לימוד שיתוף פעולה
ויישום להכרזות בבריג'

אסף עמית

לימוד שיתוף פעולה
בהכרזות בברידג'

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים במדעי המחשב

אסף עמית

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

סיון תשס"ד חיפה יוני 2004

המחקר נעשה בהנחייתו של דר' שאול מרקוביץ בפקולטה למדעי המחשב.

ברצוני להביע את תודתי לקרן ע"ש מרקו ולואיז מיטרני תרומת מר וגברת בלמונטה על תמיכתם הנדיבה, ולטכניון על העזרה כספית במהלך המחקר.

תוכן העניינים:

סימונים וקיצורים 7
1. **מבוא** 8
2. **פתרון בעיות בעזרת מודל של שותפים** 10
2.1. סימולציה בסביבה מרובת סוכנים 11
2.2. סימולציה בסביבה נצפית בחלקיות 13
2.3. דגימות מבוססות מודל 13
2.4. דגימות מבוססות מודל מוגבלות משאבים 17
3. **ייצוג אסטרטגיות בחירה בעזרת רשתות החלטה** 20
4. **לימוד שיתוף פעולה** 23
4.1. אלגוריתם למידה 23
4.1.1. יצור דוגמאות אינפורמטיביות ללמידה 23
4.1.2. תיוג הדוגמאות 24
4.1.3. לימוד בעזרת רשתות החלטה 25
4.2. לימוד עם שותפים 26
5. **אפליקציה: ללמוד להכריז בברידג'** 29
5.1. ההכרזה בברידג' 29
5.1.1. בעיות בהכרזות ברידג' 30
5.1.2. תוכניות הכרזה בברידג' 31
5.2. בידי: תוכנית הכרזה בברידג' 32
5.2.1. מרחב המצבים 32
5.2.2. רשת ההחלטה 33
5.2.3. פונקציית ההערכה 33
5.2.4. ייצוג המודלים של הסוכנים 34
5.2.5. יצירת רשתות החלטה בזמן אמת 34
5.2.6. למידה 34
6. **תוצאות אמפיריות** 36
6.1. מערך הניסויים 36
6.1.1. משתנים תלויים 37
6.1.2. משתנים בלתי תלויים 38
6.2. תוצאות הניסויים 38
6.2.1. עקומת הלמידה 39
6.2.2. השימוש ברשתות שנלמדו בעת המבחנים 40
6.2.3. השפעת הלמידה על תהליך הדגימות 41
6.2.4. השפעת הלמידה על הסימולציה 42

תוכן העניינים: (המשך)

6.3. **ניתוח פרמטרי של הלמידה** 43
6.3.1. השפעת גודל הדגימה על התוצאות 44
6.3.2. השפעת הגבלת המשאבים על התוצאות 45
6.3.3. ההשפעה של החלפות אסטרטגיות בין הסוכנים 45
6.3.4. השפעת השימוש בפונקציות הערכה שונות 46
7. **עבודות נוספות** 47
8. **סיכום** 49

נספחים:

א. חוקי משחק הברידג' 51
א.1. החלוקה 51
א.2. ההכרזה 51
א.3. המשחק 53
א.4. שיטת הניקוד 53
א.4.1. ניקוד עבור ביצוע החוזה 54
א.4.2. ניקוד עבור כישלון בחוזה 55
א.4.3. נקודות נצחון בינלאומיות (International Match Points) .. 55
ב. תכונות המשמשות לתיאור היד בבידי 57
ב.1. תכונות להערכת קלפים גבוהים 58
ב.2. תכונות להערכת החלוקה 59
ב.3. תכונות להערכת איכות הסדרות 60
ב.4. תכונות להערכת איכות היד 60
ב.5. תכונות דינמיות 61
ג. בניית רשתות החלטה בזמן אמת 62
ג.1. מבנה חוקי ההכרזה 62
ג.2. המרת חוקי הכרזה לרשתות החלטה 63
מראה מקום 69

רשימת איורים:

