστήματος
- ✓ και η λογικότητα και η σύγκλιση απαιτούνται να ισχύουν στο όριο, χωρίς να περιέχουν εγγυήσεις για ικανοποιητική απόδοση σε πεπερασμένο χρόνο

Επιπρόσθετα οι συγγραφείς συμφωνούν ότι και οι δύο ιδιότητες ορίζονται με βάση την πολιτική του πράκτορα και όχι την επιβράβευση που αποτελεί το πραγματικό μέτρο της απόδοσης.

Προτείνουν στη συνέχεια τρία νέα κριτήρια σαν στόχο μάθησης για τους αλγόριθμους πολυπρακτορικής ενισχυτικής μάθησης. Θεωρώντας επαναλαμβανόμενα παίγνια, ο μαθητευόμενος πράκτορας είναι απαραίτητο με μεγάλη πιθανότητα και σε περιορισμένο χρόνο να επιτυγχάνει τα ακόλουθα:

- ✓ εστιασμένη βελτιστότητα (targeted optimality): όταν οι αλγόριθμοι μάθησης των άλλων πρακτόρων είναι σε ένα δεδομένο σύνολο, μια μέση επιβράβευση η οποία είναι αυθαίρετα κοντά στην αξία καλύτερης απάντησης
- ✓ συμβατότητα (compatibility): στο αυτό-παίξιμο, μια μέση επιβράβευση η οποία είναι αυθαίρετα κοντά στην αξία της καλύτερης Nash-ισορροπίας
- ✓ ασφάλεια (safety): όταν οι άλλοι πράκτορες χρησιμοποιούν οποιονδήποτε άλλον αλγόριθμο μάθησης, μια μέση επιβράβευση η οποία είναι αυθαίρετα κοντά στην αξία minimax (αξία χειρότερης περίπτωσης)

Αυτά τα κριτήρια απαλλάσσουν εντελώς το στόχο μάθησης πό τις απαιτήσεις σύγκλισης. Έχουν κι αυτά όμως ένα μικρό πρόβλημα, θέτουν απαιτήσεις στη μέση επιβράβευση. Αυτό σημαίνει ότι σε κάθε δεδομένη στιγμή στο χρόνο, η απόδοση του πράκτορα μπορεί να είναι αυθαίρετα φτωχή.

Για να ξεπεράσουμε αυτό το πρόβλημα ο Bowling (2004) εισήγαγε την απαίτηση της μη-μετάνοιας (no regret) : εάν ένας πράκτορας δεν συγκλίνει σε μια σταθεροποιημένη πολιτική, τότε η μετάνοιά του (20) θα έπρεπε να είναι αρνητική ή μηδέν. Η σύγκλιση, όμως, παρέμεινε ένα επιθυμητό χαρακτηριστικό του αλγορίθμου μάθησης, επειδή οδηγεί σε σταθερότητα την εξέλιξη του παιγνίου Markov, και η σταθερότητα συνεπάγεται ακρίβεια στις εκτιμήσεις των συναρτήσεων αξίας. Όταν έχουμε καθυστερημένη επιβράβευση, η ικανότητα της ικανοποιητικής εκτίμησης των συναρτήσεων αξίας είναι πολύ σημαντική για τον πράκτορα.

Φαίνεται λοιπόν ότι η σύγκλιση και η επίτευξη καλών επιβραβεύσεων δεν είναι δύο αμοιβαία αποκλειόμενοι στόχοι. Υπάρχει κάποια σύνδεση αυτών των δύο. Για να επιτύχουμε υψηλές επιβραβεύσεις, απαιτείται ο πράκτορας να προβλέψει με ακρίβεια τη συνάρτηση αξίας, αλλά για να το επιτύχει αυτό, είναι απαραίτητη η σταθερότητα, γεγονός που σημαίνει ότι η διαδικασία μάθησης του πράκτορα πρέπει να συγκλίνει.

Σε κάθε περίπτωση, η συζήτηση για τον ακριβή στόχο της πολυπρακτορικής ενισχυτικής μάθησης δεν έχει ολοκληρωθεί και δεν έχουν δοθεί ακόμα οριστικές απαντήσεις.

<u>Εφαρμογές των τεχνικών ενός πράκτορα στην Πολυπρακτορική Ενισχυτική Μάθηση</u>

Η αλήθεια είναι ότι ο πιο απλός τρόπος να αντιμετωπίσεις τις συνέπειες της παρουσίας άλλων πρακτόρων στη διαδικασία μάθησης είναι να τους αγνοήσεις. Η προσέγγιση αυτή έχει επιφέρει κάποια καλά αποτελέσματα σε ποικιλία προβλημάτων που κυμαίνονται από απλές προσομοιώσεις σε πραγματικές πολύπλοκες εργασίες.

Ο Sen κ.ά. (1994) σε μια εργασία τους βασίστηκαν σε αυτή την προσέγγιση. Πιο συγκεκριμένα, δύο πράκτορες οι οποίοι ακολουθούσαν Q-μάθηση, εκπαιδεύονταν σε συμπληρωματικές πολιτικές προκειμένου να μετακινήσουν ένα κουτί σε μια επιφάνεια δύο διαστάσεων. Οι συγγραφείς εδώ χρησιμοποίησαν έναν πολύ συγκεκριμένο τύπο προβλήματος όπου η κατάσταση-στόχος είχε την ίδια οριζόντια συντεταγμένη με την αρχική κατάσταση. Με τον τρόπο αυτό, μπορούσαν να ποσοτικοποιήσουν την πληροφορία θέσης και να παρέχουν στιγμιαία ενίσχυση στους πράκτορες βασιζόμενοι στην οριζόντια απόσταση μεταξύ της τρέχουσας κατάστασης-θέσης του κουτιού και της κατάστασης στόχου. Απλοποίησαν δηλαδή το πρόβλημα της ενισχυτικής μάθησης με δύο τρόπους: πρώτον, ο αριθμός των καταστάσεων του διαστήματος των καταστάσεων περιορίστηκε στο μισό (από 2 σε 1) και δεύτερον, εξαλείφθηκε το πρόβλημα της καθυστερημένης επιβράβευσης. Η απλοποίηση αυτή δεν μας επιτρέπει να συμπεράνουμε αν αυτή η προσέγγιση θα μπορούσε να επεκταθεί σε πιο περίπλοκα προβλήματα, γεγονός που αποτελεί χαρακτηριστικό όλων των εργασιών που χρησιμοποιούν τεχνικές ενός πράκτορα σε απλές πολυπρακτορικές προσομοιώσεις.

Ένα μεγάλο μέρος των εργασιών που έχουν γίνει για πιο πολύπλοκα προβλήματα προέρχονται από το πεδίο των συστημάτων πολλών ρομπότ. Σε μια από αυτές ο Mataric (1996) παρουσίασε τα ερευνητικά του αποτελέσματα από μια εργασία αναζήτησης αντικειμένων πολλών ρομπότ, όπου τα ρομπότ εκπαιδεύονταν να συλλέγουν αντικείμενα τα οποία ήταν διεσπαρμένα και να τα μεταφέρουν σε μια συγκεκριμένη περιοχή (το σπίτι). Τα διαστήματα των καταστάσεων και των ενεργειών ήταν αφηρημένα σε μεγάλο βαθμό, έτσι αυτό που χρειάζονταν οι πράκτορες ήταν να μάθουν να αντιστοιχούν έναν μικρό αριθμό συνθηκών υψηλού επιπέδου (όπως π.χ. διατήρηση_αντικειμένου) σε έναν μικρό αριθμό συμπεριφορών υψηλού επιπέδου (όπως π.χ. τοποθέτηση_στο_σπίτι). Το σήμα της ενίσχυσης συντίθεται με έναν περίπλοκο τρόπο από διάφορους ξεχωριστούς στόχους και στιγμιαία στοιχεία που ονομάζονται εκτιμητές προόδου.

Μια άλλη αντιπροσωπευτική εργασία είναι η εφαρμογή ποδοσφαίρου από ρομπότ των Bonarini και Trianni (2001). Σε αυτή την εργασία η πληροφορία κατάστασης αναπαρίσταται με ασαφείς γλωσσικές μεταβλητές. Για παράδειγμα, ο πράκτορας

γνωρίζει εάν υπάρχουν ένας ή περισσότεροι συμπαίχτες γύρω του, με τον πλησιέστερο κοντά δεξιά και τον πιο απομακρυσμένο μακριά αριστερά (με τις γλωσσικές τιμές “zero”, “one”, “more”, “close”, “far”, “left”, “right”). Αυτό επαναλαμβάνεται και για τους πράκτορες της αντίπαλης ομάδας. Παρόμοιες τιμές χρησιμοποιούνται και για την μπάλα και την εσωτερική κατάσταση του πράκτορα.

Ο Crites και Barto (1998) εφάρμοσαν την ενισχυτική μάθηση στην εργασία του προγραμματισμού του ανελκυστήρα. Το μοντέλο αυτής της εργασίας είναι ένα σύστημα διακριτών γεγονότων με συνεχή χρόνο. Αν και το σήμα ενίσχυσης είναι κοινό, έχοντας σαν αποτέλεσμα μια εργασία με κοινό στόχο, διαφορετικοί ελεγκτές αποδίδονται σε κάθε ανελκυστήρα. Επιπρόσθετα η μερική παρατηρησιμότητα καθιστά πιο πολύπλοκη την εργασία (κάποια στοιχεία της κατάστασης είναι κρυφά, π.χ. οι αποστάσεις των χρηστών που περιμένουν σε κάθε πάτωμα). Οι ερευνητές χρησιμοποίησαν αποτελέσματα που ορίστηκαν σαν συνεχή διαστήματα τιμών και νευρωνικά δίκτυα για να αναπαραστήσουν τη συνάρτηση αξίας. Τα αποτελέσματα που παρήχθησαν κατά τις προσομοιώσεις ξεπέρασαν αυτά των αλγορίθμων των εφαρμογών προγραμματισμού του εμπορίου.

Χαρακτηριστικό όλων αυτών των πολυπρακτορικών εφαρμογών είναι η χρήση εμπειρικών και ευριστικών μηχανισμών προκειμένου να είναι εφικτή η εκπαίδευση των πρακτόρων σε αυτά τα πολύπλοκα προβλήματα, γεγονός που καθιστά όμως δύσκολη την ανάλυση και χρησιμότητα των τεχνικών από άλλους τομείς.

Πλήρως συνεργάσιμες πολυπρακτορικές ομάδες

Το θεωρητικό μοντέλο το οποίο περιγράφει το πλήρως συνεργάσιμο σκηνικό είναι αυτό της πολυπρακτορικής διαδικασίας λήψης απόφασης Markov. Σε αυτό το μοντέλο η βέλτιστη λύση μπορεί να βρεθεί αν θεωρήσουμε το πολυπρακτορικό σύστημα σαν έναν απλό πράκτορα και αν εκτιμήσουμε τις βέλτιστες συνδυαστικές αξίες ενέργειας με την Q-μάθηση. Αυτή η προσέγγιση επιφέρει τη βέλτιστη πολιτική αν απλώς στη συνέχεια ακολουθήσουμε άπληστη επιλογή ενέργειας. Σε αυτή την περίπτωση όμως έχουμε να αντιμετωπίσουμε το πρόβλημα ότι στα πολυπρακτορικά συστήματα οι πράκτορες έχουν πάντα κάποιο βαθμό αυτονομίας κατά την επιλογή των ενεργειών τους. Μια λύση θα μπορούσε να ήταν να αντιγράψουμε τις συναρτήσεις αξίας και τον αλγόριθμο μάθησης σε κάθε πράκτορα. Αυτό είναι εφικτό καθόσον οι ενέργειες μεταξύ των πρακτόρων είναι μετρήσιμες. Ο αλγόριθμος που βασίζεται σε αυτή τη λύση είναι γνωστός στη βιβλιογραφία ως “ομάδα-Q” (team-Q) ή “φίλος-Q” (friend-Q) (Littman 2001b). Υπάρχει όμως μια επιπλέον δυσκολία που γεννιέται από την κατανεμημένη φύση της διαδικασίας

λήψης απόφασης. Για κάποια δεδομένη κατάσταση του κόσμου, είναι πιθανό να υπάρχουν περισσότερες της μια συνδυαστικές ενέργειες οι οποίες έχουν σαν αποτέλεσμα την καλύτερη Q-αξία. Η κλασσική άπληστη επιλογή ενέργειας θα επίλυε τις ισοπαλίες με τυχαίο τρόπο. Εάν όμως ο κάθε πράκτορας επίλυε τις ισοπαλίες με τυχαίο τρόπο θα είχαμε διάφορους τρόπους επίλυσής τους, γεγονός που θα οδηγούσε τους πράκτορες στην εκτέλεση διαφορετικών συνδυαστικών ενεργειών, που θα είχαν σαν αποτέλεσμα η τελική συνδυαστική ενέργεια να μην είναι βέλτιστη. Αυτό αποτελεί μια ειδική περίπτωση του προβλήματος επιλογής ισορροπίας.

Είναι απαραίτητο να αναφέρουμε στο σημείο αυτό ότι ο Littman (2001b) διασφάλιζε τη σύγκλιση μόνο της συνάρτησης αξίας. Οι Q-συναρτήσεις που χρησιμοποιούν οι πράκτορες θα συγκλίνουν στη βέλτιστη Q*. Επιπρόσθετα σύγκλιση των πολιτικών έχουμε μόνο στην περίπτωση που οι βέλτιστες συνδυαστικές ενέργειες που επιτυγχάνουν τη Q* είναι μοναδικές σε κάθε κατάσταση.

Εν συντομία, η ομάδα-Q απαιτεί τις ακόλουθες υποθέσεις:

- το παίγνιο Markov να είναι πλήρως συνεργατικό
- οι ενέργειες να είναι μετρήσιμες μεταξύ των πρακτόρων
- οι βέλτιστες συνδυαστικές ενέργειες είναι μοναδικές σε κάθε κατάσταση του κόσμου

Οι Lauer και Riedmiller (2000) παρουσίασαν έναν αλγόριθμο τον οποίο αποκαλούσαν ‘κατανεμημένη Q-μάθηση’ (distributed Q-learning) και ο οποίος αφαιρούσε τη δεύτερη υπόθεση. Ένας πράκτορας i διατηρεί έναν Q-πίνακα $Q_i(x,u_i)$, όπου παρατηρούμε ότι δείκτης υπάρχει μόνο στη δικιά του ενέργεια και χρησιμοποιεί έναν τροποποιημένο κανόνα ανανέωσης της χρονικής διαφοράς:

$$Q_i(s_k, a_k) \leftarrow \max\{Q_i(s_k, a_{i,k}), r_{k+1} + \gamma \max_{a_i' \in A_i} Q_i(s_{k+1}, a'_i)\} \quad (22)$$

Οι συγγραφείς απέδειξαν ότι με αυτό τον κανόνα ανανέωσης υπό τις παραδοχές ότι η συνάρτηση επιβράβευσης είναι θετική και ότι όλες οι Q-αξίες αρχικοποιούνται στην τιμή 0, οι Q-αξίες που λαμβάνουν οι πράκτορες είναι οι μέγιστες των συνδυαστικών Q-αξιών:

$$Q_i(s, a_i) = \max_{\substack{a=[a_1,\ldots a_n]^T \\ a_j \in A,\, j \neq i}} Q(s,a), \forall s \in S, a_i \in A_i \quad (23)$$

Αν και η απαίτηση, η συνάρτηση επιβράβευσης να είναι θετική, φαντάζει περίεργη, δεν περιορίζει στην πραγματικότητα τη γενικότητα του αλγορίθμου. Οι πράκτορες διατηρούν

ακριβείς εκτιμήσεις $\tilde{h}_i$ της βέλτιστης πολιτικής. Με το να διαφοροποιήσουν την άπληστη πολιτική έτσι ώστε οι τροποποιήσεις να γίνονται αποδεκτές στην περίπτωση και μόνο που συμβάλλουν στη βελτίωση των Q-αξιών

$$\tilde{h}_i(s_k) \leftarrow \begin{cases} \tilde{h}_i(s_k) \text{ εάν το } \max_{a_i \in A_i} Q_i(s_k, a_i) \text{ δεν έχει τροποποιηθεί από (22)} \\ a_{i,k} \quad \text{διαφορετικά} \end{cases} \tag{24}$$

οι συγγραφείς απέδειξαν ότι η συνδυαστική πολιτική $\tilde{h} = (\tilde{h}_1, ..., \tilde{h}_n)$ πάντα έχει την άπληστη τιμή βάσει του συνδυαστικού Q-πίνακα. Στην (24) η $u_{i,k}$ είναι η ενέργεια που διενεργείται από τον πράκτορα τη χρονική στιγμή k. Καθώς η κατανεμημένη Q-μάθηση είναι off-policy και οι πράκτορες απαιτείται να ακολουθούν εξερευνητικές πολιτικές, αυτή η ενέργεια δεν είναι απαραίτητο πάντα να εκτελεστεί σύμφωνα με την $\tilde{h}_i$ και η μάθηση λαμβάνει χώρα στην (24). Τέλος οι συγγραφείς έδειξαν ότι η σύγκλιση της κατανεμημένης Q-μάθησης απορρέει από τη βασική Q-μάθηση.

Γενικά Πολυπρακτορικά Συστήματα

Η υπόθεση της πλήρους συνεργατικότητας είναι πολύ περιοριστική στις περισσότερες των περιπτώσεων. Αυτό είναι ιδιαίτερα εμφανές στην περίπτωση των ιδιοτελών πρακτόρων. Αλλά ακόμα και στις περιπτώσεις των ομάδων πρακτόρων που συνεργάζονται, μερικές φορές μπορεί τα στιγμιαία ενδιαφέροντα κάποιων να συγκρούονται. Αυτό θα μπορούσε να είναι ένας κοινός πόρος, για παράδειγμα. Αν και οι στόχοι σε υψηλό επίπεδο είναι κοινοί, οι πράκτορες θα συναγωνιστούν για να αποκτήσουν την ίδια στιγμή τον μοιραζόμενο πόρο. Το θεωρητικό μοντέλο που περιγράφει αυτού του τύπου τις εργασίες είναι αυτό του παιγνίου Markov.

Οι αλγόριθμοι που καλούνται να αντιμετωπίσουν τέτοιες καταστάσεις κάνουν κάποιες υποθέσεις:

i. Οι ενέργειες είναι μετρήσιμες μεταξύ των πρακτόρων
ii. Οι επιβραβεύσεις είναι μετρήσιμες μεταξύ των πρακτόρων

Κάποιοι αλγόριθμοι προσθέτουν κι άλλες, τις οποίες θα αναφέρουμε όπου είναι χρήσιμο. Οι πράκτορες στις περισσότερες των περιπτώσεων μαθαίνουν τις συναρτήσεις αξίας βασιζόμενοι στη συνδυαστική ενέργεια και χρησιμοποιούν

στοχαστικές πολιτικές. Το τελευταίο εξηγείται από την ύπαρξη των παιγνίων Markov για τα οποία οι λύσεις-στόχοι δεν μπορούν να εκφραστούν με ντετερμινιστικές πολιτικές.

