ΠΛΗ512 - Πολυπρακτορικά Συστήματα

Algorithm Selection in Bilateral Negotiation

Περίληψη Ερευνητικής Εργασίας

Ιωάννης Χριστοφιλογιάννης (A.M.: 2019030140)
Πέτρος Μπιμπίρης (A.M.: 2019030135)

Χειμερινό Εξάμηνο 2023

Introduction

Οι συγγραφείς αντιμετωπίζουν το πρόβλημα επιλογής αλγορίθμου (στρατηγικής) για διμερή διαπραγμάτευση μεταξύ πρακτόρων με τη δημιουργία meta-αλγορίθμων που εξετάζουν ποσοτικά την αποτελεσματικότητα των διάφορων εναλλακτικών.

Το πρόβλημα χωρίζεται σε on-line και off-line variants, με κριτήριο το αν οι συμμετέχοντες σε μια διαπραγμάτευση μπορούν να αλλάξουν στρατηγική κατά την διάρκεια της. Το off-line variant προσεγγίζεται με supervised machine learning το οποίο αντιμετωπίζει διαφορετικούς αλγορίθμους ως "black boxes" και προσπαθεί να προβλέψει την απόδοσή τους με βάση τα χαρακτηριστικά του domain. Το on-line variant μοντελοποιείται ως multi-armed bandits (MAB) setting με τις διαφορετικές στρατηγικές ως arms, και εφαρμόζεται ο αλγόριθμος Upper Confidence Bound (UCB) για να επιλεγούν στρατηγικές.

Η αξιολόγηση των μετα-αλγορίθμων γίνεται με βάση το GENIUS testbed διαπραγματεύσεων του διαγωνισμού ANAC. Η βασική ιδέα του paper είναι ότι "a little learning goes a long way" – "λίγη μάθηση" βοηθάει σε αυτό και πιθανώς σε άλλα προβλήματα απόφασης στον τομέα των πολυπρακτορικών συστημάτων.

Problem Definition

Ορίζεται μαθηματικά το πρόβλημα της επιλογής αλγορίθμου, το μοντέλο του περιβάλλοντος διαπραγματεύσεων και η διαδικασία που ακολουθείται στον ANAC. Υποδεικνύεται και αιτιολογείται ο τρόπος με τον οποίο κωδικοποιούνται τα χαρακτηριστικά των domains τα οποία χρησιμοποιούν τα μοντέλα προκειμένου να προβλέψουν την απόδοση των διαφορετικών αλγορίθμων.

Off-Line Problem Variant: Supervised Learning

Στην off-line παραλλαγή του προβλήματος η στρατηγική/πράκτορας επιλέγεται στην αρχή της διαπραγμάτευσης και δεν αλλάζει στην συνέχεια. Δημιουργείται ο "meta-agent", ο οποίος επιλέγει για κάθε νεο domain τον καλύτερο αλγόριθμο, με βάση ορισμένα στατιστικά χαρακτηριστικά του domain. Η διαδικασία επιλογής έχει ως εξής:

1. Χωρίζεται το dataset (τα domains και οι συμμετέχοντες πράκτορες στον ANAC2012) σε training και test κομμάτια.

2. Εκπαιδεύεται ο meta-agent πάνω στο training set χρησιμοποιώντας μεθόδους supervised learning.

3. Ο meta-agent έχει πλέον την δυνατότητα να "προβλέψει" την απόδοση ενός πράκτορα σε ένα domain, ακόμα και αν δεν έχει "ξαναδεί" τίποτα από τα 2.

4. Για κάθε νέο domain που αντιμετωπίζει, ο meta-agent προβλέπει την απόδοση καθενός από τους διαθέσιμους πράκτορες και επιλέγει αυτόν με την μεγαλύτερη προβλεπόμενη απόδοση.

Η πρόβλεψη απόδοσης γίνεται με μια από δύο διαφορετικές "μετρικές": είτε το (normalized) score που θα πετύχει ο πράκτορας, είτε το αν το score αυτό θα είναι αρκετά κοντά στο βέλτιστο δυνατό. Και για τις δύο μετρικές χρησιμοποιούνται συμβατικές μέθοδοι μάθησης, με μικρές αλλαγές μεταξύ των μετρικών. Κάθε agent χρησιμοποιεί μια μόνο τεχνική. Οι τεχνικές που χρισιμοποιούνται είναι: γραμμική παλινδρόμηση (Linear Regression), λογιστική παλινδρόμηση (Logistic Regression), Classification And Regression Trees (CART) και ένα νευρωνικό δίκτυο με 1 hidden layer με 4 νευρώνες.