1. אלגוריתם בסיסי לביצוע החלטות בסביבה נצפית במלואה עבור סוכן מוגבל במשאבים 10
2. עץ חיפוש המשתמש בסימולציה לחיזוי פעולות סוכנים נוספים בסביבה. 12
3. אלגוריתם סימולציה מוגבל משאבים לביצוע החלטות בסביבה מרובת סוכנים נצפית במלואה. (אלגוריתם ה-FIDM) 14
4. שילוב דגימות בחיפוש בסביבה נצפית חלקית. 15
5. אלגוריתם סימולציה מוגבל משאבים לביצוע החלטות בסביבה מרובת סוכנים נצפית חלקית. (אלגוריתם ה-PIDM) 16
6. דגימות מבוססות מודל. 18
7. דגימות מבוססות מודל מוגבלות במשאבים. (אלגוריתם ה-RBMBMC) 19
8. דוגמא לרשת החלטה. 21
9. חיפוש ברשת החלטה. 22
10. זרימת המידע בעת תהליך הלמידה. 23
11. צורות עדכון רשת ההחלטה בעת תהליך הלמידה. 25
12. החלפת רשתות החלטה בין השותפים בעת תהליך הלמידה. 28
13. עקומת הלמידה עבור 100 בעיות אקראיות. 39
14. שימוש ברשתות ההחלטה שנלמדו בעת המבחן. 41
15. השפעת הלמידה על תהליך הדגימות. 42
16. אורך ההכרזה הממוצע בעת הפעלת הסימולציות. 42
17. תוצאות הניסוי בניתוח הפרמטרי. 43
18. השפעת גודל הדגימה על התוצאות. 44
19. השפעת הגבלת המשאבים על התוצאות. 45
20. השפעת הלמידה בעת לימוד סוכנים שונים. 46
21. דוגמא לחוק הכרזה. 63
22. דוגמא לחוק הכרזה שלילי. 64
23. השלבים בבנית רשתות בזמן אמת. 64
24. צורות הזרמת צמתים בעת בניית רשתות החלטה. 67
25. האלגוריתם לבניית רשתות בזמן אמת. 68

תקציר

מזה זמן רב משתמשים חוקרים של בינה מלאכותית במשחקים כדי ללמוד את תהליך קבלת החלטות בסביבות תחרותיות מרובות-סוכנים. מרבית המחקרים עוסקים במשחקים של שני שחקנים, כגון שחמט או דמקה, שבהם יש מידע מלא על הסביבה (*Fully observable environment*). רק מקצתם עוסקים במשחקים מורכבים יותר. ניקח לדוגמא את משחק הברידג' אשר לו מאפיינים, ההופכים אותו ליותר מאתגר בתהליך קבלת ההחלטות:

1. המשחק כולל קבוצות של סוכנים משתפי-פעולה המתחרות זו נגד זו.
2. לכל אחד מהסוכנים יש רק מידע חלקי על הסביבה.
3. התקשורת בין הסוכנים היא מוגבלת.
4. מרחב החיפוש הוא אדיר, בעוד שהמשאבים המוקצים להחלטות מוגבלים.

עבודה זו מציגה מערכת מבוססת-מודל ללמידה וקבלת החלטות בסביבות מרובות סוכנים. הסוכנים מתחלקים לשתי קבוצות, כאשר הם פועלים בשיתוף פעולה בתוך הקבוצה (*cooperative agents*) ומתחרים יחד נגד הקבוצה השניה בתוך סביבה נצפית-חלקית (*partial observable*).

תרומות העבודה

לעבודה זו יש שלוש תרומות עיקריות:

1. אנו מציגים אלגוריתם (PIDM) לקבלת החלטות מבוססות-מודל בסביבות מרובות סוכנים ובתנאי אי-ודאות.
2. אנו מציגים מערכת לומדת, לאימון משותף של הסוכנים משתפי-הפעולה. התהליך מאפשר תיאום טוב יותר עם יתר הסוכנים בקבוצה וכך משפר את איכות ההחלטות במגבלת המשאבים הקיימת.
3. את הטכניקות המוצגות כאן אנו מיישמים בתחום של משחק הברידג' ויוצרים את אלגוריתם ההכרזה הראשון היודע להתאים את עצמו באופן אוטומטי לשותף, ומשתפר תוך כדי אימון עצמי.