Πολλές προσεγγίσεις που σχετίζονται με τα γενικά Πολυπρακτορικά Συστήματα προέρχονται από το πεδίο της Θεωρίας Παιγνίων

Προβλήματα Δίχως Καταστάσεις (Stateless Problems)

Υπάρχουν προσεγγίσεις Πολυπρακτορικής Ενισχυτικής Μάθησης που προέρχονται από τη Θεωρία Παιγνίων, οι οποίες σχετίζονται μόνο με επαναλαμβανόμενα στρατηγικά παίγνια. Ως τέτοια δεν ορίζουν την ιδέα της κατάστασης, και χάνουν ένα από τα βασικά χαρακτηριστικά της ενισχυτικής μάθησης: την καθυστερημένη επιβράβευση. Παρόλα αυτά, οι προσεγγίσεις αυτές είναι σχετικές με το θέμα της Πολυπρακτορικής Ενισχυτικής Μάθησης και μάλιστα μια από αυτές (η απειροστική βαθμιαία κλίση – infinitesimal gradient ascent) αποτελεί τη βάση για μια πιο γενική μέθοδο. Όλοι οι αλγόριθμοι αυτής της ενότητας ενισχύουν την υπόθεση (ii) ως εξής:

iii. Οι συναρτήσεις επιβράβευσης των πρακτόρων είναι κοινή γνώση (τυπικά κοινή γνώση σημαίνει ότι οι πράκτορες γνωρίζουν το γεγονός, όλοι οι πράκτορες γνωρίζουν ότι οι άλλοι πράκτορες γνωρίζουν το γεγονός και ούτω καθ' εξής επ' άπειρον).

Επιπρόσθετα εισάγουν και μια νέα υπόθεση:

iv. Οι πράκτορες παίζουν ένα επαναλαμβανόμενο στρατηγικό παίγνιο

Σχεδόν όλες αυτές οι προσεγγίσεις έχουν σχεδιαστεί για παίγνια με δύο μόνο πράκτορες.

Οι Littman και Stone (2001) επικεντρώθηκαν στα επαναλαμβανόμενα παίγνια δύο πρακτόρων και εισήγαγαν δύο επιθετικές στρατηγικές 'ηγέτη' οι οποίες προσπαθούν να επάγουν συμπεριφορά οπαδού στον αντίπαλο. Με αυτές τις στρατηγικές οι πράκτορες προσπαθούν να δημιουργήσουν μια ασύμμετρη κατάσταση Stackelberg σε ένα αυθεντικό συμμετρικό παίγνιο. Το κίνητρο των συγγραφέων ήταν ότι οι στρατηγικές καλύτερης απάντησης που χρησιμοποιούνται στις προσεγγίσεις της Θεωρίας Παιγνίων για την πολυπρακτορική μάθηση είναι βασικά στρατηγικές 'οπαδού' και ότι θα επιτυγχάνονταν καλύτερα αποτελέσματα με την εφαρμογή επιθετικών στρατηγικών 'ηγέτη', οι οποίες έμμεσα επάγουν τη συνεργασία στον αντίπαλο. Οι δύο στρατηγικές ονομάζονται 'Δυνάστης' (Bully) και 'Νονός' (Godfather) και βασίζονται αντίστοιχα στο πείσμα και στις απειλές. Θα μελετήσουμε το παίγνιο από την οπτική γωνία του πρώτου

πράκτορα. Ο Δυνάστης είναι μια ντετερμινιστική στρατηγική η οποία επιλέγει τις ενέργειές της με :

$$a_{1,k} = \arg\max_{a_1 \in A_1} \bar{p}_1(a_1, \arg\max_{a_2 \in A_2} \bar{p}_2(a_1, a_2)), \forall k \geq 0 \quad (25)$$

Ο Δυνάστης υποθέτει ότι ο αντίπαλος θα επιλέξει την καλύτερη απάντηση στη στρατηγική του και γι' αυτό το λόγω παίζει έτσι ώστε να μεγιστοποιήσει την επιβράβευσή του με αυτή την υπόθεση.

Η στρατηγική του Νονού επιλέγει μια συνδυαστική ενέργεια η οποία επιφέρει μεγαλύτερα οφέλη από αυτά της minimax (επίπεδο ασφαλείας). Στη συνέχεια εκτελεί το δικό του κομμάτι της συνδυαστικής ενέργειας. Εάν ο αντίπαλος δεν εκτελέσει το δικό του κομμάτι, ο Νονός εκτελεί την ενέργεια η οποία θα μειώσει το όφελος του αντιπάλου στο επίπεδο ασφαλείας. Με απλά λόγια ο Νονός προειδοποιεί τον αντίπαλό του 'παίξε το κομμάτι σου, ή ανεξάρτητα με το τι θα κάνεις δεν θα κερδίσεις περισσότερα από το επίπεδο ασφαλείας'. Εάν [a_1*, a_2*] είναι η συνδυαστική ενέργεια που επελέγη από τον Νονό:

$$\begin{cases} a_1 * & \text{εάν } k=0 \text{ ή } a_{2,k-1} = a_2 * \\ a_{1,k} = \arg\min_{a_1 \in A} \max_{a_2 \in A_2} \bar{p}_2(a_1, a_2) & \text{διαφορετικά} \end{cases} \quad (26)$$

Οι συγγραφείς δοκίμασαν τις στρατηγικές ηγέτη σε διάφορα επαναλαμβανόμενα παίγνια και έδειξαν ότι λειτουργούν καλύτερα από τις στρατηγικές καλύτερης απάντησης όταν 'συνεργάζονται' με τους αντιπάλους.

Οι Conitzer και Sandholm (2003) επικέντρωσαν το ενδιαφέρον τους στους στόχους μάθησης της λογικότητας και της σύγκλισης, όπως ορίστηκαν από τους Bowling και Veloso (2002). Παρουσίασαν έναν αλγόριθμο που ονομάζεται AWESOME από το "Adapt When Everybody is Stationary, Otherwise Move to Equilibrium". Ο πράκτορας εναλλάσσεται μεταξύ των στρατηγικών ισορροπίας και καλύτερης απάντησης βασιζόμενος στην εμπειρία του. Λειτουργεί σύμφωνα με δύο υποθέσεις: α) ότι όλοι οι άλλοι πράκτορες είναι σταθεροποιημένοι και β) ότι οι άλλοι πράκτορες εκτελούν το κομμάτι μιας προϋπολογισμένης ισορροπίας που τους αντιστοιχεί. Οι υποθέσεις τους μεταβάλλονται εξαρτώμενες από την ανάλυση των αλλαγών στις εμπειρικές κατανομές των ενεργειών των άλλων πρακτόρων μετά από διαδοχικές μαθησιακές επαναλήψεις. Η μέθοδος αποδείχτηκε ότι ασυμπτωτικά λειτουργεί σαν καλύτερη απάντηση απέναντι σε σταθεροποιημένους πράκτορες και συγκλίνει σε αυτό-παίξιμο για όλα τα επαναλαμβανόμενα παίγνια.

Ο Singh κ.ά. (2000b) εστίασαν την προσοχή τους στην βαθμιαία αυξητική ανανέωση της πολιτικής σε επαναλαμβανόμενα στρατηγικά παίγνια με δύο πράκτορες οι οποίοι εκτελούν δύο ενέργειες. Σε αυτά τα παίγνια η στοχαστική στρατηγική του ενός πράκτορα μπορεί να μοντελοποιηθεί με ένα πραγματικό αριθμό από το διάστημα [0,1]. Αν υποθέσουμε ότι η στρατηγική του πράκτορα 1 αναπαρίσταται με το α, με τιμές από 0 έως και 1, τότε ο πράκτορας επιλέγει την πρώτη ενέργεια με πιθανότητα α και τη δεύτερη ενέργεια με πιθανότητα 1-α. Παρόμοια η στοχαστική στρατηγική του δεύτερου πράκτορα περιγράφεται με μια παράμετρο β. Ο κανόνας ανανέωσης είναι τότε:

$$\begin{cases} a_{k+1} = a_k + \delta \dfrac{\partial V_1(a_k, \beta_k)}{\partial a} \\ \beta_{k+1} = \beta_k + \delta \dfrac{\partial V_2(a_k, \beta_k)}{\partial \beta} \end{cases} \quad (27)$$

όπου V_i (α_k,β_k) είναι το προσδοκώμενο όφελος του πράκτορα i δεδομένου ότι οι πράκτορες ακολουθούν στρατηγικές α_k, β_k και δ είναι το μέγεθος του βήματος της μεθόδου κλίσης. Το κύριο αποτέλεσμα που αποδείχθηκε από τους συγγραφείς ήταν ότι θεωρώντας ένα απειροστικό μέγεθος βήματος ($\lim_{\delta \to 0}$) το όφελος των πρακτόρων συγκλίνει στο όριο στο όφελος Nash.

Προβλήματα Πολλαπλών Καταστάσεων (Multiple-state Problems)

Προκειμένου να τυγχάνει πραγματικού ενδιαφέροντος, ένας αλγόριθμος Πολυπρακτορικής Ενισχυτικής Μάθησης είναι απαραίτητο να αντιμετωπίζει προβλήματα με μη-κενό διάστημα καταστάσεων και καθυστερημένη επιβράβευση (παίγνιο Markov). Ένας τέτοιος αλγόριθμος τυπικά θέτει άλλη μια απαίτηση εκτός από τις δυο υποθέσεις που είδαμε προηγουμένως: Όλοι οι πράκτορες χρησιμοποιούν τον ίδιο αλγόριθμο μάθησης. Πολλοί αλγόριθμοι Πολυπρακτρικής Ενισχυτικής Μάθησης για τα Παίγνια Markov προέρχονται από την Q-μάθηση και έχουν ένα κοινό σκελετό. Αυτός ο σκελετός δίδεται στον επόμενο αλγόριθμο. Αναφέρεται σε έναν πράκτορα, που προσδιορίζεται ως πράκτορας i. Οι τελείες '.' αποδίδουν το 'όλες οι τιμές του αντίστοιχου παράγοντα'. Ο αλγόριθμος είναι σχεδόν όμοιος με τον βασικό αλγόριθμο Q-μάθησης. Υπάρχουν όμως κάποιες σημαντικές διαφορές που θα αναλύσουμε.

Αλγόριθμος *Γενικός Πολυπρακτορικής Q-μάθησης για πράκτορα i*
Είσοδος: ρυθμός μάθησης α, παράγοντας μείωσης γ

1. $Q_i(s,a) \leftarrow 0, \forall s \in S, a \in A, where A = \times_{j \in N} A_j$
2. $h_i(s,a_i) = 1/|A_i|, \forall a_i \in A_i$
3. παρατήρησε αρχική κατάσταση s

4. **βρόχος**
5. $h_i(s,\cdot) \leftarrow solve_i(Q_1(s,\cdot),...,Q_n(s,\cdot))$
6. σχεδίασε a_i, σύμφωνα με $h_i(s,a_i)$
7. εφάρμοσε a_i, με κατάλληλη εξερεύνηση
8. παρατήρησε άλλες ενέργειες $a_j, j \in N, j \neq i$, επιβραβεύσεις $r_j, j \in N$ και επόμενη κατάσταση s'
9. $Q_i(s,a) \leftarrow Q_i(s,a) + a[r_i + \gamma \cdot eval_i(Q_1(s',\cdot),...,Q_n(s',\cdot)) - Q_i(s,a)]$
10. **για** $j \in A, j \neq i$
11. $Q_j(s,a) \leftarrow Q_j(s,a) + a[r_j + \gamma \cdot eval_j(Q_1(s',\cdot),...,Q_n(s',\cdot)) - Q_j(s,a)]$
12. **τέλος_για**
13. $s \leftarrow s'$
14. **τέλος_βρόχου**

Στη γραμμή 5 ανανεώνεται η πολιτική του πράκτορα στην κατάσταση s, η h_i (s,.) με τη στρατηγική που παράγεται από τη $solve_i$, από τους Q-πίνακες όλων των πρακτόρων. Έτσι, αντί της συνηθισμένης άπληστης πολιτικής, ένας πράκτορας ακολουθεί μια στοχαστική πολιτική που υπολογίζεται με βάση τους Q-πίνακες όλων των πρακτόρων. Η γραμμή 9 είναι μια ανανέωση χρονικής διαφοράς, αλλά η τιμή της επόμενης κατάστασης που χρησιμοποιείται δεν είναι η τιμή της άπληστης σε αυτή την κατάσταση, όπως στη βασική Q-μάθηση. Αντιθέτως, αυτή η τιμή δίδεται από τη συνάρτηση $eval_i$ η οποία λαμβάνει υπόψη τους Q-πίνακες των άλλων πρακτόρων.

Όπως γίνεται αντιληπτό από τα δύο αυτά βήματα, είναι απαραίτητοι οι Q-πίνακες όλων των πρακτόρων. Έτσι ένας πράκτορας απαιτείται να διατηρεί τους Q-πίνακες και των άλλων πρακτόρων εκτός από τον δικό του. Αυτό γίνεται με το βήμα μοντελοποίησης στις γραμμές 10-12. Η αναγκαιότητα μετρήσιμων επιβραβεύσεων (ii) και η υπόθεση του αυτό-παιξίματος (v) είναι προφανή στη γραμμή 11. Οι επιβραβεύσεις των άλλων πρακτόρων απαιτούνται για να γίνει η ανανέωση, και οι άλλοι πράκτορες επιβάλλεται να χρησιμοποιούν τον ίδιο αλγόριθμο μάθησης προκειμένου η αναβάθμιση να ανταποκρίνεται στην πραγματικότητα. Είναι επίσης απαραίτητο να σχολιάσουμε λίγο και τη γραμμή 7. Στο σημείο αυτό ο πράκτορας ενεργεί βάσει της στοχαστικής του πολιτικής. Η ενέργεια του πράκτορα εκτελείται στο περιβάλλον ταυτόχρονα με τις ενέργειες των άλλων πρακτόρων. Ο πράκτορας εφαρμόζει στην κατάσταση x τη στρατηγική που παράγεται από τη solve, αλλά μερικές φορές εκτελεί ενέργειες με εξερεύνηση. Και είναι αναγκαία αυτή η εξερεύνηση (αν και οι πολιτικές των πρακτόρων είναι στοχαστικές) διότι τίποτα δεν εμποδίζει τις στρατηγικές που παράγονται από τη solve να δίδουν βάρος μηδενικής πιθανότητας σε κάποιες εκ των ενεργειών. Μια

κατάλληλη πολιτική εξερεύνησης, αντιθέτως δεν θα ανέθετε βάρος μηδενικής πιθανότητας σε καμιά ενέργεια, εκτός ίσως από το όριο.

Πολλοί αλγόριθμοι πολυπρακτορικής μάθησης αποτελούν εφαρμογές του σκελετού του Γενικού Αλγορίθμου Πολυπρακτορικής Q-Μάθησης για τον πράκτορα i.

Εάν το παίγνιο είναι πλήρως ανταγωνιστικό μπορούμε να εφαρμόσουμε την ιδέα του minimax (Littman, 2001b). Σε αυτή την περίπτωση, ο πρώτος πράκτορας (ο παίκτης) επιλέγει τη στρατηγική η οποία μεγιστοποιεί το όφελός του με την υπόθεση ότι ο άλλος πράκτορας (ο αντίπαλος) θα δράσει έτσι ώστε να ελαχιστοποιήσει αυτό το όφελος. Η τιμή της επόμενης κατάστασης που χρησιμοποιείται για την ανανέωση θεωρείται επίσης ότι λαμβάνει την τιμή της χειρότερης περίπτωσης. Από την οπτική γωνία του παίκτη και θεωρώντας ντετερμινιστική επιλογή ενέργειας για τον αντίπαλο η τιμή αυτή είναι:

$$\begin{cases} eval_1(Q(s,\cdot)) = \max\limits_{h_1(s,\cdot)\in\Pi(A_1)} \min\limits_{a_2\in A_2} \sum\limits_{a_1\in A_1} h_1(s,a_1)Q(s,a_1,a_2) \\ solve_1(Q(s,\cdot)) = \arg\max\limits_{h_1(s,\cdot)\in\Pi(A_1)} \min\limits_{a_2\in A_2} \sum\limits_{a_1\in A_1} h_1(s,a_1)Q(s,a_1,a_2) \end{cases} \quad (28)$$

Αξίζει να σημειωθεί ότι εξ' αιτίας του γεγονότος ότι $p_1 = -p_2$, ισχύει $Q_1 = -Q_2$ και επομένως ο πράκτορας έχει τη δυνατότητα να αποθηκεύει μόνο ένα Q-πίνακα, ο οποίος δηλώνεται ως Q στην (28). Η μοντελοποίηση του Q-πίνακα του αντιπάλου δεν είναι απαραίτητη. Ο αλγόριθμος minimax μπορεί να επεκταθεί ώστε να συμπεριλάβουμε περισσότερους από δύο πράκτορες υποθέτοντας ότι όλοι οι άλλοι πράκτορες θα δράσουν εναντίον του μαθητευόμενου. Ο Littman το 2001 παρουσίασε αυτή την επέκταση. Συνδύασε τον minimax-Q και τον team-Q σε έναν απλό αλγόριθμο που τον ονόμασε "φίλος ή εχθρός Q-μάθηση" (friend of foe Q-learning) και παρουσίασε τις συνθήκες σύγκλισης αυτού του αλγορίθμου για τις συγκεκριμένες κλάσεις των παιγνίων Markov όπου αυτός συγκλίνει. Το όνομα του αλγορίθμου βασίζεται στο γεγονός ότι κάποιοι πράκτορες σχεδιάστηκαν ως φίλοι και κάποιοι σαν εχθροί.

Ο αλγόριθμος Nash Q- μάθησης (Nash Q-learning) λειτουργεί σε μια πιο γενική κλάση Παιγνίων Markov (Hu & Wellman,2003). Όπως προδίδει και το όνομά του, ο αλγόριθμος υπολογίζει μια Nash-ισορροπία σε κάθε κατάσταση του Παιγνίου Markov και την χρησιμοποιεί στην ανανέωση της χρονικής διαφοράς και της πολιτικής.

$$\begin{cases} eval_i(Q_1(s,\cdot),...,Q_n(s,\cdot)) = V_i^{NE[Q_1(s,\cdot),...,Q_n(s,\cdot)]}(s) \\ solve_i(Q_1(s,\cdot),...,Q_n(s,\cdot)) = NE_i[Q_1(s,\cdot),...,Q_n(s,\cdot)] \end{cases} \quad (29)$$

Η έκφραση NE_i αναφέρεται στο μερίδιο της στρατηγικής της Nash-ισορροπίας που αντιστοιχεί στον πράκτορα i.