Όλοι οι meta-agents πετυχαίνουν στατιστικά σημαντικές αυξήσεις στην απόδοση σε σχέση με τον μέσο πράκτορα που συμμετείχε στον ANAC2012, και επιλέγουν τον βέλτιστο πράκτορα για το αντίστοιχο domain περίπου στο 35% των φορών. Στον ANAC2012 παρ'ολ'αυτά ο meta-agent που συμμετείχε τερμάτισε 6ος. Δεν παρατηρούνται σημαντικές διαφορές στην απόδοση μεταξύ των meta agents.

On-Line Problem Variant: Multi-Armed Bandit Approach

Το πρόβλημα προσεγγίζεται ως Multi-Armed Bandit (MAB) setting όπου κάθε αλγόριθμος είναι ένα arm. Μοντελοποιείται ο MAB-agent, ο οποίος αντιμετωπίζει κάθε αλγόριθμο σαν μια άγνωστη πιθανοτική κατανομή που προσφέρει "rewards" και προσπαθεί να μαζέψει όσο περισσότερα rewards γίνεται, επιλέγοντας ένα arm σε κάθε negotiation round. Τέτοιου είδους προβλήματα έχουν μελετηθεί αρκετά και ως αποτέλεσμα υπάρχουν αλγόριθμοι που προσφέρουν (θεωρητικές) εγγυήσεις απόδοσης και βελτιστότητας, ένας από τους οποίους είναι ο Upper Confidence Bound (UCB), ο οποίος επιλέγεται σε αυτή την εργασία.

Δημιουργούνται 2 MAB-agents, οι οποίοι διαφέρουν ως προς την πληροφορία που έχουν για τους διαθέσιμους αλγορίθμους πριν ξεκινήσουν να επιλέγουν. Ο ένας ονομάζεται MAB-pure και δεν διαθέτει καμία πληροφορία για τους διαθέσιμους αλγορίθμους. Ο δεύτερος ονομάζεται MAB-prior και χρησιμοποιεί την τεχνική που περιγράφηκε στο off-line variant (με ένα Classification And Regression Tree ως μέθοδο μάθησης) προκειμένου να αποκτήσει κάποια αρχική πληροφορία για τους αλγορίθμους μεταξύ των οποίων καλείται να επιλέξει. Σημειώνεται ότι ο MAB-pure δεν χρειάζεται training data ή ξεχωριστή περίοδο μάθησης, καθώς σχηματίζει το μοντέλο του για κάθε arm κατά την διάρκεια της διαδικασίας επιλογής (aka on-line), από τα score που πετυχαίνει με κάθε επιλογή.

Ο MAB-pure πετυχαίνει παρόμοια αποτελέσματα με τον meta-agent μετά από περίπου το 60% των test rounds, κατά την διάρκεια των οποίων δίνει περισσότερο βάρος στο να "εξερευνήσει" τις διάφορες εναλλακτικές παρά στο να εκμεταλλευτεί αυτές που θεωρεί καλύτερες. Ο MAB-prior αποδίδει σημαντικά καλύτερα από τον meta-agent, βελτιώνοντας συνεχώς την

απόδοση του όσο περνάει ο χρόνος, υπογραμμίζοντας την δυνατότητα του να προσαρμόζεται στο περιβάλλον του.

Ο prior-MAB συμμετείχε στον ANAC2013, στον οποίο είχε το ίδιο score με τον νικητή του διαγωνισμού αλλά κατέλαβε την δεύτερη θέση λόγω μεγαλύτερου variance στην απόδοση του. Το αυξημένο αυτό variance αντικατοπτρίζει το ότι ο αλγόριθμος UCB κάνει κάποιες επιλογές με σκοπό όχι να αυξήσει την απόδοση του αλλά να εξερευνήσει καλύτερα το περιβάλλον.