סקירה כללית

בתחילת העבודה אנו מציגים את אלגוריתם ה-PIDM (*Partial Information Decision Making*). זהו אלגוריתם חוזה, המקבל מודלים של הסוכנים השונים ומחזיר את הפעולה הטובה ביותר, שהוא מוצא במשאבים המוגבלים שהוקצו לשם כך. האלגוריתם משתמש בסימולציה של כל אחת מהפעולות האפשריות על מנת לבחור את הפעולה שתניב את התועלת הגבוהה ביותר.

שתי הבעיות המרכזיות המפריעות לנו לביצוע הסימולציה הם אי-הוודאות לגבי הפעולות של כל אחד מהסוכנים, ואי-הוודאות על המצב המלא של הסביבה.

בבעיה הראשונה אנו מטפלים על ידי שימוש במודל של הסוכנים האחרים בסביבה. לכל סוכן יש פונקצית הערכה למצב, התייחסות לשאר הסוכנים ואסטרטגיית של אופן בחירת הפעולה שלו. היישום של מודל זה לעץ החיפוש מאפשר לנו לצמצם את מקדם החיפוש בכל אחד מהצמתים בעץ, כך שאנו מפתחים אותו רק עבור הפעולות, התואמות את אסטרטגיית הבחירה של הסוכן.

אי-הוודאות על מצב הסביבה מטופלת על ידי מנגנון דגימות מבוססות-מודל (אלגוריתם ה-RBMBMC - *Bounded Model Based Monte-Carlo Sampling*). אילו יכולנו לבחון את כל האפשרויות של מצבי הסביבה, היינו יכולים למצוא את הפעולה בעלת התוחלת הרווחית ביותר. זה לא מתאפשר בגלל המשאבים המוגבלים שלנו ולכן אנו מייצרים דגימות של מצבים אפשריים של הסביבה. בחירת הפעולה מתבססת על תוחלת הרווח של כל הפעולות האפשריות בהתבסס על אותן דגימות.

באלגוריתם זה אנו מייצרים באופן אקראי דגימות של מצבים אפשריים של הסביבה, על פי המידע החשוף לסוכן. כל אחת מדגימות אלה נבדקות באמצעות המודלים של הסוכנים האחרים מול הפעולות שננקטו על ידי הסוכנים ואסטרטגיות הבחירה שלהם. כיון שהמשאבים מוגבלים, אנו לוקחים רק דגימות שתואמות במידה הרבה ביותר את אסטרטגיות הבחירה של הסוכנים השונים בסביבה.

מבנה הנתונים המשמש אותנו כדי לשמור את אסטרטגיית הבחירה של כל סוכן היא רשת החלטה (Decision net). רשת ההחלטה היא גרף מכוון, שכל אחד מהצמתים בה ממפה קבוצת מצבים אפשריים של הסביבה לפעולות אפשריות של הסוכן, המותאמות לאותם מצבים על פי אסטרטגיית הפעולה של הסוכן. הרשת בנויה בצורה היררכית, כאשר הקבוצה המשויכת לכל צומת היא תת-קבוצה של המצבים המשויכים לצמתים המובילים אליה.

כדי למצוא את הפעולות האפשריות של הסוכן, אנו יורדים בצמתים ברשת בהתאם לסט המצבים המתאים לכל צומת. קבוצת הפעולות בצמתים הספציפיים ביותר שהגענו אליהם משמשת כקבוצת הפעולות הנבדקת. באופן דומה אנו מבצעים את תהליך הדגימות, כאשר לכל דגימה אנו בודקים אם הפעולה שביצע הסוכן תואמת לאחת הפעולות שהוצעו בתהליך החיפוש ברשת ההחלטה.