Η τιμή V_i μιας κατάστασης για έναν πράκτορα i ακολουθώντας τη συνδυαστική πολιτική, είναι το άθροισμα των Q-πινάκων των πρακτόρων σταθμισμένων με την πιθανότητα των αντίστοιχων συνδυαστικών ενεργειών που υπαγορεύονται από τη συνδυαστική πολιτική σε αυτή την κατάσταση.

$$V_i^{(h_1,...,h_n)}(s) = \sum_{a_1 \in A_1,...,a_n \in A_n} h_1(s,a_1)...h_n(s,a_n)Q_i(s,a_1,...,a_n) \quad (30)$$

Στη Nash-μάθηση ο κάθε πράκτορας είναι απαραίτητο να μοντελοποιήσει τους Q-πίνακες όλων των άλλων. Ο αλγόριθμος αποδείχθηκε ότι συγκλίνει, με δεδομένο ότι οι πράκτορες χρησιμοποιούν την ίδια Nash-ισορροπία για να υπολογίσουν την eval σε όλες τις καταστάσεις του παιγνίου. Είναι φανερό ότι ο αλγόριθμος Nash-Q-μάθηση καλείται να αντιμετωπίσει το πρόβλημα της επιλογής ισορροπίας.

Αντικαθιστώντας τη Nash-ισορροπία με μια συσχετιζόμενη ισορροπία καταλήγουμε στην Q-μάθηση με συσχετιζόμενη ισορροπία (Greenwald & Hall, 2003). Το πλεονέκτημα είναι ότι η μια συσχετιζόμενη ισορροπία υπολογίζεται μέσω του γραμμικού προγραμματισμού. Οι συγγραφείς πειραματίστηκαν με 4 είδη ισορροπιών, με τις συσχετίσεις μεταξύ των πρακτόρων εκφρασμένες στην αντικειμενική συνάρτηση γραμμικού προγραμματισμού: την *λειτουργική (utilitarian)* όπου μεγιστοποιείται το άθροισμα των επιβραβεύσεων των πρακτόρων, την *ισόνομη (egalitarian)* όπου μεγιστοποιείται η ελάχιστη τιμή των επιβραβεύσεων των πρακτόρων, την *ρεπουμπλικανή (republican)* όπου μεγιστοποιείται η μέγιστη τιμή των επιβραβεύσεων των πρακτόρων και την *φιλελεύθερη (libertarian)* όπου μεγιστοποιείται ανεξάρτητα η μέγιστη τιμή της επιβράβευσης του κάθε πράκτορα.

Παρόμοια, χρησιμοποιώντας την ισορροπία Stackelberg στις ανανεώσεις αποκτούμε ασύμμετρη πολυπρακτορική Ενισχυτική Μάθηση (Kononen, 2003). Αυτός ο αλγόριθμος μεταχειρίζεται καταστάσεις όπου ο πράκτορας-ηγέτης μπορεί να επιβάλλει επιλογές ενεργειών στους πράκτορες-οπαδούς. Αν και οι πράκτορες δεν χρησιμοποιούν ίδιους αλγορίθμους, είναι απαραίτητο να ακολουθούν τα συμπληρωματικά τους μέλη στην

ασύμμετρη διαδικασία μάθησης, γεγονός που καθιστά την υπόθεση του αυτό-παιξίματος απαραίτητη.

Μια παραλλαγή της πολυπρακτορικής Q-μάθησης η οποία αφαιρεί την υπόθεση του αυτό-παιξίματος είναι ο αλγόριθμος που είναι γνωστός ως αλγόριθμος της 'εκτεταμένης βέλτιστης απάντησης' (extended optimal response) που παρουσιάστηκε από τους Suematsu και Hayashi το 2002. Ο αλγόριθμος σχεδιάστηκε για παίγνια με δυο πράκτορες. Θα τον περιγράψουμε από την οπτική του πρώτου πράκτορα. Ο στόχος του αλγορίθμου είναι να φτάσουμε σε μια Nash ισορροπία όταν ο δεύτερος πράκτορας εκπαιδεύεται, αλλά να την αξιοποιούμε χρησιμοποιώντας την καλύτερη απάντηση όταν επιδεικνύει σταθεροποιημένη συμπεριφορά: αυτή είναι η εκτεταμένη βέλτιστη απάντηση. Ο πράκτορας χρησιμοποιεί την επόμενη ευριστική συνάρτηση solve:

$$solve_1(Q_1(s,\cdot),Q_2(s,\cdot) = \arg\max_{h_1(s,\cdot)\in\Pi(A_1)} [V_1^{(h_1,\tilde{h_2})}(s) - \varsigma\rho(s,Q_2(s,\cdot),h_1)] \quad (31)$$

όπου ς είναι μια συντονιστική παράμετρος, $\tilde{h_2}$ είναι μια εκτίμηση της πολιτικής του δεύτερου πράκτορα, και ρ είναι μια συνάρτηση απόστασης η οποία προσεγγίζει τη βελτίωση του οφέλους το οποίο επιτυγχάνει ο πράκτορας τροποποιώντας την πολιτική του δεδομένης της πολιτικής h_1

$$\rho(s,Q_2(s,\cdot),h_1) = \max_{h_2(s,\cdot)\in\Pi(A_2)} V_2^{(h_1,h_2)}(s) - V_2^{(h_1,\tilde{h_2})}(s) \quad (32)$$

Οι τιμές V των συνδυαστικών πολιτικών υπολογίζονται από την (30). Η πραγματική ανανέωση χρονικής διαφοράς που χρησιμοποιείται από τον πράκτορα είναι η ανανέωση SARSA (6), επομένως η eval είναι απλά η Q-τιμή της επόμενης συνδυαστικής ενέργειας που λαμβάνεται. Το πρώτο μέρος του παράγοντα βελτιστοποίησης στην (31) αναφέρεται στο κομμάτι της συμπεριφοράς του πράκτορα που σχετίζεται με την καλύτερη απάντηση, ενώ το δεύτερο μέρος κατευθύνει το σύστημα προς μια Nash ισορροπία μειώνοντας έμμεσα την επιθυμία του αντιπάλου να παρεκκλίνει από την πολιτική του.

3.4 Θέματα στην πραγματική Ενισχυτική Μάθηση

Τα θεωρητικά αποτελέσματα δείχνουν ενθαρρυντικά. Όταν, όμως η Πολυπρακτορική Ενισχυτική Μάθηση (και η ενισχυτική μάθηση γενικότερα) εφαρμόζεται σε εργασίες της πραγματικής ζωής ή σε ρεαλιστικές προσομοιώσεις, προκύπτουν κάποια προβλήματα τα οποία δεν τα έχουμε περιγράψει παραπάνω. Θα αναφερθούμε με συντομία σε αυτά στη συνέχεια:

Μεγάλα και/ή συνεχή διαστήματα καταστάσεων και ενεργειών

Όταν τα διαστήματα καταστάσεων και ενεργειών είναι μεγάλα και/ή έχουν συνεχείς διαστάσεις, η αναπαράσταση των συναρτήσεων αξίας με χρήση πινάκων δεν είναι εφικτή. Σε αυτές τις περιπτώσεις συνήθως καταφεύγουμε στη χρήση τεχνικών από την μάθηση με επίβλεψη, όπως π.χ. τα νευρωνικά δίκτυα (Sutton & Barto, 1998).

Σε αυτή την περίπτωση όμως προκύπτει ένα καινούριο πρόβλημα. Το πρόβλημα είναι ότι πλέον δεν ισχύουν πολλές από τις θεωρητικές εγγυήσεις σύγκλισης. Μια κλάση μεθόδων οι οποίες διατηρούν τις ιδιότητες σύγκλισης είναι η αναζήτηση άμεσης πολιτικής με gradient descent. Όμως και αυτές οι μέθοδοι παρουσιάζουν κάποιες δυσκολίες. Είναι αργές και παγιδεύονται σε τοπικά βέλτιστα (Baird 1995, Baird & Moore 1998).

Μερική μετρησιμότητα

Σε πολλές περιπτώσεις οι πράκτορες δεν έχουν τη δυνατότητα να μετρήσουν πλήρως την κατάσταση του περιβάλλοντος και σαν συνέπεια αυτού, η υπόθεση Markov παραβιάζεται. Εξαρτάται από τη σοβαρότητα της παραβίασης για το εάν ή όχι οι εκπαιδευόμενοι πράκτορες θα την αγνοήσουν. Εάν δεν μπορούμε να την αγνοήσουμε διαλέγουμε μεταξύ δύο εναλλακτικών τύπων λύσεων που συναντάμε στη βιβλιογραφία. Στην πρώτη, διατηρούμε κατανομές πιθανότητας των πεποιθήσεων στην κατάσταση της “υποβόσκουσας” διεργασίας Markov τις οποίες ανανεώνουμε με τις διάφορες παρατηρήσεις χρησιμοποιώντας ένα πλαίσιο Bayes. Αυτό συνεπάγεται υψηλά υπολογιστικά κόστη. Στη δεύτερη, παρέχεται στον πράκτορα η ικανότητα να θυμάται αλληλουχίες καταστάσεων με την ελπίδα ότι αυτές οι αλληλουχίες θα εμπεριέχουν αρκετή πληροφορία, προκειμένου να επιτευχθεί η ιδιότητα Markov (Whitehead & Lin, 1995).

Προγενέστερη γνώση

Στα πραγματικά προβλήματα, η ενισχυτική μάθηση χωρίς μοντέλο δεν είναι εφικτή. Η εξερεύνηση ενός μεγάλου διαστήματος ενεργειών-καταστάσεων, η καθυστερημένη ενίσχυση, και η αβεβαιότητα, επιβραδύνουν τη μάθηση ή την καθιστούν αδύνατη. Επειδή δεν είναι δυνατή η κατασκευή ακριβών μοντέλων της διεργασίας, καταφεύγουμε στη χρήση προγενέστερης γνώσης ώστε να διευκολυνθεί η διαδικασία μάθησης. Αυτό μπορούμε να το επιτύχουμε με διάφορους τρόπους:

- ✓ Αρχικοποίηση: Η συνάρτηση αξίας μπορεί να αρχικοποιηθεί έτσι ώστε να εμπεριέχει προηγούμενη γνώση.

- ✓ Τοπικά σήματα ενίσχυσης: Μπορούν να βοηθήσουν απαλείφοντας το πρόβλημα της δομικής και χρονικής εκχώρησης πίστωσης.
- ✓ Μάθηση: Ένα άτομο μπορεί να λάβει τον έλεγχο κάποιου πράκτορα κάποια χρονική στιγμή και να τον κατευθύνει προς το στόχο.
- ✓ Διαμόρφωση: Αρχικά ο πράκτορας εκπαιδεύεται σε πολύ απλές εργασίες και στη συνέχεια σταδιακά αυξάνεται η δυσκολία.
- ✓ Αποσύνθεση προβλήματος: Ακολουθείται στη μάθηση κατά επίπεδα (Stone & Veloso, 1998).
- ✓ Αντιδράσεις: Προσδίδονται στους πράκτορες με προσρόφηση. Θα αποτελέσουν θεμέλιοι λίθοι για μια πιο πολύπλοκη συμπεριφορά. (Mataric, 1996)

3.5 Επίλογος

Ένα χαρακτηριστικό στοιχείο αυτής της εξεταζόμενης επιστημονικής περιοχής είναι οι αντιγνωμίες που παρατηρούνται όσον αφορά τη διατύπωση του κατάλληλου στόχου της Πολυπρακτορικής Ενισχυτικής Μάθησης. Ένα επιπλέον στοιχείο αποτελεί το χάσμα μεταξύ των θεωρητικών αποτελεσμάτων και των πρακτικών εφαρμογών. Μέθοδοι οι οποίες αποδεικνύονται θεωρητικά, έντονα επηρεασμένες από τη θεωρία παιγνίων, επιβάλλουν πολύ περιοριστικές υποθέσεις προκειμένου να αποδείξουν τα αποτελέσματά τους. Υποθέσεις, όμως, που δεν ισχύουν στις περισσότερες των πρακτικών καταστάσεων.

Πρόβλημα δημιουργείται και από την απαίτηση για μια ρητή αναπαράσταση με χρήση πινάκων των διαστημάτων καταστάσεων-ενεργειών. Το πρόβλημα αυτό διογκώνεται στα πλαίσια της πολυπρακτορικής μάθησης καθώς εκτινάσσονται εκθετικά τα διαστήματα με τον αριθμό των πρακτόρων. Επιπλέον, ένα αξιοσημείωτο στοιχείο είναι ότι οι αξιολογήσεις των διαφόρων αλγορίθμων πραγματοποιούνται σε πολύ απλά προβλήματα, κάποια από αυτά χωρίς καταστάσεις, και κάποια με πολύ μικρά διαστήματα καταστάσεων, χωρίς να δίνονται οδηγίες για το πώς οι αλγόριθμοι θα επεκταθούν σε προβλήματα πραγματικών μεγεθών.

Τέλος ένα θέμα που προκύπτει, είναι ότι ο στόχος της επικεντρώνεται στη βελτιστότητα, αγνοώντας άλλες επιθυμητές ιδιότητες της συμπεριφοράς των πρακτόρων, όπως η θετική προσαρμογή στις αλλαγές του περιβάλλοντος, ή μια σχεδόν μονότονη βελτίωση της απόδοσης κατά τη διαδικασία μάθησης. Πολλοί αλγόριθμοι παρέχουν εγγυήσεις ασυμπτωτικής σύγκλισης αμελώντας την βραχυπρόθεσμη απόδοση του πράκτορα.

ΚΕΦΑΛΑΙΟ 4
ΜΟΝΤΕΛΟ ΜΑΘΗΣΗΣ ΣΕ ΕΙΚΟΝΙΚΕΣ ΑΓΟΡΕΣ

4.1 Σενάριο

Παρακινούμενοι από τα ενδιαφέροντα χαρακτηριστικά της ενισχυτικής μάθησης, προσπαθήσαμε να υλοποιήσουμε μια εικονική αγορά στην οποία θα μελετούσαμε τα οφέλη της χρήσης της από κάποιον υποψήφιο αγοραστή και πως αυτή θα επηρέαζε τη συμπεριφορά. Πρόκληση πολύ ενδιαφέρουσα και λίγο ριψοκίνδυνη, αν αναλογιστεί κανείς τον πολύ δυναμικό χαρακτήρα των εικονικών αγορών. Ας τα πάρουμε όμως τα πράγματα από την αρχή…

Στο σενάριο μας θα διαμορφώσουμε μια εικονική αγορά πληροφορίας στην οποία οι υποψήφιοι αγοραστές συνεργάζονται με τους ενδιάμεσους προκειμένου να αποκτήσουν τα προϊόντα που ανταποκρίνονται καλύτερα στις απαιτήσεις τους. Η επιλογή του ενδιαμέσου προσδιορίζεται από την υπόδειξη του αλγορίθμου της Q-μάθησης. Τα χαρακτηριστικά των επιθυμητών επιλογών αποτυπώνονται στις επιβραβεύσεις που αποδίδονται στον αγοραστή κατά την εκπαίδευση. Οι ενδιάμεσοι στο σενάριό μας λειτουργούν ως brokers. Διατηρούν την ανωνυμία των πωλητών με τους οποίους πραγματοποιείται η συναλλαγή.

Οντότητες του προγράμματος

Στο πρόγραμμά μας μελετούμε τις αλληλεπιδράσεις μεταξύ υποψηφίων αγοραστών(buyer) και ενδιαμέσων οντοτήτων(mediators) οι οποίοι ουσιαστικά αποτελούν αντιπροσώπους των πωλητών που συμμετέχουν στην εικονική αγορά. Πιο αναλυτικά στο πρόγραμμά μας συναντούμε:

Marketplace

Αποτελεί την κύρια κλάση της εικονικής μας αγοράς. Εδώ κατασκευάζονται οι οντότητες μας και διαμορφώνεται το σκηνικό στο οποίο αυτές θα αλληλεπιδράσουν.

Mediators

Η οντότητα (πράκτορας) που τοποθετείται μεταξύ αγοραστών και πωλητών. Στο πρόγραμμα αυτό λειτουργεί ως broker. Έχει διαθέσιμες προσφορές προϊόντων τα οποία ανταποκρίνονται στις ανάγκες των αγοραστών. Ο αριθμός των mediators που κατασκευάζονται κατά την εκκίνηση του προγράμματος της εικονικής αγοράς, το πλήθος και το είδος των προϊόντων εξαρτώνται από ένα αρχείο. Στο αρχείο αυτό αναγράφονται με λεπτομέρεια τα χαρακτηριστικά (ο αριθμός, οι κωδικοί, οι τιμές, οι

σχετικότητες, οι ποσότητες και οι χρόνοι απόκτησης των διαθέσιμων προϊόντων) κάθε ενός mediator. Επιλέξαμε αυτό τον τρόπο για να διευκολυνθεί η διαδικασία αξιολόγησης του προγράμματος. Είναι όμως πολύ εύκολη η τροποποίησή του ώστε όλες οι παραπάνω παράμετροι να πάρουν τυχαίες τιμές κατά την εκκίνηση.Κατά την εκτέλεση του προγράμματος παρέχεται η δυνατότητα εισαγωγής και απομάκρυνσης προϊόντων καθώς και μεταβολής των διαθέσιμων προσφορών από τους πωλητές σε τυχαίες χρονικές στιγμές.

Buyer

Η οντότητα (πράκτορας) η οποία προβαίνει στην εκτέλεση αγορών εκ μέρους μας σύμφωνα με κάποιες προτιμήσεις που έχουμε περιγράψει. Διαχωρίζεται σε δύο τμήματα. Το πρώτο τμήμα είναι υπεύθυνο για την κατασκευή των Q-πινάκων κάνοντας χρήση των πληροφοριών που αποκτά από την επικοινωνία με τους ενδιάμεσους. Ο τρόπος με τον οποίο επιτυγχάνεται το χτίσιμο των πινάκων εκμάθησης αναλύεται στη συνέχεια. Σημαντικό στοιχείο αποτελεί ότι οι τιμές των στοιχείων των Q-πινάκων πρέπει να αντικατοπτρίζουν τις ‘τρέχουσες’ συνθήκες στην εικονική αγορά.

Το δεύτερο τμήμα όταν καταστεί ανάγκη αγοράς ενός προϊόντος συμβουλεύεται τις εγγραφές των Q-πινάκων για να επιλέξει τον πιο συμφέροντα ενδιάμεσο και να συνδιαλλαγεί μαζί του. Η επιτυχία ή αποτυχία της αποστολής του εξαρτάται από τη δυνατότητα προσαρμογής στις έντονες και διαρκείς μεταβολές που παρατηρούνται σε ένα δυναμικό περιβάλλον όπως αυτό των εικονικών αγορών η οποία αποτυπώνεται στις τιμές των Q-πινάκων που του προσφέρει το τμήμα κατασκευής.

