ייצור תכונות רקורסיבי על ידי אינדוקציה מבוססת ידע

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר

מגיסטר למדעים במדעי המחשב

ליאור פרידמן

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

אדר התשע"ז ,חיפה, מרץ 2017

ארצה להודות לשאול מרקוביץ' על הסבלנות, הגמישות והעזרה הרבה לכל אורך הדרך.

תודה רבה הולכת גם למדור לימודי מוסמכים ולמזכירות הפקולטה, שהסכימו באדיבות לאשר דחייה אחר דחייה. תודה רבה מקרב לב.

לבסוף, ארצה להודות מקרב לב להוריי, חברי, וחברתי על התמיכה, הסבלנות, והדחיפה הרבה. בלעדיכם זה לא היה קורה.

תקציר

בשנים האחרונות, אנו רואים שימוש רב ונפוץ בשיטות למידת מכונה במגוון רב של תחומי עיסוק כגון ביולוגיה, גיאולוגיה ורפואה. רוב השיטות הללו מסתמכות על השיטה האינודקטיבית: בהינתן קבוצה של דוגמאות מתוייגות, הן מנסות למצוא רעיון מאחד המסביר את התיוג בצורה מדוייקת ככל הניתן. למידת מכונה הוכיחה את עצמה עבור משימות בהן כמות גדולה של מידע מתוייג קיים וזמין, וניתן להגדיר תכונות ברורות ובעלות יכולת הפרדה טובה. למרות זאת, במקרים רבים אין מתאים אוסף התכונות למשימה זאת.

דרך אחת נפוצה להתמודד עם בעיה זו היא לייצר תכונות, למשל על ידי שילוב תכונות קיימות במגוון דרכים. אם זאת, דרכים אלו מוגבלות בכך שהן מסתמכות רק על המידע המקודד בתכונות הקיימות. כאשר בני אדם ניגשים ללמידה (אינדוקטיבית), הם לרוב מסתמכים על ידע עולם חיצוני לבעיה על מנת לפתור אותה בצורה אפקטיבית. בהינתן שני חולים במחלה גנטית מארצות המוצא פולין ורומניה, למשל, אלגוריתם למידה לא יצליח להשתמש במידע רפואי על מנת להסיק מסקנות ולהכליל מדוע הם חולים. רופא, לעומת זאת, ישתמש בהבנתו ובידע החיצוני שלו על מנת להסיק כי חולים שארצות המוצא שלהן במזרח אירופה הינם בסיכון.

בעבודתנו, אנו מציגים אלגוריתם לייצור תכונות המסוגל לנצל ידע חיצוני בצורה הדומה לזו של בני אדם. אנו מניחים בעבודתנו כי מקור ידע זה נתון בצורה רלציונית (או בצורת שלשות עובדות, הדבר שקול). האלגוריתם מתייחס לערכי התכונות הקיימות כאובייקטים בעלי משמעות, ובונה בעיית למידה מעליהם, תוך שימוש בתיוג קיים ובמקור הידע על מנת לייצר תכונות לבעיה חדשה זו. לאחר בנייתה, האלגוריתם משתמש בשיטות מוכרות של למידת מכונה על מנת לייצר מסווג לבעיה. מסווג זה ישמש אותנו כתכונה חדשה לבעיה המקורית. בדוגמה לעיל, למשל, האלגוריתם שלנו ייקח את ערך התכונה "ארץ מוצא", ויבנה בעזרתן מסווג שיסיק מסקנה דומה לזו של הרופא האנושי.

בעת ביצוע תהליך זה, ישנם מספר אלמנטים אליהם יש לשים לב: ראשית, יש לייצר תיוג לדוגמאות בבעיית הלמידה המיוצרת. אנו עושים זאת על ידי לקיחת התיוג הנפוץ ביותר עבור הערך בבעיה המקורית. נושא נוסף הוא השימוש במסווג הנוסף. לאחר יצירת המסווג, נוכל להכיל אותו כתכונה לבעיה המקורית על ידי בדיקת הערך של המסווג על ערכי התכונה הנבחרת. לבסוף, עלינו לדאוג לבעיה הפוטנציאלית של התאמת-יתר. בבעיות למידה בכלל, ובבעיות מועטות דוגמאות במיוחד, קיים סיכון של יצירת מסווג המותאם לדוגמאות המוצגות ולא לקונספט כללי. על מנת למנוע התאמת-יתר, התעלמנו מבעיות נוצרות אם הן היו קטנות מאוד. בנוסף, העובדה כי ניתן להשוות את התכונה הנוצרת לתכונות קיימות מקלה על הסרת תכונות בעלות ביסוס חלש בקבוצת הדוגמאות.

התמקדנו בעבודתנו בעיקר בתחום סיווג הטקסטים, בעיה בה יש להפריד טקסטים לקטגוריות שונות לפי התוכן שלהם. בבעיה מסוג זה, ניתן להשתמש במילות הטקסט כישויות בעלות משמעות, ולהצמידן לישויות במקורות ידע רלציוניים כגון

Freebase ו-YAGO. ביצענו מגוון רחב של ניסויים על מנת לבחון את האפקטיביות של השיטה שלנו ושל התכונות המיוצרות, כולל מגוון של בעיות למידה, אלגוריתמי למידה שונים המשתמשים בתכונות המיוצרות, וכן השוואה לשיטות אחרות לשימוש במקור ידע רלציוני לייצור תכונות. כמו כן, ביצענו ניתוח איכותי וכמותי של התוצאות ממגוון כיוונים. התוצאות שלנו מראות כי השימוש בידע חיצוני מגדיל בצורה משמעותית את הדיוק של מספר אלגוריתמי למידה מוכרים. מעבר לכך, התוצאות מראות כי השיטה שלנו טובה מהותית משיטות לא-מתוייגות אחרות אליהן השוונו את תוצאותינו.