על ההכרזות במשחק הברידג'

בעבודה זו אנו מיישמים את אלגוריתם ה-PIDM ותהליך הלמידה לשלב ההכרזות (*Bidding*) במשחק הברידג'. בשלב זה של המשחק ניתן להבחין במספר מאפיינים שמקשים על קבלת ההחלטות של השחקן:

1. הוא צריך לשתף פעולה עם שחקן נוסף, וביחד להתחרות מול זוג יריבים.
2. כל אחד מהשחקנים רואה את הקלפים ביד שלו ואת ההכרזות (פעולות) שננקטו על ידי השחקנים האחרים. הוא אינו רואה את הקלפים שבידי השחקנים האחרים.
3. התקשורת בין השחקנים היא מוגבלת ביותר.
4. מסגרת הזמן המוקצבת לכל החלטה אינה מאפשרת לשחקן לסרוק את מרחב החיפוש האדיר של המצבים האפשריים. (ישנם כ8.45×10^{16} אפשרויות שונות למצב המלא של הקלפים כאשר היד של השחקן ידועה)

התקשורת בין השחקנים נעשית על ידי ההכרזות, באמצעותן הם מעבירים מסרים על הקלפים שהם מחזיקים ביד (כגון מספר הקלפים מסדרת העלה, או מספר האסים שהשחקן מחזיק בידו). מספר מהלכי ההכרזה האפשריים בברידג' הוא גדול מכדי שיהיה ניתן לנתח כל אחד בנפרד. האסטרטגיה של בחירת ההכרזות של כל אחד מהזוגות במשחק הברידג' ידועה כ"שיטת ההכרזה" של הזוג. בנוסף לכך ניתן לאפיין כל שחקן ביכולת הערכה לגבי הפעולות שלו במצבים שאינם מכוסים על ידי שיטת ההכרזה של הזוג.

בעזרת המודל של השחקן, המכיל את אסטרטגיית הבחירה של הזוג, אנו מממשים את אלגוריתמי ה-PIDM וה-RBMBMC. מרחב המצבים בצמתים השונים מייצג את הידיים האפשריות שמחזיק השחקן, ומרחב הפעולות מייצג את ההכרזות האפשריות העומדות בפני השחקן.

תהליך הלמידה

תהליך הלמידה המתואר בעבודה כולל שלושה שלבים עיקריים:

בשלב הראשון יוצרים דוגמאות עבור הלמידה. הדוגמאות נוצרות על ידי בחירה אקראית של מצבים אפשריים של הסביבה, תוך סינון מצבים טריוויאליים, שבהם אסטרטגיית הבחירה של הסוכן מציעה עבורם רק פעולה אפשרית אחת. בעבודה זו הגרלנו חלוקות ידיים באופן אקראי. לכל אחד ממהלכי ההכרזה נמדדה כמות המשאבים שנדרשה לשם קבלת ההחלטה על כל הכרזה. בסופו של התהליך מוינו מהלכי ההכרזה על פי סדר יורד של משאבים שנדרשו לשם קבלת ההחלטות.

בשלב השני מתייגים כל אחת מהדוגמאות האלה, על ידי הצמדה של פעולה "נכונה" לכל אחת מהן. לשם כך אנו נעזרים בתהליך ה-PIDM המוצג לעיל, כאשר היתרון הוא שבאפשרותנו להשקיע יותר משאבים בתהליך הסימולציה מאשר בעת היישום.
בשלב זה של תהליך הלימוד בעבודה זו, יצרנו דוגמאות אינפורמטיביות של ידיים לכל מצב ובצענו תיוג שלהן. הדוגמאות הוכנסו לרשת ההחלטה והעץ עודכן בעזרת מנגנון הסיווג ID3. לאחר כל העדכון החליפה הרשת שנוצרה בעת הלמידה את העץ שהיה מבוסס על אסטרטגיית ההכרזה של הזוג בלבד.

בשלב השלישי מתבצע ניתוח הדוגמאות המתויגות ובניית רשת החלטה מעודכנת, הכוללת את המידע על הדוגמאות החדשות. אנו מעדנים את הרשת על ידי הוספה של צמתים המהווים חיתוך של צמתים קודמים ברשת או הפרדה של צמתים שהכילו אי-ודאות לגבי הפעולה המוצעת לאותו מצב.