Βοηθητικές κλάσεις

Q_Message

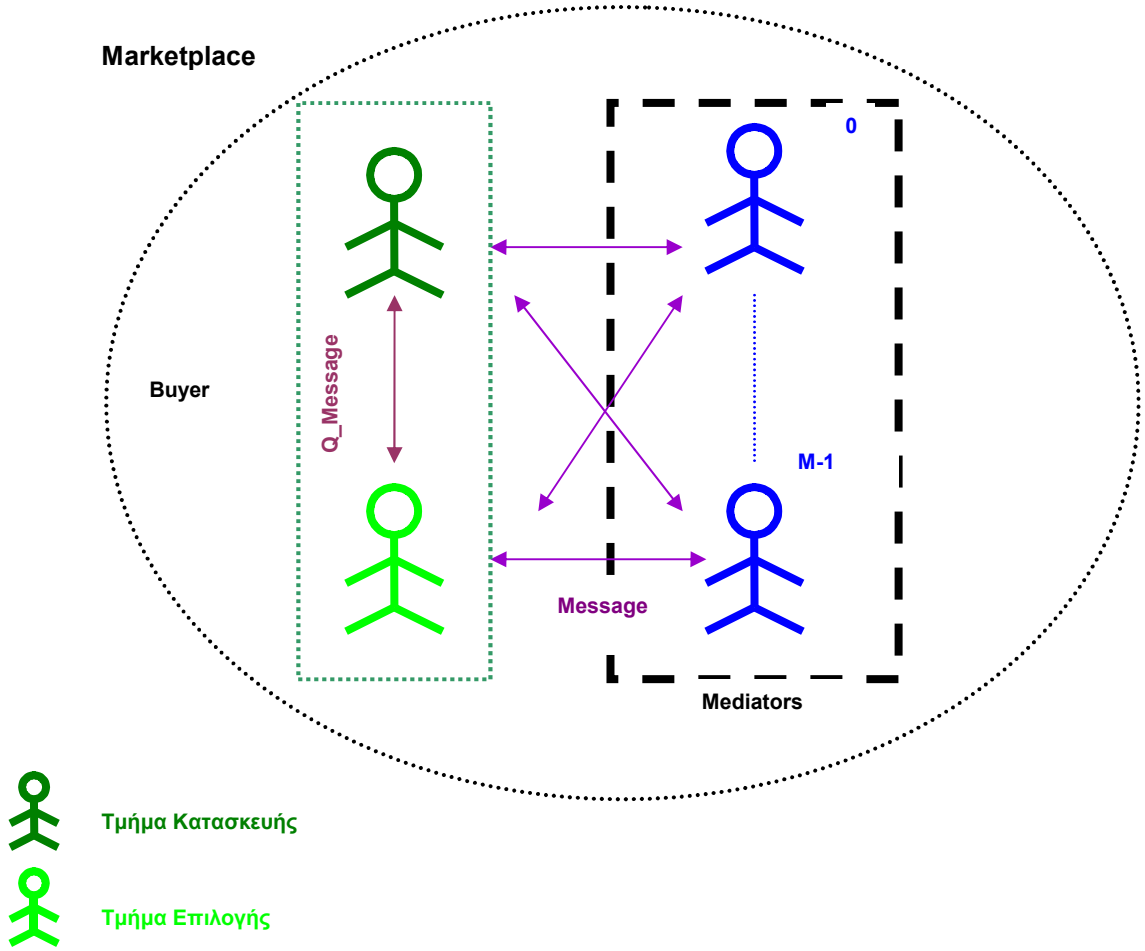
Αντικείμενα αυτής της κλάσης αποτελούν τα μηνύματα που ανταλλάσουν το τμήμα κατασκευής των πινάκων του Buyer και το τμήμα επιλογής του κατάλληλου ενδιαμέσου. Ουσιαστικά το τμήμα επιλογής ζητά από το τμήμα κατασκευής την καινούρια έκδοση των Q-πινάκων καθώς έχει επέλθει μια σημαντική αλλαγή στην εικονική μας αγορά. Ζητά τους καινούριους πίνακες για να επιλέξει την πιο κατάλληλη ενέργεια την τρέχουσα χρονική στιγμή.

Message

Αντικείμενα αυτής της κλάσης αποτελούν τα μηνύματα που ανταλλάσουν ο Buyer με τους Mediators. Κατά την εκπαίδευση – χτίσιμο των Q-πινάκων το τμήμα κατασκευής του Buyer αποστέλλει τέτοια μηνύματα στους Mediators με την ελπίδα να αποδώσουν

τιμές στα χαρακτηριστικά με τα οποία αυτός υπολογίζει τις Q-αξίες, εφόσον αυτοί έχουν το προϊόν. Πανομοιότυπα, κατά την αγορά των προϊόντων το τμήμα επιλογής ενδιαμέσου του Buyer με ένα τέτοιο μήνυμα δηλώνει την πρόθεση αγοράς ενός προϊόντος από τον Mediator στον οποίο το αποστέλλει και ο Mediator στη συνέχεια του απαντά με τα επιπλέον χαρακτηριστικά του προϊόντος, με την προϋπόθεση αυτό να είναι εκείνη τη χρονική στιγμή διαθέσιμο.

Στο επόμενο σχήμα φαίνεται πως αλληλεπιδρούν οι διάφορες οντότητες του προγράμματός μας.



Εικόνα 6. Κλάσεις Μοντέλου

Ας ξετυλίξουμε, λοιπόν σιγά – σιγά το σενάριό μας…

Κατά την έναρξη της εκτέλεσης της κύριας κλάσης (Marketplace) δημιουργείται το σκηνικό της αγοράς. Δημιουργούνται τόσοι Mediators όσοι αναγράφονται στο αρχείο των ενδιαμέσων, το τμήμα κατασκευής και το τμήμα επιλογής του πράκτορα Buyer Κάθε Mediator έχει συγκεκριμένο αριθμό προϊόντων με συγκεκριμένα χαρακτηριστικά οι τιμές των οποίων καθορίζονται από το παραπάνω αρχείο. Το τμήμα επιλογής ενδιαμέσου ξεκινά αμέσως ‘δουλειά’ και προβαίνει στην αγορά ενός συγκεκριμένου

αριθμού προϊόντων (π.χ. 400) τον οποίο εμείς προσδιορίζουμε. Ο αριθμός αυτός είναι πολύ σημαντικός για το πρόγραμμά μας καθώς αποτελεί μια σημαντική παράμετρος αξιολόγησής του. Απαραίτητη προϋπόθεση όμως για την πραγματοποίηση μιας αγοράς είναι η κατοχή μιας έγκυρης έκδοσης των Q-πινάκων οι οποίοι θα του υποδείξουν την καλύτερη κάθε φορά επιλογή. Αν έχει παρατηρηθεί μια σημαντική αλλαγή (π.χ. προσθήκη – αφαίρεση ενός προϊόντος) στο περιβάλλον της εικονικής μας αγοράς το τμήμα επιλογής ζητά από το τμήμα κατασκευής να του διαθέσει τους Q-πίνακες οι οποίοι ανταποκρίνονται σε αυτές τις αλλαγές.

Q-πίνακας

Κάθε στοιχείο του Q-πίνακα αποτελεί την αξία που μπορεί να αποκτηθεί από την πραγματοποίηση μιας ενέργειας στην τρέχουσα κατάσταση. Ο πίνακας Q που χρησιμοποιούμε στην εφαρμογή έχει τρεις διαστάσεις. Η πρώτη διάσταση αναφέρεται στον κωδικό του προϊόντος, η δεύτερη στην τρέχουσα κατάσταση και η τρίτη στην ενέργεια που εκτελεί ο πράκτορας σε αυτήν την κατάσταση. Επομένως στην πρώτη διάσταση μπορούμε να έχουμε έναν από τους N διακριτούς κωδικούς προϊόντων που μπορεί να είναι διαθέσιμοι στην αγορά, ενώ στην δεύτερη και στην τρίτη μια τιμή μεταξύ των M+1 διακριτών καταστάσεων, όπου M ο αριθμός των mediators και η μια παραπάνω αναφέρεται στην κατάσταση αγοράς του προϊόντος.

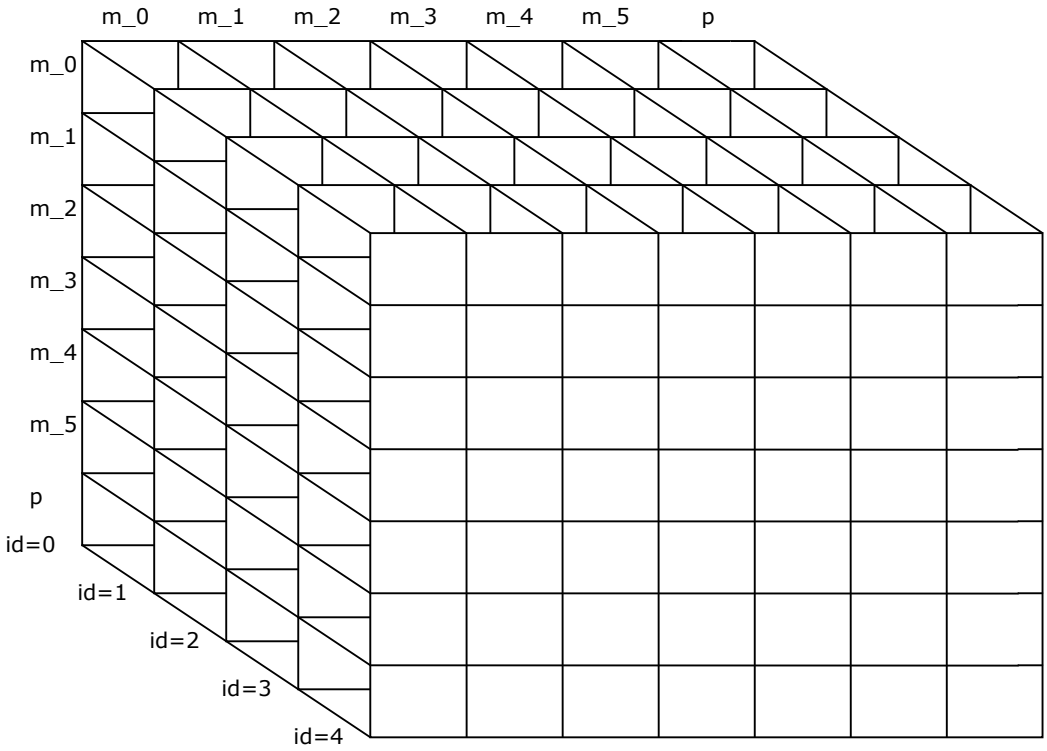
Συμπεραίνουμε λοιπόν ότι ο αριθμός των στοιχείων του πίνακα Q δίνεται από τον τύπο

$$NEl = N * (M+1)^2 \quad (33)$$

Στο σχήμα που ακολουθεί, παρουσιάζεται η γραφική αναπαράσταση του Q-πίνακα στην περίπτωση που στην εικονική αγορά δραστηριοποιούνται 6 mediators οι οποίοι μπορούν να παραθέσουν προσφορές για 5 προϊόντα που υπάρχουν στην αγορά. Η κατάσταση – ενέργεια p αναφέρεται στην περίπτωση αγοράς ενός προϊόντος.

Ενημέρωση του πίνακα εκμάθησης (Q-πίνακας)

Πριν την έναρξη των αγορών και κάθε φορά που εισάγεται ή εξάγεται ένα προϊόν στην αγορά πραγματοποιείται σχεδόν άμεση ενημέρωση των τιμών του Q-πίνακα. Το τμήμα κατασκευής του Buyer ενεργοποιείται κάθε φορά που παρατηρείται μια τέτοια σημαντική αλλαγή στην αγορά και ακολουθεί την κατάλληλη διαδικασία προκειμένου να ενημερώσει τις Q-αξίες του πίνακα. Λαμβάνει ένα Q_message με την τελευταία έκδοση του Q-πίνακα και επιστρέφει στη συνέχεια μια καινούρια έκδοση.



Εικόνα 7. Q-πίνακας

Ο πίνακας εκμάθησης Q, όπως είπαμε μας υποδεικνύει την επιλογή της καλύτερης ενέργειας προκειμένου να αγοράσουμε ένα συγκεκριμένο προϊόν υπό τις παρούσες συνθήκες σύμφωνα με τις προτιμήσεις μας. Οι συνθήκες όμως σε ένα τέτοιο δυναμικό περιβάλλον μεταβάλλονται διαρκώς επιβάλλοντας με αυτό τον τρόπο και την ταυτόχρονη ανανέωση του πίνακα Q. Στην εφαρμογή μας έχουμε τη δυνατότητα ενημέρωσης του πίνακα εκμάθησης στις περιπτώσεις:

- δημιουργίας της αγοράς
- εισαγωγής νέου κωδικού προϊόντος στην αγορά
- εισαγωγής προϊόντος που είναι ήδη διαθέσιμο στην αγορά από κάποιον mediator
- αφαίρεσης ενός προϊόντος από κάποιον mediator, το οποίο παρέχεται όμως στην αγορά από κάποιον άλλο
- πλήρης απομάκρυνση (μη διαθεσιμότητας) ενός προϊόντος από την αγορά

Το τμήμα κατασκευής των πινάκων είναι απαραίτητο να μεταχειριστεί με διαφορετικό τρόπο τις πιθανές αλλαγές που μπορούν να παρουσιαστούν στην αγορά. Διαφορετικά θα λειτουργήσει κατά την έναρξη λειτουργίας της εικονικής αγοράς, ενώ θα διαφοροποιήσει την αντίδρασή του κατά την προσθήκη ή αφαίρεση ενός προϊόντος από

έναν mediator (περικλείεται και η περίπτωση εισαγωγής νέου προϊόντος) και κατά την απομάκρυνση ενός προϊόντος από την αγορά.

Το τμήμα κατασκευής :

- στην περίπτωση έναρξης λειτουργίας της αγοράς πρέπει να ενημερώσει τις τιμές όλων των στοιχείων του Q-πίνακα
- στην περίπτωση εισαγωγής ή αφαίρεσης ενός προϊόντος τις τιμές των στοιχείων που αντιστοιχούν στο id του συγκεκριμένου προϊόντος
- ενώ κατά την εξάλειψη ενός προϊόντος από την αγορά δεν χρειάζεται να ενημερώσει κανένα στοιχείο

προκειμένου να οδηγήσει το τμήμα επιλογής στην αγορά των προϊόντων με τα επιθυμητά χαρακτηριστικά

Σε κάθε περίπτωση που απαιτεί ενημέρωση των τιμών των Q-αξιών ο Buyer επικοινωνεί με τους mediators ώστε να καταγραφούν όλα εκείνα τα στοιχεία που θα βοηθήσουν στην κατασκευή των πινάκων εκμάθησης για ένα πλήθος συγκεκριμένων προϊόντων. Πιο αναλυτικά, αποστέλλει ένα message σε έναν συγκεκριμένο mediator για κάποιο προϊόν, και ο δεύτερος του απαντά με τη σειρά του πάλι με ένα μήνυμα message με τις τιμές των χαρακτηριστικών βάσει των οποίων διαμορφώνονται οι τιμές των στοιχείων του Q-πίνακα.

Οι παράγοντες που λαμβάνονται υπόψη για την ενημέρωση των τιμών των Q-αξιών είναι οι επόμενοι:

- Βαθμός σχετικότητας προϊόντος. Δείχνει το πόσο σχετικό είναι το προϊόν που μπορεί να προσφέρει ο ενδιάμεσος με το προϊόν του οποίου η ζήτηση θα υπάρξει από τους αγοραστές.
- Τιμή προϊόντος. Το τίμημα για την ολοκλήρωση της αγοράς.
- Χρόνος απόκρισης (αγοράς). Σε πόσο χρόνο θα καταστεί δυνατή η ολοκλήρωση της αγοράς από την στιγμή που ο αγοραστής θα αποφασίσει να προχωρήσει σ' αυτή.
- Αριθμός μετακινήσεων μέχρι την τελική απόφαση. Πόσες μετακινήσεις θα κάνει ένας αγοραστής με βάση τους πίνακες εκμάθησης ώστε να φτάσει να αποφασίσει την αγορά του συγκεκριμένου προϊόντος από ένα συγκεκριμένο ενδιάμεσο.

Και ουσιαστικά αποτελούν τα κριτήρια μιας επιτυχημένης επιλογής. Θέλουμε ένα προϊόν σχετικό με αυτό που ζητάμε σε χαμηλή τιμή και σε μικρό χρονικό διάστημα. Επιπρόσθετα όμως θέλουμε να επισπεύσουμε και τη διαδικασία επιλογής. Δεν θέλουμε να χρονοτριβούμε προκειμένου να επιλέξουμε τον κατάλληλο ενδιάμεσο.

Αλγόριθμοι που χρησιμοποιούνται κατά την κατασκευή των πινάκων:

Υπολογισμός της επιβράβευσης r

Η επιβράβευση r για κάθε μία ενέργεια που πραγματοποιεί ο Buyer προκύπτει από το άθροισμα τριών επί μέρους επιβραβεύσεων α) της επιβράβευσης σχετικότητας, β) της επιβράβευσης τιμής και γ) της επιβράβευσης χρόνου απόκρισης.

Επιβράβευση Σχετικότητας

Η επιβράβευση σχετικότητας είναι ανάλογη με το βαθμό σχετικότητας του παρεχόμενου προϊόντος σε σύγκριση με το επιθυμητό και αντιστρόφως ανάλογη με τον αριθμό των μεταβάσεων μεταξύ των ενδιαμέσων που απαιτούνται προκειμένου να αποκτήσουμε το προϊόν. Με απλά λόγια όσο πιο σχετικό είναι το παρεχόμενο προϊόν με το επιθυμητό και όσο λιγότερες μεταβάσεις μεταξύ των ενδιαμέσων απαιτούνται για την απόκτησή του τόσο μεγαλύτερη και η επιβράβευση σχετικότητας. Αξίζει να σημειωθεί ότι έχουμε θέσει ένα όριο σχετικότητας κάτω από το οποίο η επιβράβευση είναι μηδενική.

Επιβράβευση Τιμής

Αν η τιμή του προϊόντος είναι μικρότερη από το μισό valuation που έχει ο Buyer τότε παίρνει επιβράβευση αλλιώς όχι. Όσο μεγαλύτερη είναι η απόσταση από το μισό valuation (προς τα κάτω) τόσο μεγαλύτερη είναι η επιβράβευση. Αξίζει να σημειωθεί ότι η τιμή της σταθεράς μπορεί να κατευθύνει και την αγορά ή μη του προϊόντος.

Επιβράβευση Χρόνου Απόκρισης

Ομοίως και σε αυτή την περίπτωση, αν ο χρόνος απόκρισης είναι μικρότερος από κάποιο επιθυμητό όριο ο πράκτορας επιβραβεύεται για την επιλογή του, σε διαφορετική περίπτωση όχι. Η επιβράβευση είναι τόσο μεγαλύτερη όσο μεγαλύτερη είναι και η απόσταση του χρόνου απόκρισης από το επιθυμητό όριο

Υπολογισμός του ρυθμού μάθησης

Μια σημαντική παράμετρος στην Q-μάθηση είναι ο ρυθμός μάθησης. Στο πρόγραμμά μας αυτός είναι αντιστρόφως ανάλογος με τη ρίζα του αριθμού των επεισοδίων που εκτελούνται κατά την εκπαίδευση. Ο ρυθμός μάθησης, επομένως, μειώνεται καθώς αυξάνεται ο αριθμός των επεισοδίων. Για πολύ μεγάλο αριθμό, γίνεται ανεπαίσθητος.

Υπολογισμός των Q-αξιών

Για τον υπολογισμό των αξιών των Q-πινάκων χρησιμοποιήσαμε τον κανόνα ανανέωσης (13) που είδαμε στο προηγούμενο κεφάλαιο.

Σημαντικές παρατηρήσεις

- Οι επιβραβεύσεις στον πίνακα μειώνονται κατά 5% όταν αφορούν καταστάσεις οι οποίες οδηγούν σε mediators που δεν έχουν το προϊόν (μπορεί να ελεγχθούν τα αποτελέσματα για διάφορες τιμές του ποσοστού).
- Το τμήμα κατασκευής των πινάκων του Buyer που κάνει την εκπαίδευση αντιπροσωπεύει όλους τους υποψήφιους αγοραστές και συνεπώς θεωρούμε ότι υπάρχει μια μέγιστη τιμή για το valuation την οποία και χρησιμοποιούμε στην κατασκευή του πίνακα εκμάθησης.
- Μια μελλοντική προέκταση θα μπορούσε να μεριμνήσει για τις περιπτώσεις εισαγωγής – εξαγωγής ενός mediator από την αγορά.

Υπολογισμός του αριθμού των επεισοδίων

Κατά τη διάρκεια της εκπαίδευσης απαιτείται ένας μεγάλος αριθμός επεισοδίων προκειμένου οι τιμές των στοιχείων των Q-πινάκων να ανταποκρίνονται στα πραγματικά χαρακτηριστικά της εικονικής αγοράς. Η μία επιλογή θα ήταν να αποδώσουμε μια μεγάλη τιμή ως αριθμό επεισοδίων η οποία θα παρέμενε σταθερή καθ' όλη τη διάρκεια 'ζωής' της αγοράς (στατική απόδοση). Μια άλλη επιλογή θα ήταν όμως αυτή την τιμή να την υπολογίζουμε κάθε φορά που απαιτείται ανανέωση των πινάκων λαμβάνοντας υπόψη τις τρέχουσες συνθήκες της αγοράς (δυναμική απόδοση).

Όπως αναφέραμε και σε προηγούμενη παράγραφο καθώς αυξάνει ο αριθμός των επεισοδίων μειώνεται ο ρυθμός μάθησης. Επιπρόσθετα ο χρόνος κατασκευής του πίνακα Q αυξάνεται καθώς αυξάνεται ο αριθμός των επαναλήψεων που απαιτούνται για την κατασκευή του. Αυτά τα δύο στοιχεία συνυπολογίσαμε μαζί με το μέγεθος του πίνακα Q προκειμένου να καταλήξουμε σε έναν τύπο ο οποίος θα προσδιορίζει τον αριθμό των επαναλήψεων που χρειάζονται κατά την περίοδο της αρχικής εκπαίδευσης. Πιο συγκεκριμένα ο αριθμός των επεισοδίων υπολογίζεται από τον τύπο:

$$NEp = c * N * (M+1)^2 \quad (34)$$

που δεν είναι τίποτα άλλο από τον πολλαπλασιασμό του μεγέθους (33) με μια σταθερά c. Ύστερα από σειρά προσομοιώσεων καταλήξαμε ότι η τιμή 4 μας δίνει αρκετά καλά αποτελέσματα. Ίσως εμπεριέχεται κάποιο ρίσκο σε αυτό το σημείο ακριβείας των τιμών

του πίνακα Q, καθώς περιορίζουμε τον αριθμό των επαναλήψεων, αλλά αν σκεφτούμε πάλι ότι ο ρυθμός μάθησης ελαττώνεται σημαντικά καθώς αυξάνεται ο αριθμός των επαναλήψεων καθώς και το χρονικό όφελος που επιφέρει αυτή η αλλαγή μπορούμε να το παραβλέψουμε. Το χρονικό όφελος είναι πιο σημαντικό στις περιπτώσεις εισαγωγής – αφαίρεσης ενός προϊόντος στην αγορά από κάποιον mediator. Μπορούμε να παραλείψουμε τον παράγοντα N από το γινόμενο υπολογισμού των επεισοδίων. Στην περίπτωση μάλιστα που κάποιο προϊόν δεν είναι διαθέσιμο από κανέναν πωλητή ο χρόνος είναι μηδενικός.

Φυσικά προσομοιώσεις έγιναν και για σταθερό αριθμό επεισοδίων. Με την επιλογή μιας σταθερής τιμής επεισοδίων (π.χ. 100000), στην περίπτωση αγορών με μικρό αριθμό mediators και διαθέσιμων προϊόντων έχουμε καλύτερη ποιότητα Q-αξιών (περισσότερες επισκέψεις σε μια συγκεκριμένη τιμή), αλλά έχουμε μεγάλες επιβραδύνσεις στην ενημέρωση του Q-πίνακα, ενώ στην περίπτωση αγορών με μεγάλο αριθμό mediators και διαθέσιμων προϊόντων έχουμε καλύτερους χρόνους ενημέρωσης του πίνακα, αλλά λιγότερο ακριβείς Q-αξίες, οι οποίες θα μας οδηγήσουν στη συνέχεια στη μη βέλτιστη αγορά προϊόντων.

Για να ολοκληρώσουμε το θέμα είναι απαραίτητο να πούμε ότι στην βιβλιογραφία δεν αναφέρεται κάπου ρητά ο επαρκής αριθμός των επεισοδίων που απαιτούνται. Περιοριζόμαστε στο ‘μεγάλος αριθμός επεισοδίων’. Ο τύπος (34) λοιπόν θα μπορούσε να θεωρηθεί κατά κάποιον τρόπο ως κάτω όριο του αριθμού των επεισοδίων που επαρκούν για την εκπαίδευση.

Συμπεριφορά του αγοραστή κατά την διαδικασία αγοράς προϊόντων

Στην παράγραφο αυτή θα παρουσιασθεί η διαδικασία αγοράς προϊόντων από το τμήμα επιλογής ενδιαμέσου του Buyer. Απαραίτητη προϋπόθεση σε κάθε περίπτωση αποτελεί η κατοχή του πλήρως ενημερωμένου Q-πίνακα.

Ξεκινά την προσπάθεια αγοράς ενός προϊόντος με τυχαία επιλογή του πρώτου mediator (αρχική κατάσταση). Θεωρούμε με αυτή την επιλογή ότι ο Buyer μπορεί να συνδιαλέγεται με αυτόν τον mediator όταν του ανατίθεται η απόκτηση του προϊόντος. Με βάση το προϊόν και τον πίνακα εκμάθησης επιλέγει την καλύτερη επόμενη μετάβαση και την εκτελεί.

Αν η καταλληλότερη κίνηση είναι η αγορά τότε υποβάλλει αίτημα αγοράς, μέσω ενός αντικειμένου της κλάσης message στον αντίστοιχο mediator (broker) και περιμένει απάντησή του.

Αν ο mediator έχει διαθέσιμο το προϊόν τότε ικανοποιεί το αίτημά του και του το παρέχει μαζί με όλα τα χαρακτηριστικά του. Διαφορετικά του επιστρέφει μήνυμα αδυναμίας ολοκλήρωσης της συναλλαγής λόγω μη διαθεσιμότητας πωλητών.

Στο σενάριο χρησιμοποιείται μια παράμετρος ποσότητας η οποία είναι η διαθεσιμότητα πωλητών για το συγκεκριμένο προϊόν. Για κάθε ενδιάμεσο υπάρχει ένας συγκεκριμένος αριθμός πωλητών για το προϊόν με τους οποίους συνεργάζεται και μπορεί να επικοινωνήσει για να είναι ικανός να διαθέσει το προϊόν στους αγοραστές. Τυχαία εξετάζουμε τις περιπτώσεις της εισόδου ή εξόδου πωλητών από κάποιους mediators.

Στην περίπτωση αδυναμίας αγοράς προϊόντος από τον συγκεκριμένο mediator, το τμήμα επιλογής του Buyer βρίσκει με βάση τον πίνακα την δεύτερη καλύτερη εναλλακτική και μεταβαίνει σ' αυτή. Αν όμως στη νέα κατάσταση η καλύτερη μετάβαση είναι η οπισθοδρόμηση στον προηγούμενο mediator τότε δεν μπορεί να πραγματοποιηθεί η αγορά. Το μήνυμα αδυναμίας αγοράς θα εμφανιστεί ακόμη και αν το προϊόν υπάρχει σε κάποιον άλλο, αλλά όχι με τα χαρακτηριστικά που επιθυμούμε. Η αδυναμία αγοράς από άλλο mediator οφείλεται στις χαμηλές τιμές του στον πίνακα που με την σειρά τους οφείλονται σε χαμηλό βαθμό σχετικότητας ή πολύ υψηλή τιμή ή πολύ μεγάλο χρόνο απόκρισης ή συνδυασμός αυτών και συνεπώς η αγορά του προϊόντος από κάποιον άλλο ενδιάμεσο δεν είναι συμφέρουσα.

Συμπεριφορά αγοράς κατά τη διάρκεια εκτέλεσης

Μια εικονική αγορά έχει έναν αρκετά δυναμικό χαρακτήρα. Παρατηρούνται έντονες και συνεχείς αλλαγές στις οποίες ο Buyer είναι απαραίτητο να προσαρμοστεί έτσι ώστε να αποδειχθεί ωφέλιμος για την πραγματοποίηση συναλλαγών.

Κατά την εκτέλεση του προγράμματος, οι mediators είναι πιθανό είτε να εισάγουν στην αγορά ένα καινούριο προϊόν, το οποίο γίνεται διαθέσιμο για πρώτη φορά, είτε να έχουν πλέον διαθέσιμο έναν κωδικό προϊόντος το οποίο δεν κατείχαν προηγουμένως αλλά ήδη υπήρχε στην αγορά. Κατ' αντιστοιχία, οι mediators μπορεί να μην έχουν την ικανότητα να μας προσφέρουν κάποιο προϊόν το οποίο κατείχαν πριν, και αν αυτό είναι γενικό, αυτό το προϊόν να μην υπάρχει πλέον στην αγορά. Επιπρόσθετα μια πιο συνήθης αλλαγή στην οποία πρέπει να προσαρμοστεί ο πράκτορας αγορών είναι η διαρκής μεταβολή της διαθεσιμότητας προμηθευτών η οποία επηρεάζει σημαντικά τη συναλλαγή.

Το πρόγραμμα αυτό για να ανταπεξέλθει μεν στις αλλαγές που αφορούν την εισαγωγή ή την απομάκρυνση ενός προϊόντος στην αγορά ειδοποιεί τον πράκτορα που

πραγματοποιεί την εκπαίδευση να δημιουργήσει καινούριους πίνακες Q, ενώ για τις διαφοροποιήσεις στις τιμές της διαθεσιμότητας ο πράκτορας Buyer ανατρέχει στην δεύτερη καλύτερη εναλλακτική επιλογή η οποία ικανοποιεί τις απαιτήσεις μας (αν υπάρχει φυσικά).

4.2 Επίλογος

Στο κεφάλαιο αυτό προσπαθήσαμε να περιγράψουμε το σενάριο της εικονικής αγοράς που μελετήσαμε. Αναφερθήκαμε στις οντότητες που λαμβάνουν μέρος σε αυτό το σκηνικό και πως αυτές αλληλεπιδρούν. Ο πράκτορας Buyer πραγματοποιεί αγορές εκ μέρους μας σε αυτό το δυναμικό περιβάλλον. Χρησιμοποιεί τον αλγόριθμο της Q-μάθησης προκειμένου να επιλέξει τον κατάλληλο ενδιάμεσο που διαθέτει το προϊόν που ανταποκρίνεται καλύτερα στις απαιτήσεις και τις ανάγκες μας. Χαρακτηριστικά που κατευθύνουν την επιλογή του είναι ο βαθμός σχετικότητας του προϊόντος, η τιμή του, ο χρόνος απόκρισης του καθώς και ο αριθμός των μεταβάσεων που απαιτούνται για την απόκτησή του. Οι Mediators παίζουν το ρόλο των Brokers στο συγκεκριμένο σενάριο και αποκρύπτουν οποιαδήποτε πληροφορία για τους πωλητές των προϊόντων. Ο αριθμός τους είναι σταθερός καθ' όλη τη διάρκεια λειτουργίας της αγοράς.

Το σενάριό μας μελετά τις περιπτώσεις εισαγωγής και εξαγωγής προϊόντων από την αγορά σε τυχαίες στιγμές στο χρόνο. Μια μελλοντική επέκταση η οποία θα μελετούσε τις αλλαγές στην εικονική αγορά από την είσοδο ή έξοδο ενός mediator θα κρινόταν ενδιαφέρουσα.

ΚΕΦΑΛΑΙΟ 5
ΤΕΚΜΗΡΙΩΣΗ

5.1 Εισαγωγή

Καθώς ολοκληρώθηκε το στάδιο της κωδικοποίησης ακολούθησε στη συνέχεια ένα πολύ κρίσιμο για την εφαρμογή μας στάδιο. Το στάδιο της αξιολόγησης. Ήταν απαραίτητο να συγκρίνουμε τη συμπεριφορά και τις επιδόσεις του προγράμματος με κάποιο μοντέλο πρότυπο προκειμένου να εκμεταλλευθούμε τα οφέλη που προκύπτουν και να προσπαθήσουμε να βελτιώσουμε τις όποιες αδυναμίες του.

Το πρώτο ερώτημα που προκύπτει, λοιπόν, είναι ποιο μοντέλο θα αποτελέσει το μέτρο σύγκρισης. Ορίσαμε σαν μέτρο σύγκρισης το μοντέλο όπου ο αγοραστής κάθε φορά που επιθυμεί να αποκτήσει ένα προϊόν ρωτά όλες τις ενδιάμεσες οντότητες για τη διαθεσιμότητα του συγκεκριμένου προϊόντος και στη συνέχεια επιλέγει αυτό που ανταποκρίνεται καλύτερα στις απαιτήσεις του. Σε κάθε αγορά λοιπόν απαιτούνται Μ + 1 βήματα, όπου Μ είναι ο αριθμός των ενδιαμέσων. Στο μοντέλο αυτό ο αγοραστής δεν χρησιμοποιεί καθόλου γνώσεις από προηγούμενες εμπειρίες του στην εικονική αγορά.

Στην Εικόνα 8 παρουσιάζεται η αλληλεπίδραση του αγοραστή με τους ενδιάμεσους. Τα αμφίδρομα βέλη υποδηλώνουν τις ερωτήσεις – απαντήσεις των ενδιαμέσων για τη διαθεσιμότητα ενός προϊόντος ενός το μπλε την απόφαση αγοράς από έναν συγκεκριμένο ενδιάμεσο.

Το δεύτερο ερώτημα που προκύπτει αμέσως είναι τα κριτήρια βάσει των οποίων θα γίνει η αξιολόγηση. Μερικά θα μπορούσαν να είναι τα ακόλουθα:

α) ποιο μοντέλο μας οδηγεί στις οικονομικότερες προτάσεις;

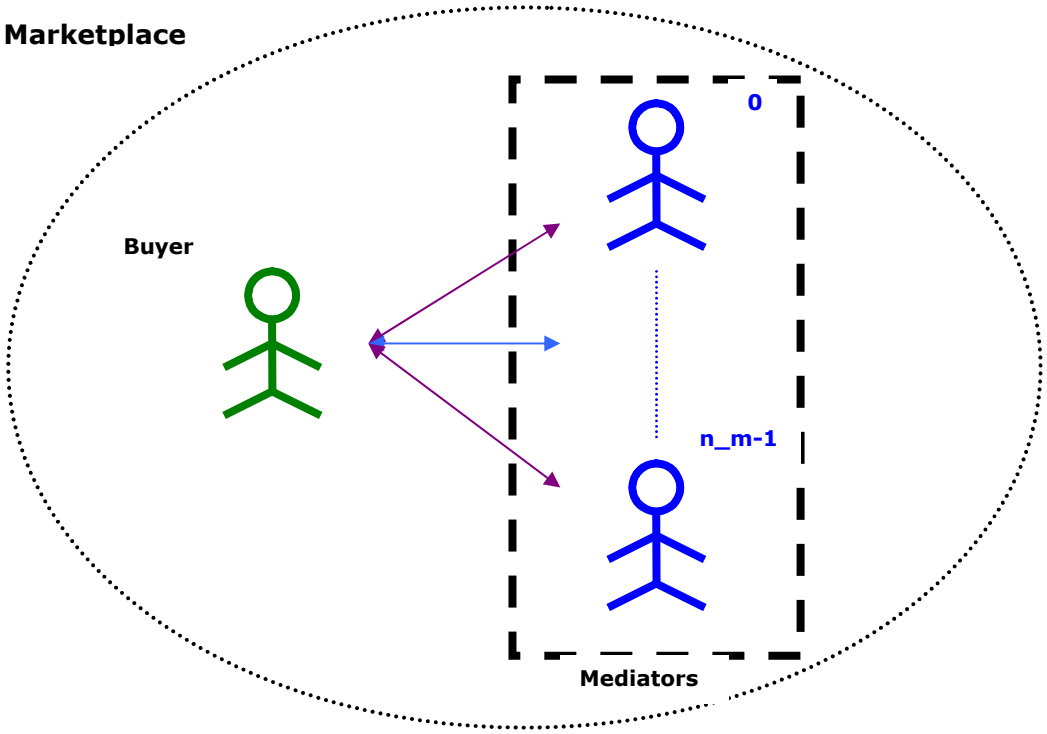
β) ποιο μοντέλο μας παρέχει προϊόντα τα οποία παρουσιάζουν το μεγαλύτερο βαθμό σχετικότητας με τις απαιτήσεις μας;

γ) ποιο μοντέλο λαμβάνει πιο γρήγορες αποφάσεις, πραγματοποιεί πιο γρήγορα τις αγορές εκ μέρους μας;

Στα πρώτα δύο η απάντηση μοιάζει να είναι εύκολη. Στην περίπτωση του μοντέλου που αποτελεί το μέτρο σύγκρισης καθόσον έχουμε εξαντλητική αναζήτηση έχουμε και την εγγύηση της βέλτιστης επιλογής από άποψη τιμής ή σχετικότητας. Και στο δικό μας μοντέλο όμως αυτές οι απαιτήσεις μπορούν να ικανοποιηθούν όταν δώσουμε τις κατάλληλες επιβραβεύσεις και τον κατάλληλο αριθμό επεισοδίων κατά τη διάρκεια της εκπαίδευσης. Εάν θέλουμε να επιλέγουμε πάντα την οικονομικότερη πρόταση θα

δώσουμε μια μεγάλη επιβράβευση, ενώ κάτι αντίστοιχο είναι απαραίτητο να κάνουμε και στην περίπτωση που θέλουμε το προϊόν με τη μεγαλύτερη σχετικότητα.

Το τρίτο ερώτημα όμως, παραμένει ανοικτό, και απαιτεί περαιτέρω διερεύνηση. Θα αποτελέσει ουσιαστικά το κριτήριο σύγκρισης των δύο μοντέλων. Θα πρέπει να κατασκευάσουμε διάφορα σενάρια στα οποία θα καταγράψουμε τους χρόνους στους οποίους τα μοντέλα επιτυγχάνουν την αγορά των προϊόντων.