דגש מושם על האימון המשותף של הסוכנים. לאימון כזה יש יתרונות לעומת תהליך למידה רגיל של כל סוכן בנפרד. התקשורת החופשית בזמן האימון מאפשרת לסוכן לקבל במישרין את ההחלטות של הסוכן השותף בכל מצב נתון, במקום לעשות סימולציות שלהן, וכך נבנית רשת החלטה מדוייקת יותר.
שיפור משמעותי נוסף בתהליך הלמידה הוא תהליך הלמידה האיטרטיבי, שבו נעשתה החלפת מידע ותיאום שוטף של אסטרטגיות הפעולה בין השותפים, כך שאסטרטגיה זו תוקנה והוחלפה מדי פעם. הדבר נעשה על ידי עצירה של תהליך הלמידה, תיאום אסטרטגית הבחירה בין השותפים, וחזרה להמשך הלמידה. לשיטה זו יש שני יתרונות עיקריים - דיוק ויעילות: תהליך הדגימה משתפר עקב מידע מדויק יותר על האסטרטגיה של שותפי הסוכן וכך לבצע הערכות יותר טובות על המצבים האפשריים שבהם נמצא השותף. כמו כן, כאשר הסוכן מבצע את הסימולציות, מספר הפעולות החזויות עבור השותפים מצטמצם, דבר המקטין את המשאבים הדרושים לביצוע החיפוש.

<u>בחינת המערכת</u>

כדי לבחון את המערכת נערך מבחן שהתבסס על בעיות הכרזה אקראיות. בשלב הראשון הוגרלו 2000 חלוקות ידיים לשם מציאת המצבים האפשריים של ידיים. לאחר מכן נלמדו 100 מהלכי ההכרזה שדרשו את המשאבים הגדולים ביותר.

בכל אחת מבעיות ההכרזה הושוותה התוצאה שהושגה אל התוצאה האופטימלית האפשרית של הבעיה, ותורגמה לניקוד בשיטת IMP.

לאחר לימוד 100 ידיים המערכת הראתה שיפור משמעותי ממוצע של כ1- IMP לבעיה. זו תוצאה יפה מאוד, לאור העובדה שהתחרויות מוכרעות לעיתים קרובות בהפרשים של פחות מ0.1- IMP לבעיה.

ומה הלאה?

נושא חשוב באלגוריתם ה-PIDM הוא הקצאת המשאבים בשלבים השונים שלו. יש צורך לחלק את המשאבים בין מנגנון הדגימות (אלגוריתם ה-RBMBMC) ובין הקצאת המשאבים להמשך הסימולציה בכל אחד מהצמתים בעץ הסימולציה.

בעבודה זו השתמשנו בחלוקה פשוטה של המשאבים באופן שווה בין ענפי העץ השונים. האלגוריתם מאפשר לנו גמישות רבה בחלוקת המשאבים וזה בסיס למחקר נוסף על אופן חלוקת המשאבים האופטימלית.

העבודה שנעשתה היא כללית לגבי אופי הסביבה שבה אנו עובדים. האלגוריתם המוצע ותהליך הלמידה מותאם למספר כלשהו של קבוצות סוכנים מתחרות. על מנת להתאים את האלגוריתם לסביבות נוספות יש להגדיר את המשתנים המודולרים הבאים, בלי צורך לשנות את האלגוריתם:

1. סט תכונות מאפיינות של הסביבה.
2. אסטרטגיית התחלתית לבחירת פעולות של הסוכנים (לא הכרחי שכן ניתן להתחיל ללא אסטרטגיה כלל).
3. פונקצית הערכה לגבי מצב הסביבה.

אנו מאמינים שהאלגוריתמים המוצגים בעבודה זו יעזרו לשפר את יכולתנו לבנות סוכנים, אשר מסוגלים ללמוד את הסביבה ולהתאים עצמם אליה באופן עצמאי, תוך כדי שיתוף פעולה עם סוכנים שותפים והתחרות בקבוצות אחרות של סוכנים בעלי מטרות מנוגדות.