Εικόνα 8. Μοντέλο Σύγκρισης

Χρόνοι απόκτησης προϊόντων

Όπως αναφέραμε προηγουμένως, το μοντέλο που θα χρησιμοποιήσουμε σαν μέτρο σύγκρισης πραγματοποιεί μια αγορά προϊόντος σε M + 1 βήματα, οπότε ο συνολικός χρόνος αγοράς N προϊόντων δίνεται από τον τύπο:

$$T = (M+1) * N * t_m \quad (35)$$

όπου ο χρόνος t_m αντιστοιχεί στο χρόνο επικοινωνίας του αγοραστή με τους ενδιαμέσους. Στο δικό μας μοντέλο ο συνολικός χρόνος απόκτησης N προϊόντων υπολογίζεται από τον τύπο:

$$T_p = \sum_{i=1}^{N} T_{ci} + \sum_{j=1}^{K} M_j \quad (36)$$

όπου T_{ci} είναι ο χρόνος κατασκευής και ανανέωσης του πίνακα i (ο πιο καθοριστικός παράγοντας διαμόρφωσης του τελικού συνολικού χρόνου), M_j ο χρόνος αναζήτησης της

επόμενης κίνησης βάσει των Q - πινάκων και ο χρόνος επικοινωνίας με τους ενδιαμέσους, και K ο αριθμός των μετακινήσεων που απαιτούνται για την απόκτηση των προϊόντων.

5.2 Προσομοιώσεις

Εκ των προτέρων γνωρίζαμε ότι το μεγαλύτερο ποσοστό του χρόνου που θα κατανάλωνε το μοντέλο μας για την αγορά ενός προϊόντος θα αναφέρονταν στη διαδικασία εκμάθησης, στη διαδικασία κατασκευής/ενημέρωσης των Q-πινάκων. Ο αλγόριθμος της Q-μάθησης αναφέρει ότι απαιτείται ένας μεγάλος αριθμός επεισοδίων κατά τη διαδικασία της εκπαίδευσης προκειμένου οι Q-αξίες να συγκλίνουν προς κάποια συγκεκριμένη τιμή. Αλλά ακόμα και εάν δεν φτάσουμε σε αυτό επίπεδο, ένας μεγάλος σχετικά αριθμός επεισοδίων μας διασφαλίζει καλύτερη ποιότητα Q-αξιών. Στις Q-αξίες θα καθρεπτίζονταν πιστότερα τα χαρακτηριστικά του ιδανικού προϊόντος. Ποιος όμως θα ήταν αυτός ο αριθμός που θα διασφάλιζε καλή ποιότητα Q-αξιών και ταυτόχρονα δεν θα επιβάρυνε πολύ, χρονικά, τη διαδικασία της αγοράς; Ένα ερώτημα που μας απασχόλησε καθ' όλη τη διάρκεια των προσομοιώσεων. Κατά τη διάρκεια των προσομοιώσεων ακολουθήσαμε δύο διαφορετικούς δρόμους. Στον πρώτο αρκεστήκαμε στην στατική απόδοση μιας σταθερής τιμής ενός μεγάλου αριθμού επεισοδίων, ενώ στο δεύτερο επιχειρήσαμε αυτός ο αριθμός να υπολογίζεται δυναμικά. Στη δεύτερη περίπτωση ο αριθμός των επεισοδίων υπολογιζόταν κάθε φορά που γινόταν κάποια ενημέρωση των πινάκων βάσει του μεγέθους τους και χρησιμοποιώντας τη σχέση (34).

Αποτελέσματα

Πραγματοποιήσαμε πάνω από 60 set προσομοιώσεων για διάφορους συνδυασμούς τιμών των διάφορων βασικών παραμέτρων του σεναρίου μας. Τα αποτελέσματα ήταν ενθαρρυντικά και επιδεικνύουν μια σημαντική μείωση του χρόνου που απαιτείται για την απόκτηση 400 προϊόντων. Η μόνη περίπτωση όπου το μοντέλο μας επιφέρει χρόνους που κυμαίνονται στα ίδια επίπεδα με το μοντέλο σύγκρισης είναι όταν χρησιμοποιήσαμε έναν πολύ μεγάλο αριθμό επεισοδίων.

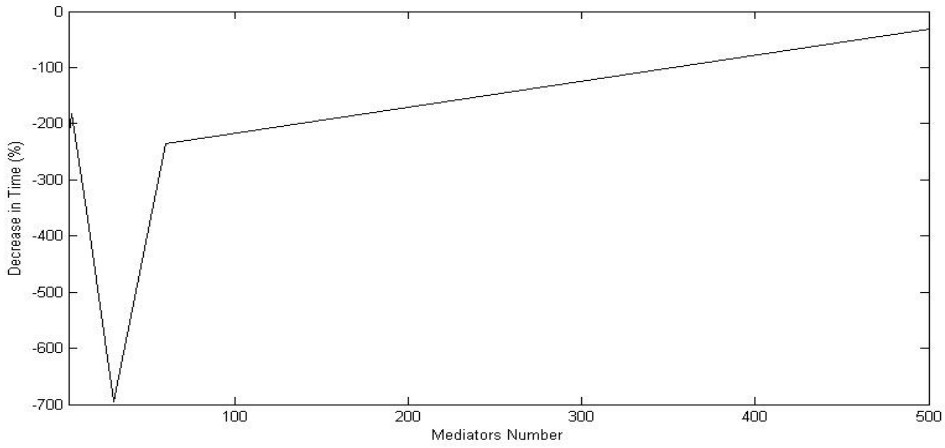
Α) Σταθερός αριθμός επεισοδίων

Ο Πίνακας 1 δείχνει ενδεικτικά τη μείωση του χρόνου απόκτησης των προϊόντων όταν στο μοντέλο μας είχαμε 4, 6, 15, 30, 60 και 500 ενδιαμέσους που ο καθένας μπορεί να προτείνει μέχρι 10 κωδικούς προϊόντων και είχαμε 10000 επεισόδια για την κατασκευή των Q-πινάκων.

Πίνακας 1. Αποτελέσματα με 10 κωδικούς προϊόντων και σταθερό αριθμό επεισοδίων

Αριθμός Ενδιαμέσων	Χρόνος Κατασκευής Πινάκων (ms)	Μέσες Κινήσεις	Συνολικός Χρόνος Απόκτησης (ms)	Αύξηση/Μείωση στο Χρόνο Απόκτησης
4	656	1.66	1982	-101.82%
6	608	1.73	1986	-181.97%
15	1342	1.74	2732	-368.52%
30	1747	1.72	3117	-695.64%
60	13104	1.81	14548	-235.44%
500	304357	1.87	305847	-31.05%

Παρατηρούμε τη σημαντική μείωση του χρόνου απόκτησης του προϊόντος που φτάνει στη μέγιστη τιμή της όταν έχουμε 30 ενδιαμέσους. Στη συνέχεια όμως, η ψαλίδα συνεχώς κλείνει. Η αύξηση του αριθμού των ενδιαμέσων επιφέρει αύξηση του χρόνου κατασκευής των Q-πινάκων και κατ' επέκταση του χρόνου απόκτησης των προϊόντων από το μοντέλο μας. Έπεται η γραφική παράσταση (Εικόνα 9) που αποτυπώνει τα αποτελέσματα του προηγούμενου πίνακα.



Εικόνα 9.Μείωση χρόνου για 10 κωδικούς προϊόντων και σταθερός αριθμός επεισοδίων

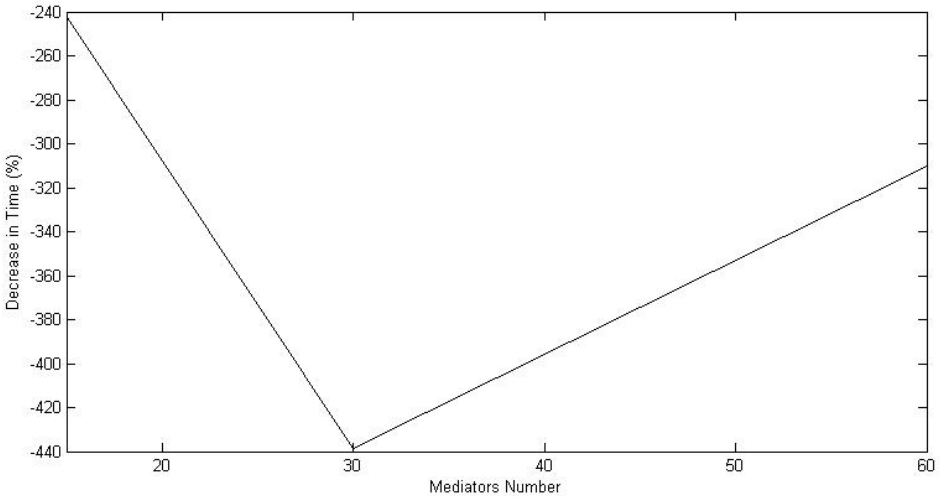
Στη συνέχεια παρουσιάζεται ένας πίνακας που δείχνει ενδεικτικά τα αποτελέσματα στην περίπτωση που στο μοντέλο μας είχαμε 15, 30 και 60 ενδιαμέσους, κάθε ένας από τους οποίους μπορεί να προτείνει στον αγοραστή μέχρι 40 διαφορετικά προϊόντα και είχαμε 10000 επεισόδια κατά τη διάρκεια εκπαίδευσης του Buyer. Η διαφορά παραμένει εξίσου σημαντική. Η τάξη της διαφοράς κυμαίνεται σε τριψήφια επίπεδα με μέγιστη τιμή στην περίπτωση που είχαμε 30 mediators. Η περαιτέρω αύξηση του αριθμού των mediators

επιβαρύνει σημαντικά το χρόνο απόκτησης των προϊόντων καθώς εισάγεται σημαντική επιβράδυνση κατά την κατασκευή των πινάκων της εκπαίδευσης.

Πίνακας 2. Αποτελέσματα με 40 κωδικούς προϊόντων και σταθερό αριθμό επεισοδίων

Αριθμός Ενδιαμέσων	Χρόνος Κατασκευής Πινάκων (ms)	Μέσες Κινήσεις	Συνολικός Χρόνος Απόκτησης (ms)	Αύξηση/Μείωση στο Χρόνο Απόκτησης
15	2377	1.71	3739	-242.34%
30	3245	1.7	4603	-438.78%
60	10546	1.7	11902	-310.02%

Ακολουθεί η γραφική παράσταση αυτών των αποτελεσμάτων. Παρατηρούμε την μέγιστη τιμή της διαφοράς όταν στην αγορά μετέχουν 30 mediators και πως η διαφορά μικραίνει καθώς αυξάνεται στη συνέχεια ο αριθμός τους.



Εικόνα 10.Μείωση χρόνου για 40 κωδικούς προϊόντων και σταθερός αριθμός επεισοδίων

Β) Μεταβλητός αριθμός επεισοδίων

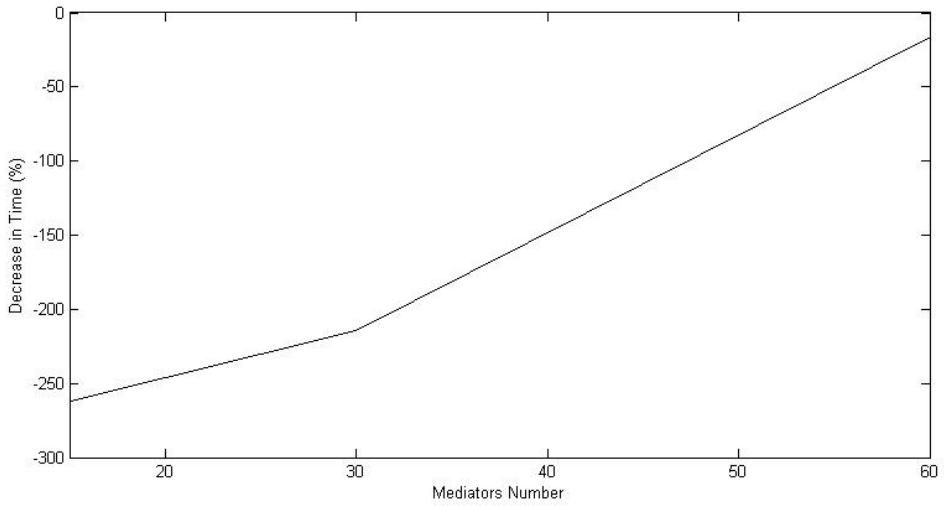
Πραγματοποιήσαμε και προσομοιώσεις κατά τις οποίες ο αριθμός των επεισοδίων υπολογίζεται δυναμικά κάθε φορά που κατασκευάζεται ή ενημερώνεται ο Q-πίνακας. Η σκέψη μας ήταν να προσδιορίσουμε τον αριθμό των επεισοδίων της εκπαίδευσης βασιζόμενοι στις διαστάσεις του προβλήματος. Ο αριθμός των ενδιαμέσων και ο αριθμός των διαθέσιμων προϊόντων θα καθόριζε τώρα τον αριθμό των επεισοδίων και δεν θα αποδίδαμε μια πολύ μεγάλη τιμή όπως προηγουμένως για να εκπαιδεύσουμε τον πράκτορα. Τα αποτελέσματα κι εδώ ήταν ενθαρρυντικά. Ακολουθεί ένας πίνακας που παρουσιάζει τη μείωση που επέφερε το μοντέλο μας στην περίπτωση που είχαμε 15, 30, 60 ενδιαμέσους κάθε ένας από τους οποίους είχε τη δυνατότητα να συστήσει

στην αγορά μέχρι 40 προϊόντα. Παρατηρούμε ότι όσο αυξάνει ο αριθμός των ενδιαμέσων μειώνεται η διαφορά μεταξύ του μοντέλου μας και του μοντέλου πρότυπο. Η μέγιστη τιμή της διαφοράς παρατηρείται στην περίπτωση που στην αγορά δρουν 15 ενδιάμεσοι.

Πίνακας 3. Αποτελέσματα με 40 κωδικούς προϊόντων και μεταβλητό αριθμό επεισοδίων

Αριθμός Ενδιαμέσων	Χρόνος Κατασκευής Πινάκων (ms)	Μέσες Κινήσεις	Συνολικός Χρόνος Απόκτησης (ms)	Αύξηση/Μείωση στο Χρόνο Απόκτησης
15	2123	1.77	3533	-262.30%
30	6473	1.76	7879	-214.76%
60	40434	1.76	41838	-16.64%

Ακολουθεί η γραφική παράσταση (Εικόνα 11) η οποία αποτυπώνει οπτικά τη μείωση που επιτελείται με τη βοήθεια του μοντέλου μας κάνοντας χρήση τον δυναμικό υπολογισμό των επεισοδίων.



Εικόνα 11.Μείωση χρόνου για 40 κωδικούς προϊόντων και μεταβλητός αριθμός επεισοδίων

Σχέση αριθμού επεισοδίων ποιότητας επιλογών προϊόντων

Εκ των προτέρων γνωρίζαμε ότι ένας μεγάλος αριθμός επεισοδίων θα είχε σαν αποτέλεσμα καλύτερη ποιότητα επιλογών προϊόντων από τον πράκτορα – αγοραστή. Ταυτόχρονα όμως ένας μεγάλος αριθμός επεισοδίων συνεπάγεται και μεγάλους χρόνους κατασκευής των πινάκων και ακολούθως μεγάλους χρόνους απόκτησης των αγορών. Θα ήταν επομένως ενδιαφέρον να δούμε πως η αύξηση του αριθμού των επεισοδίων επιδρά στην ποιότητα των αγορών. Ακολουθεί πίνακας (Πίνακας 4) που

παρουσιάζει τα ποιοτικά χαρακτηριστικά των αγορών για διάφορους αριθμούς επεισοδίων.

Πίνακας 4. Ποιοτικά χαρακτηριστικά αγορών για διάφορους αριθμούς επεισοδίων

Αριθμός Επεισοδίων	Κινήσεις	Μέση Σχετικότητα	Μέση Τιμή Απόκτησης
50	825	0.624	11.67
100	737	0.699	10.03
1000	693	0.769	10.42
10000	707	0.813	6.19
100000	668	0.832	6.24
500000	682	0.843	7.47
1000000	675	0.842	6.54

Αν ρίξουμε μια προσεκτική ματιά στον πίνακα θα δούμε ότι ο δεκαπλασιασμός π.χ. του αριθμού των επεισοδίων δεν επιφέρει ταυτόχρονα τόσο μεγάλη βελτίωση ως προς τους ποιοτικούς δείκτες των αγορών. Ένα συμπέρασμα αρκετά χρήσιμο, διότι μας επιτρέπει να ρισκάρουμε λίγο ως προς των αριθμό των επεισοδίων κατά τη διάρκεια της εκπαίδευσης. Θα μπορούσαμε να χρησιμοποιήσουμε έναν μικρό σχετικά αριθμό επεισοδίων με αποτέλεσμα μικρές απώλειες ως προς την ποιότητα, αλλά με ταυτόχρονη μεγάλη μείωση του χρόνου κατασκευής των πινάκων.

5.3 Επίλογος

Η φάση των προσομοιώσεων θα αποτελούσε το στάδιο κατά το οποίο θα εξάγονταν χρήσιμα συμπεράσματα για το μοντέλο μας και τα οποία θα αποτελούσαν ερεθίσματα για περαιτέρω μελέτη. Στο σύνολο των προσομοιώσεων, το μοντέλο μας επέφερε σημαντική μείωση του χρόνου που απαιτείται για την απόκτηση προϊόντων, η τιμή της οποίας κάποιες φορές έφτανε και σε τριψήφια επίπεδα. Παράλληλα όμως παρατηρήσαμε ότι η διαφορά αυτή μειώνεται καθώς αυξάνεται ο αριθμός των ενδιαμέσων λόγω της σημαντικής επιβάρυνσης κατασκευής των Q-πινάκων. Εκτός όμως από τη σύγκριση του των χρόνων απόκτησης προϊόντων του μοντέλου μας και του μοντέλου προτύπου, μελετήσαμε επιπρόσθετα και τις διαφορές της ποιότητας των επιλογών του μοντέλου μας καθώς μεταβάλλεται ο αριθμός των επεισοδίων. Μια μεγάλη μείωση του αριθμού των επεισοδίων δεν συνεπάγεται μια εξίσου μεγάλη σχετική μείωση της ποιότητας των προϊόντων που αποκτώνται.

ΚΕΦΑΛΑΙΟ 6
ΣΥΜΠΕΡΑΣΜΑΤΑ - ΕΠΙΛΟΓΟΣ

6.1 Συμπεράσματα

Μετά το πέρας των προσομοιώσεων και την καταγραφή των αποτελεσμάτων και των παραμέτρων που παρουσιάζουν κάποιο ενδιαφέρον για εμάς, έφτασε η στιγμή της κρίσης και σύγκρισης του μοντέλου μας. Η διαδικασία αυτή μας οδήγησε στην εξαγωγή χρήσιμων συμπερασμάτων - παρατηρήσεων τα οποία παρουσιάζουμε αναλυτικότερα στη συνέχεια.

Χρόνος πραγματοποίησης αγορών

Αποτελεί την παράμετρο που θα έκρινε ουσιαστικά την επιτυχία ή την αποτυχία του μοντέλου μας. Η σύγκριση των χρόνων των δύο μοντέλων θα καταδείκνυε αν πράγματι το υπολογιστικό κόστος κατασκευής των Q-πινάκων θα μας ωφελούσε χρονικά κατά τη διαδικασία απόκτησης προϊόντων. Τα αποτελέσματα ήταν άκρως ενθαρρυντικά για εμάς στην πλειοψηφία των προσομοιώσεων. Είτε στην περίπτωση που είχαμε σταθερό αριθμό επεισοδίων, είτε στην περίπτωση δυναμικού υπολογισμού κάθε φορά που γινόταν η κατασκευή ή η ανανέωση των τιμών των Q-πινάκων είχαμε σημαντική μείωση των χρόνων. Μάλιστα η μείωση όταν κάναμε αναφορά σε σχετικά μεγέθη με χρήση ποσοστών έφτανε σε επίπεδα τριψήφιων αριθμών. Μειώσεις της τάξης του 695,64% και 438,78% ήταν αξιοσημείωτες.

Ο αρχικός μας φόβος ότι το μεγάλο υπολογιστικό κόστος που επιφέρει η κατασκευή και συντήρηση, ανανέωση των Q-πινάκων εξαλείφθηκε με την καταγραφή των αποτελεσμάτων. Η χρονική ωφέλεια ήταν σημαντική και θα ξεδιάλυνε όποιες αρχικές αμφιβολίες.

Αριθμός επεισοδίων

Όπως ήταν γνωστό, η μεγαλύτερη χρονική επιβάρυνση για την αγορά προϊόντων χρησιμοποιώντας το μοντέλο μας θα οφειλόταν στο μεγάλο χρόνο κατασκευής των πινάκων. Σε ένα συγκεκριμένο σκηνικό ενδιαμέσων και διαθέσιμων προϊόντων ο χρόνος αυτός επηρεάζεται από τον αριθμό των επεισοδίων.

Στην προσπάθειά μας να περιορίσουμε όσο το δυνατόν περισσότερο την χρονική επιβάρυνση που επιφέρει η κατασκευή/ανανέωση των πινάκων χωρίς όμως να επηρεάζουμε σημαντικά την ποιότητα των αγορών ως προς τα επιθυμητά χαρακτηριστικά των προϊόντων θελήσαμε να διερευνήσουμε αν μια σημαντική μείωση

του αριθμού των επεισοδίων, με άμεσο όφελος την αξιόλογη μείωση του χρόνου αγορών, θα είχε σημαντική επίπτωση και στην ποιότητα των αγορών μας, γεγονός μη επιθυμητό βεβαίως.

Τα αποτελέσματα και σε αυτή την περίπτωση, ήταν θετικά. Μια σημαντική μείωση του αριθμού των επεισοδίων, έχοντας πάντα κατά νου τις διαστάσεις του προβλήματος, δεν επιφέρει ταυτόχρονη μεγάλη μείωση στα ποιοτικά χαρακτηριστικά των αγορών, γεγονός που μας δίνει ένα περιθώριο ρίσκου, περιορισμού του αριθμού αυτού, με άμεσο όφελος τη σημαντική μείωση του χρόνου κατασκευής των πινάκων. Ένα χαρακτηριστικό παράδειγμα που αποδεικνύει αυτό τον ισχυρισμό για κάποιο συγκεκριμένο σκηνικό ήταν ο πενταπλασιασμός του αριθμού των επεισοδίων (από 100.000 σε 500.000) που βελτιώνει τον μέσο παράγοντα σχετικότητας μόλις κατά 1,3% (από 0,832 σε 0,843).

Κινήσεις για αγορά

Ένα άλλο στοιχείο το οποίο χρήζει ενδιαφέροντος είναι ο αριθμός των κινήσεων που χρειάζεται ο πράκτορας Buyer για να αποκτήσει ένα προϊόν. Είναι απαραίτητο να σημειώσουμε ότι όταν λέμε κινήσεις αναφερόμαστε στις μετακινήσεις στους διαφορετικούς ενδιάμεσους που απαιτούνται για την εύρεση του πιο κατάλληλου παρόχου, δηλαδή, του ενδιαμέσου που μας προσφέρει το προϊόν που ανταποκρίνεται καλύτερα στις δικές μας απαιτήσεις από τη στιγμή όμως που έχει κατασκευαστεί – ενημερωθεί ο Q-πίνακας. Ο αριθμός αυτός παραμένει στα ίδια επίπεδα άσχετα από το πλήθος των ενδιάμεσων ή των προϊόντων που διαπραγματεύονται. Μετά από το σύνολο των προσομοιώσεων ο μέσος όρος βημάτων για την αγορά των 400 προϊόντων ισούται με 688,56 το οποίο ισοδυναμεί με περίπου 1,72 βήματα ανά προϊόν. Και μάλιστα η μονάδα αναφέρεται στην δήλωση της επιθυμίας αγοράς του προϊόντος από κάποιον συγκεκριμένο ενδιάμεσο. Ο αριθμός αυτός στην περίπτωση του μοντέλου προτύπου ανέρχεται σε 51 όταν έχουμε για παράδειγμα 50 ενδιάμεσες οντότητες.

6.2 Κριτική του Μοντέλου

Αν θέλαμε να κάνουμε μια γενική αποτίμηση των αποτελεσμάτων που παρήχθησαν θα λέγαμε ότι αυτά κρίνονται ικανοποιητικά. Στην πλειοψηφία των προσομοιώσεων είχαμε αξιόλογη μείωση του χρόνου αγοράς ενός προϊόντος, στοιχείο βαρύνουσας σημασίας για την ερευνητική περιοχή των αγορών πληροφορίας. Οι αρχικοί μας ενδοιασμοί για την χρονική επιβάρυνση που θα επέφερε η κατασκευή των Q-πινάκων, προκειμένου να κατασκευάσουμε το μοντέλο της αγοράς, δεν επιβεβαιώθηκαν. Οι σχετικοί χρόνοι μειώθηκαν σε ποσοστά που έφταναν και τριψήφια μεγέθη.

Αν μάλιστα κάποιος θελήσει να ακολουθήσει μια πιο ριψοκίνδυνη πολιτική ως προς τον αριθμό των επεισοδίων που απαιτούνται για την κατασκευή των πινάκων, μειώνοντας τον σημαντικά θα είχε ακόμα μεγαλύτερα οφέλη από άποψη χρόνου, χωρίς ταυτόχρονα να έχει αξιόλογη αλλοίωση των ποιοτικών χαρακτηριστικών των αγορών. Παρόλα αυτά τα θετικά μηνύματα που λάβαμε από τα αποτελέσματα των προσομοιώσεων είναι σίγουρο ότι υπάρχουν κάποιες ενδιαφέρουσες επεκτάσεις οι οποίες θα προσέδιδαν νέα χαρακτηριστικά στο μοντέλο μας και θα έδιδαν ερέθισμα για περαιτέρω διερεύνηση.

Μελλοντικές Προεκτάσεις

Α) Δυναμική είσοδος – έξοδος mediator

Στο μοντέλο μας έχουμε έναν συγκεκριμένο αριθμό ενδιαμέσων στους οποίους δίδεται η δυνατότητα να εισάγουν/εξάγουν σε τυχαίες χρονικές στιγμές προϊόντα από την αγορά. Αν φέρουμε στην μνήμη μας λίγο τη δομή των Q-πινάκων που υιοθετήσαμε για αυτό το μοντέλο, η είσοδος ενός καινούριου προϊόντος στην αγορά θα επηρέαζε την πρώτη διάσταση τους. Θα είχαμε προσθήκη ενός καινούριου κωδικού και ταυτόχρονα ενημέρωση των Q-αξιών που αντιστοιχούν σε αυτόν τον κωδικό. Στην περίπτωση της απομάκρυνσης ενός προϊόντος εξ ολοκλήρου από την αγορά θα είχαμε αφαίρεση ενός κωδικού προϊόντος. Στην περίπτωση, τώρα, της εισόδου/εξόδου ενός προϊόντος που ήταν ή θα είναι διαθέσιμο αντίστοιχα μετά από αυτή την μεταβολή στην αγορά έχουμε απλά ενημέρωση των Q-αξιών.

Σε μια αγορά πληροφορίας, όμως, είναι απαραίτητο να διερευνηθεί και η δυναμική είσοδος/έξοδος ενδιαμέσων οντοτήτων σε τυχαίες χρονικές στιγμές. Τι αλλαγές θα επέφερε μια τέτοια επιπρόσθετη δυνατότητα στο μοντέλο μας; Για τι άλλο θα έπρεπε να μεριμνήσουμε; Σίγουρα η είσοδος και η έξοδος μιας ενδιάμεσης οντότητας θα επηρέαζε τη δεύτερη και τρίτη διάσταση των Q-πινάκων καθώς θα είχαμε ταυτόχρονη προσθήκη/αφαίρεση ενός στοιχείου σε αυτές.

Β) Χρήση αλγορίθμου Q(λ)

Στο μοντέλο μας για την απόδοση τιμών στα στοιχεία των Q-πινάκων χρησιμοποιούμε τον αλγόριθμο της Q-μάθησης (Q-learning). Ο κανόνας ανανέωσης της Q-μάθησης, όμως, διαδίδει την πληροφορία μόνο για ένα βήμα κατά μήκος της τροχιάς του πράκτορα στο διάστημα των καταστάσεων.

Τι οφέλη θα καρπωνόμασταν εάν χρησιμοποιούσαμε τον αλγόριθμο Q(λ); Με τη χρήση αυτού του αλγορίθμου ο πράκτορας αποδίδει στην τροχιά που σχηματίζει με την κίνησή του ένα ίχνος εκλεξιμότητας. Σε κάθε βήμα ανανεώνει όχι μόνο την τελευταία Q-αξία

αλλά ολόκληρο το σύνολο των Q-αξιών που αντιστοιχούν στην τροχιά, αναλογικά στις τιμές της εκλεξιμότητά τους. Θα είχαμε πράγματι κάποιο χρονικό όφελος;

Γ) Συμμετοχή πολλαπλών αγοραστών

Στο μοντέλο μας κυρίαρχο ρόλο παίζει αναμφίβολα ο πράκτορας Buyer. Ένα τμήμα του επιβαρύνεται με την κατασκευή των Q-πινάκων και το άλλο με την πραγματοποίηση αγορών βασιζόμενο σε αυτούς τους πίνακες εκ μέρους του χρήστη του.

Μια ενδιαφέρουσα επέκταση του μοντέλου μας θα ήταν η συμμετοχή πολλών πρακτόρων αγοραστών Buyer. Και πιο συγκεκριμένα, θα ήταν σίγουρα ενδιαφέρον να μελετούσαμε τα οφέλη και τις δυσκολίες βέβαια που θα συνεπαγόταν η χρήση παραλληλίας για την κατασκευή των πινάκων από διάφορους πράκτορες Buyer. Θα μπορούσαμε να μεταδώσουμε τη γνώση (τους Q-πίνακες) σε περισσότερους Buyer;

Δ) Άλλες προεκτάσεις

Σίγουρα αυτές είναι κάποιες προεκτάσεις που παρουσιάζουν κάποιο ενδιαφέρον αλλά δεν είναι οι μόνες. Κάποιος θα μπορούσε να διαφοροποιήσει τον τρόπο απόδοσης τιμών στους Q-πίνακες. Θα μπορούσε να διαφοροποιήσει τους παράγοντες που διαμορφώνουν την επιβράβευση για κάθε ενέργεια των πρακτόρων αγοραστών. Θα μπορούσε να μελετήσει διαφορετικό τρόπο αποτύπωσης του μοντέλου στους Q-πίνακες, να δει με άλλη οπτική γωνία τις διακριτές καταστάσεις του μοντέλου. Θα μπορούσε να μελετήσει λεπτομερέστερα την επίπτωση του αριθμού των επεισοδίων που είναι απαραίτητα για την κατασκευή των Q-πινάκων στο συνολικό χρόνο απόκτησης των προϊόντων. Και η καταγραφή δεν μπορεί να σταματήσει εδώ… Είναι ανεξάντλητη.

6.3 Επίλογος

Κλείνοντας, θα προσπαθήσουμε να επαναφέρουμε στη μνήμη μας τα θέματα με τα οποία ασχοληθήκαμε κατά τη διάρκεια συγγραφής αυτής της εργασίας. Ξεκινήσαμε με τη μελέτη των έννοιών Πράκτορας Λογισμικού, Πολυπρακτορικά Συστήματα και Αγορές Πληροφορίας. Πιο συγκεκριμένα, μελετήσαμε τα χαρακτηριστικά των πρακτόρων λογισμικού και τους κατηγοριοποιήσαμε με βάση αυτά. Όταν σταθήκαμε στα πολυπρακτορικά συστήματα, είδαμε τα διάφορα χαρακτηριστικά τους, και με ιδιαίτερο ενδιαφέρον μελετήσαμε τα προβλήματα της επικοινωνίας και διαπραγμάτευσης που συναντώνται μεταξύ των πρακτόρων που συμμετέχουν σε αυτά. Στη συνέχεια κάνοντας προέκταση σε μια αγορά πληροφορίας όπως αυτή που θέλαμε να μοντελοποιήσουμε

εξετάσαμε τις διάφορες οντότητες που συμμετέχουν σε αυτές καθώς και τον τρόπο που πρέπει να συνεργάζονται – επικοινωνούν για να επιτύχουν τους στόχους τους.

Στη συνέχεια ασχοληθήκαμε με την λεπτομερή μελέτη και καταγραφή των τεχνικών της Ενισχυτικής Μάθησης. Διερευνήσαμε τους διάφορους αλγορίθμους και τα χαρακτηριστικά τους και στην περίπτωση συστημάτων που αφορούν έναν πράκτορα καθώς και σε συστήματα με πολλούς πράκτορες.

Ο επόμενος σταθμός ήταν η μοντελοποίηση μιας αγοράς πληροφορίας κάνοντας χρήση των γνώσεων που αποκομίσαμε προηγουμένως. Σε μια τέτοια αγορά μετέχουν τρεις τύποι οντοτήτων. Οι αγοραστές, οι πωλητές και κάποιοι ενδιάμεσοι. Μελετήσαμε το κομμάτι της αλληλεπίδρασης των αγοραστών με τις ενδιάμεσες οντότητες. Στο μοντέλο μας έχουμε μία οντότητα αγοραστή, την Buyer . Αποτελείται από δύο τμήματα. Το πρώτο έχει την ευθύνη κατασκευής πινάκων κάνοντας χρήση του αλγορίθμου της Q-μάθησης λαμβάνοντας υπόψη τις προτιμήσεις του χρήστη για να αποδώσει τις κατάλληλες επιβραβεύσεις στις διάφορες ενέργειες – μεταβάσεις του πράκτορα, ενώ το δεύτερο χρησιμοποιεί αυτόν τον πίνακα για να αποκτήσει τα αγαθά που επιθυμεί ο χρήστης, επιλέγοντας κάθε φορά την καλύτερη ενέργεια που επιδεικνύει ο Q-πίνακας.

Ήταν απαραίτητο στη συνέχεια να υποβάλλουμε το μοντέλο μας σε μια σειρά από προσομοιώσεις προκειμένου να καταγράψουμε τις επιδόσεις του στα διάφορα χαρακτηριστικά που μας ενδιέφεραν και να τις συγκρίνουμε με ένα μοντέλο πρότυπο το οποίο προέβαινε σε εξαντλητική αναζήτηση στους διάφορους ενδιαμέσους και επέλεγε το προϊόν που καλύτερα ανταποκρινόταν στις απαιτήσεις του χρήστη. Το μοντέλο μας επέφερε σημαντική μείωση στο χρόνο που απαιτούνταν για την απόκτηση του προϊόντος είτε χρησιμοποιούσαμε έναν σταθερό είτε μεταβλητό αριθμό επεισοδίων για την κατασκευή/ανανέωση των Q-πινάκων. Αξίζει επίσης να αναφέρουμε ότι με το μοντέλο μας (αφού έχει κατασκευασθεί ο Q-πίνακας) απαιτούνται κατά μέσο όρο μόλις 1,72 κινήσεις για την απόκτηση ενός προϊόντος, ενώ στην περίπτωση του μοντέλου πρότυπο οι κινήσεις αυτές είναι N+1, όπου N ο αριθμός των ενδιαμέσων.

Και επειδή το τέλος μιας εργασίας αποτελεί ερέθισμα για περαιτέρω μελέτη και έρευνα, σημειώσαμε στο τέλος μια σειρά προεκτάσεων οι οποίες θα προσέδιδαν κάποιες επιπλέον ενδιαφέρουσες δυνατότητες.

ΟΡΟΛΟΓΙΑ

Αγγλικός Όρος	Ελληνική Απόδοση
Accessible	προσβάσιμο
Accumulating	προσαυξημένο
Action	ενέργειας
Adaptivity	προσαρμοστικότητα
Agent	πράκτορας
Autonomy	αυτονομία
Belief	πεποίθηση
Benevolence	αγαθή προαίρεση
blackboard systems	συστήματα μαυροπίνακα
Bully	δυνάστης
Continuous	συνεχής
Contractor	εργολάβος
Convergence	Σύγκλιση
correlated equilibria	συσχετισμένες ισορροπίες
credit assignment	εκχώρηση πίστωσης
delayed reinforcement	καθυστερημένη ενίσχυση
Deliberative	συλλογιστικός
Desire	επιθυμία
Deterministic	αιτιοκρατικός
Directed	κατευθυνόμενος
Discrete	διακριτός
Distributed	κατανεμημένος
Dynamic	δυναμικός
Effectors	μηχανισμός δράσης
eligibility trace	ίχνος εκλεξιμότητας
Environment	περιβάλλον
Episodic	επεισοδιακός
Equilibrium	ισορροπία
expected discount return	προσδοκώμενο φθίνον αποτέλεσμα
expert systems	έμπειρα συστήματα
Exploit	αξιοποιώ
Exploration	εξερεύνηση
Follower	οπαδός
friend-Q	φίλος-Q
fully competitive game	πλήρως ανταγωνιστικό παίγνιο
Goal	στόχος
Godfather	νονός
Greedy	άπληστος
Infinite	ανοικτός
infinite horizon	ατελής ορίζοντας
Intention	πρόθεση
Leader	ηγέτης
learning classifier system	σύστημα εκμάθησης ταξινομητή
learning rate	ρυθμός μάθησης
Manager	διαχειριστής
Markov decision process	διαδικασία λήψης απόφασης Markov
Markov game	παίγνιο Markov
message passing systems	συστήματα ανταλλαγής μηνυμάτων
Mobility	κινητικότητα
Negotiation	διαπραγμάτευση
Object	αντικείμενο

Performance	εκτέλεση
policy iteration	επανάληψη πολιτικής
pro-activeness	προνοητικότητα
Q-learning	Q-μάθηση
Q-table	Q-πίνακας
Q-value	Q-αξία
Rational	λογικός
Rationality	λογικότητα
Reactiveness	αντιδραστικότητα
recency factor	παράγοντας συχνότητας
Regret	Μετάνοια
reinforcement learning	ενισχυτική μάθηση
Reward	ανταμοιβή
rule discovery	ανακάλυψη κανόνα
self-play	αυτό-παίξιμο
Sensors	αισθητήρας
social ability	κοινωνικότητα
State	κατάσταση
Static	στατικός
strategic game	στρατηγικό παίγνιο
supervised learning	μάθηση με επίβλεψη
Team-Q	ομάδα-Q
temporal difference methods	μέθοδοι χρονικής διαφοράς
trial and error	προσπάθεια–και–λάθος
Undirected	μη κατευθυνόμενος
Value iteration	επανάληψη αξίας
Veracity	ειλικρίνεια

ΑΝΑΦΟΡΕΣ

1. Aron R., Sundararajan A., & Viswanathan S., Intelligent Agents in Electronic Markets for Information Goods: Customization, Preference and Pricing, *Decision Support Systems, vol. 41(4)*, 2006, p. 764-786.
2. Bakos, Y., The Emerging Role of Electronic Marketplaces on the Internet, *Communications of the ACM,vol. 41(8),* August 1998, p. 35-42.
3. Basar, T., An equilibrium theory for multiperson decision making with multiple probabilistic models. *IEEE Transactions on Automatic Control*, 30(2), 1985, p. 118-132.
4. Baird, L., Residual algorithms: Reinforcement learning with function approximation. *In Proceedings Twelfth International Conference on Machine Learning (ICML-95)*, Tahoe City, USA, 1995, p. 30-37.
5. Baird, L. and Moore, A., Gradient descent for general reinforcement learning. *In Advances in Neural Information Processing Systems 11 (NIPS-98)*, Denver, USA, 1998, p. 968-974.
6. Banerjee, B. and Peng, J., Adaptive policy gradient in multiagent learning. *In Proceedings 2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-03)*, Melbourne, Australia, 2003, p. 686-692.
7. Barbosa, G. P. & Silva, Q. B, An Electronic Marketplace Architecture Based on Technology of Intelligent Agents and Knowledge. Lecture Notes in Computer Science, “E-Commerce Agents, Marketplace Solutions, Security Issues, and Supply and Demand”, 2001, p. 39-60.
8. Bonarini, A. and Trianni, V., Learning fuzzy classifier systems for multi-agent coordination. *Information Sciences*, 136, 2001, p. 215-239.
9. Booker, L. B., Classifier systems that learn internal world models. *Machine Learning*, 3, 1988, p. 161-192.
10. Boutilier, C., Planning, learning and coordination in multiagent decision processes. *In Proceedings Sixth Conference on Theoretical Aspects of Rationality and Knowledge (TARK-96)*, De Zeeuwse Stromen, The Netherlands, 1996, p. 195-210.

11. Bowling, M. 2004, Convergence and no-regret in multiagent learning. In *Advances in Neural Information Processing Systems 17 (NIPS-04)*, Vancouver, Canada, 1996, p. 209-216.

12. Bowling, M. and Veloso, M., Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2), 2002, p. 215-250.

13. Boyan, J. A. and Littman, M. L., Packet routing in dynamically changing networks: A reinforcement learning approach. In *Advances in Neural Information Processing Systems 6 (Neural Information Processing Systems 1993, NIPS-93*), Denver, Colorado, USA , 1993, p. 671-678.

14. Brooks, R. A., A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation 2* (1), 1986, p. 14-23.

15. Busoniu, L., De Schutter, B., and Babuska, R., Multiagent reinforcement learning with adaptive state focus. In Proceedings *17th Belgian-Dutch Conference on Artificial Intelligence (BNAIC-05)*, Brussels, Belgium, 2005, p. 35-42.

16. Busoniu, L., De Schutter, B., and Babuska, R., Learning and Coordination in Dynamic Multiagent Systems. Technical Report 05-019, 2005.

17. Chalkiadakis, G. and Boutilier, C., Coordination in multiagent reinforcement learning: A Bayesian approach. *In Proceedings 2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-03)*, Melbourne, Australia, 2003, p. 709-716.

18. Clouse, J., Learning from an automated training agent. In *Working Notes Workshop on Agents that Learn from Other Agents, Twelfth International Conference on Machine Learning (ICML-95)*, Tahoe City, USA, 1995.

19. Chavez A.and Maes P., Kasbah: An Agent Marketplace for Buying and Selling Goods, In *Proceedings the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM '96)*, London, 22-24 April 1996, p. 75-90.

20. Collins J., Jamison S., Mobasher B. & Gini M., "A Market Architecture for Multi-Agent Contracting," *In Proceedings of the 2nd International Conference on Autonomous Agents, New York,* 1998, p. 285-292.

21. Conitzer, V. and Sandholm, T., AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents.

In *Proceedings Twentieth International Conference on Machine Learning (ICML-03)*, Washington, USA, 2003, p. 83-90.

22. Crites, R. H. and Barto, A. G., Elevator group control using multiple reinforcement learning agents. *Machine Learning*, 33(23), 1998, p. 235-262.

23. Davies N. and Weeks R., Jasper: Communicating Information Agents, In *Proceedings of the 4th International Conference on the World Wide Web*, Boston, USA, December 1995.

24. De Jong, E., An accumulative exploration method for reinforcement learning. In *Notes Workshop on Multiagent Learning, 14th National Conference on Artificial Intelligence (AAAI-97)*, Providence, Rhode Island, 1997.

25. Decker K, Sycara K and Williamson M, Middle-Agents for the Internet. *In Proceedings of the 15th International Joint Conference on Artificial Intelligence,* 1997.

26. Dorigo, M. and Bersini, H., A comparison of Q-learning and classifier systems. In *From Animals to Animats 3. Proceedings 3rd International Conference on Simulation of Adaptive Behavior (SAB-94)*, Brighton, United Kingdom, 1994, p. 248-255.

27. Etzioni O. and Weld D., A Softbot-Based Interface to the Internet, Communications of the ACM 37(7), p. 72-76, 1994.

28. Ferguson I. A., *Touring Machines: An Architecture for Dynamic, Rational, Mobile Agents*, PhD Thesis, Computer Laboratory, University of Cambridge,UK, 1992.

29. Fulda, N. and Ventura, D., Dynamic joint action perception for Q-learning agents. In *Proceedings 2003 International Conference on Machine Learning and Applications (ICMLA- 03)*, Los Angeles, USA, 2003, p. 73-79.

30. Franklin S. and Graesser A., Is it an agent or is it just a program? A Taxonomy for Autonomous Agents, *Proc ECAI '96 Workshop on Agent Theories, Architectures and Languages, Intelligent Agents III*, LNAI, Vol. 1193, Springer, August 12-13 1997, p. 21-36.

31. Graat G., "*Agent-Based Information Retrieval Supportedby Information Markets*," Msc Thesis, Dep. Of Computer Science, University of Maastricht, 2003.

32. Greenwald, A. and Hall, K., Correlated-Q learning. In *Proceedings Twentieth International Conference on Machine Learning (ICML-03)*, Washington, USA, 2003, p. 242-249.

33. Guestrin, C., Lagoudakis, M. G., and Parr, R., Coordinated reinforcement learning. In *Proceedings Nineteenth International Conference on Machine Learning (ICML-02)*, Sydney, Australia, 2002, p. 227-234.

34. Haynes, T. and Sen, S., Evolving behavioral strategies in predators and prey. In Weiss, G. and Sen, S., editors, Adaptation and Learning in Multi-Agent Systems, Springer Verlag, 1996, p. 113-126.

35. Hu, J. and Wellman, M. P., Learning about other agents in a dynamic multiagent system. *Journal of Cognitive Systems Research*, 1, 2001, p. 67-79.

36. Hu, J. and Wellman, M. P., Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4, 2003, p. 1039-1069.

37. Kaelbling, L. P., Littman, M. L., and Moore, A. W., Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 1996, p. 237-285.

38. Klusch M., Information Agent Technology for the Internet: a Survey, *Data & Knowledge Engineering*, vol. 36(3), 2001, p. 337-372.

39. Kok, J. R., Spaan, M. T. J., and Vlassis, N., Non-communicative multi-robot coordination in dynamic environment. *Robotics and Autonomous Systems*, 50(2-3), 2005a, p. 99-114.

40. Kok, J. R., 't Hoen, P. J., Bakker, B., and Vlassis, N., Utile coordination: Learning interdependencies among cooperative agents. In *Proceedings IEEE Symposium on Computational Intelligence and Games (CIG-05)*, Colchester, United Kingdom, 2005b, p. 29-36.

41. Kok, J. R. and Vlassis, N., Sparse cooperative Q-learning. In *Proceedings Twenty-first International Conference on Machine Learning (ICML-04)*, Banff, Canada, 2004, p. 481-488.

42. Kononen, V., Asymmetric multiagent reinforcement learning. In *Proceedings IEEE/WIC International Conference on Intelligent Agent Technology (IAT-03)*, Halifax, Canada, 2003, p. 336-342.

43. Kozierok R. and Maes P., A Learning Interface Agent for Scheduling Meetings, *Proceedings of the ACM-SIGCHI International Workshop on Intelligent User Interfaces*, Florida, 1993, p. 81-93.

44. Lauer, M. and Riedmiller, M., An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings Seventeenth International*

Conference on Machine Learning (ICML-00), Stanford University, USA, 2000, p. 535-542.

45. Lieberman, H., Letizia: An Agent that Assists Web Browsing, In *Proceedings of IJCAI 95*, AAAI Press, 1995.

46. Littman, M. L., Friend-or-foe Q-learning in general-sum games. In *Proceedings Eighteenth International Conference on Machine Learning (ICML-01)*, Williams College, Williamstown, USA, 2001a, p. 322-328.

47. Littman, M. L., Value-function reinforcement learning in Markov games. *Journal of Cognitive Systems Research*, 2, 2001b, p. 55-66.

48. Littman, M. L. and Stone, P., Implicit negotiation in repeated games. In *Proceedings 8th International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001)*, Seattle, USA, 2001, p. 96-105.

49. Mataric, M. J., Learning in multi-robot systems. In Weiss, G. and Sen, S., editors, *Adaptation and Learning in Multi-Agent Systems*, Springer Verlag, 1996, p. 152-163.

50. Mataric, M. J., Learning social behavior. *Robotics and Autonomous Systems*, 20, 1997, p. 191-204.

51. Moore, A. W. and Atkeson, C. G., Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13, 1993, p. 103-130.

52. Muller J. P., A Conceptual Model for Agent Interaction, In Deen S. M. (ed.) *Proceedings of the 2nd International Working Conference on Cooperative Knowledge Based Systems (CKBS-94)*, Keele University: Dake Centre, 1994, p. 213-234.

53. Nwana H.S., “Software Agents: An Overview,” *KnowledgeEngineering Review*, *vol. 11(3)*, Sept. 1996, p. 1-40.

54. Nwana H.S. and Wooldridge M., Software Agent Technologies, *BT Technology Journal, 14(4)*, 1996.

55. Nwana H., Lee L., Jennings N., Coordination in Software Agent Technologies, *BT Technology Journal*, 14(4), 1996.

56. Oliveira, E., Fonseca, J. M. & Jennings, N. R. Learning to be Competitive in the Market. In *Proc. of the AAAI Workshop on Negotiation: Settling Conflicts and Identifying Opportunities*, Orlando, Florida, USA, 1999.

57. Peng, J. and Williams, R. J., Incremental multi-step Q-learning. *Machine Learning*, 22(1-3), 1996, p. 283-290.

58. Plaza, E. and Ontanon, S., Cooperative multiagent learning. In Alonso, E., Kudenko, D., and Kazakov, D., editors, Adaptive Agents and Multi-Agent Systems, Springer, 2003, p. 1-17.

59. Powers, R. and Shoham, Y., New criteria and a new algorithm for learning in multiagent systems. In *Advances in Neural Information Processing Systems 17 (NIPS-04),* Vancouver, Canada, 2004, p. 1089-1096.

60. Price, B. and Boutilier, C., Accelerating reinforcement learning through implicit imitation. *Journal of Artificial Intelligence Research*, 19, 2003, p.569-629.

61. Pynadath, D. V. and Tambe, M., The communicative multiagent team decision problem: Analyzing teamwork theories and models. *Journal of Artificial Intelligence Research*, 16, 2002, p.389-423.

62. Russell, S. and Norvig, P., Artificial Intelligence: A Modern Approach. Prentice Hall, Englewood Cliffs, USA, 1995.

63. Russell, S. and Norvig, P., Artificial Intelligence: A Modern Approach. Prentice Hall, 2nd edition, Englewood Cliffs, USA, 2003.

64. Rhodes B. J. and Starner T., Remembrance Agent: A Continuously Automated Information Retrieval System, In *Proceedings the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM 96)*, London, 22-24 April, 1996, p. 487-496.

65. Schaerf, A., Shoham, Y., and Tennenholtz, M., Adaptive load balancing: A study in multi-agent learning. *Journal of Artificial Intelligence Research*, 2, 1995, p. 475-500.

66. Sen, S., Sekaran, M., and Hale, J., Learning to coordinate without sharing information. In *Proceedings 12th National Conference on Artificial Intelligence (AAAI-94)*, Seattle, USA, 1994, p. 426-431.

67. Sen, S. and Weiss, G., Learning in multiagent systems. In Weiss, G., editor, Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, chapter 6, MIT Press, 1999, p. 259-298.

68. Shardanand U. and Maes P., Social Information Filtering for Automating Word of Mouth, *In Proceedings of CHI-95*, Denver, CO., May, 1995.

69. Sheth B. and Maes P., Evolving Agents for Personalised Information Filtering, In *Proceedings of the 9th IEEE Conference on Artificial Intelligence for Applictions,* 1993.

70. Shoham, Y., Powers, R., and Grenager, T., Multi-agent reinforcement learning: A critical survey. Technical report, Computer Science Dept., Stanford University, California, USA, 2003.

71. Singh, M. P. and Huhns, M. N., Challenges for machine learning in cooperative information systems. In Weiss, G., editor, Distributed Artificial Intelligence Meets Machine Learning, Learning in Multi-Agent Environments, Springer, p. 11-24, Selected papers from the ECAI'96 Workshop LDAIS, Budapest, Hungary, and the ICMAS'96 Workshop LIOME, Kyoto, Japan, 1996.

72. Singh, S., Jaakkola, T., Littman, M. L., and Szepesvari, C., Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3), 2000a, p. 287-308.

73. Singh, S., Kearns, M., and Mansour, Y., Nash convergence of gradient dynamics in general-sum games. In *Proceedings 16th Conference on Uncertainty in Artificial Intelligence (UAI-00)*, San Francisco, USA, 2000b, p. 541-548.

74. Spaan, M. T. J., Vlassis, N., and Groen, F. C. A., High level coordination of agents based on multiagent Markov decision processes with roles. *In Workshop on Cooperative Robotics*, 2002 IEEE/RSJ *International Conference on Intelligent Robots and Systems (IROS-02)*, p. 66-73, Lausanne, Switzerland, 2002, p. 66-73.

75. Stone, P. and Veloso, M., A layered approach to learning client behaviors in the RoboCup soccer server. *Applied Artificial Intelligence*, 12, 1998, p. 165-188.

76. Stone, P. and Veloso, M., Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork. *Artificial Intelligence*, 110(2), 1999, p. 241-273.

77. Stone, P. and Veloso, M., Multiagent systems: A survey from the machine learning perspective. *Autonomous Robots*, 8(3), 2000, p.345-383.

78. Suematsu, N. and Hayashi, A., A multiagent reinforcement learning algorithm using extended optimal response. In *Proceedings 1st International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-02)*, Bologna, Italy, 2002, p. 370-377.

79. Sutton, R. S., Integrated modeling and control based on reinforcement learning and dynamic programming. In *Advances in Neural Information Processing Systems 3 (The 1991 Neural Information Processing Systems Conference, NIPS-91)*, Denver, Colorado, USA, 1991, p. 471-478.

80. Sutton, R. S. and Barto, A. G., Reinforcement Learning: An Introduction. MIT Press, Cambridge, USA, 1998.

81. Tambe, M., Towards flexible teamwork. Journal *of Artificial Intelligence Research*, 7, 1997, p. 83-124.

82. Tan, M., Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings Tenth International Conference on Machine Learning (ICML-93)*, Amherst, USA, 1993, p. 330-337.

83. Tesauro, G., Pricing in Agent Economies Using Neural Networks and Multi-Agent Q-Learning. *In Proc. of the Workshop on Learning About, From and With other Agents (IJCAI '99),* 1999.

84. Thrun, S., The role of exploration in learning control. In White, D. and Sofge, D., editors, Handbook for Intelligent Control: Neural, Fuzzy and Adaptive Approaches. Van Nostrand Reinhold, 1992.

85. Touzet, C. F., Robot awareness in cooperative mobile robot learning. *Autonomous Robots*, 8(1), 2000, p. 87-97.

86. Tsvetovatyy, M., Gini, M., Mobasher, M. & Wieckowski, Z. MAGMA: An Agent-Based Virtual Market for Electronic Commerce. *Applied Artificial Intelligence*, vol. 11(6), 1997, p. 501-523.

87. URL1: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-5/www/pleiades.html

88. Vlahavas I., Kefalas P., Bassiliades N., Refanidis I., Kokkoras F., Sakellariou H., Artificial Intelligence. Gartaganis Publications, Thessaloniki, Greece, EU, 2002..

89. Vlassis, N., A concise introduction to multiagent systems and distributed AI. Technical report, University of Amsterdam, The Netherlands. URL: http://www.science.uva.nl/ vlassis/cimasdai/cimasdai.pdf, 2003.

90. Walker, A. and Wooldridge, M., Understanding the emergence of conventions in multi-agent systems. *In Proceedings 1st International Conference on Multi-Agent Systems (ICMAS-95)*, San Francisco, USA, 1995, p. 384-390.

91. Watkins, C. J. C. H. and Dayan, P., Technical note: Q-learning. *Machine Learning*, 8, 1992, p.279-292.

92. Whitehead, S. D. and Lin, L.-J. Reinforcement learning of non-Markov decision processes. *Artificial Intelligence*, 73(1-2), 1995, p.271-306.

93. Wooldridge M. and Jennings N., Intelligent Agents: Theory and Practice, *Knowledge Engineering Review*, 10(2), 1995.

94. Wooldridge M. and Jennings N., Software Engineering with Agents: Pitfalls and Pratfalls, *IEEE Internet Computing*, May-June 1999.

95. Wooldridge M., On agent-based software engineering, *Artificial Intelligence*, vol .117, 2000, p. 277-296.

96. Yarom, I., Rosenschein, J. S. & Gldman, C. V. The Role of Middle-Agents in Electronic Commerce. *IEEE Intelligent Systems*, vol. 18, no. 6, 2003, p. 15-21.