241 ייעול אלגוריתם ליצירת קבצי נתונים אורידינליים עם תבנית מונוטונית ורעש לא מונוטוני

ניר לוצקי ואילאיל בהט

1. המוטיבציה לפרויקט

המוטיבציה לביצוע הפרויקט נבעה מקושי שמצאנו בהערכת ביצועי מודלים מדליית מידע בבעיות החלטה אורדינאליות. בעיות אלה בעלות חשיבות רבה בתחום קבלת ההחלטות בעולם העסקי, שכן הן מכסות טווח רחב של נושאים בהם המשתנה התלוי הוא אורדינלי, למשל: הערכת השקעות, דירוג אשראי, הערכת איכות קורס, בחירת נתיבי תחבורה וכו'. בעיות אלה צוברות תאוצה ועניין בקרב חוקרים, ולראייה כמות המאמרים שנכתבה בתחום בשנים האחרונות. להפתעתנו, על אף הקושי בהערכת ביצועי מודלים, תחום זה לא נחקר עד היום בהיקף מתאים. הבעיה שכיום כל מאמר מדעי על מודל וביצועיו לא כולל דיווח על רמות הרעש הלא מונוטוני בנתונים שעליהם נבדק המודל. מכאן נולד הצורך בכלי לייצור נתונים אורדינאליים מלאכותיים עם רעש ברמות משתנות, שלפיו ייבדקו ביצועי מודלים.

2. סקירת ספרות

עד היום, מחקרים בודדים עסקו באלגוריתמים ליצירת נתונים אורדינליים עם תבנית מונוטונית. רק מחקר אחד למיטב ידיעתנו עסק בכלי דומה עם תוספת של רמות רעש לא מונוטוני בעוצמות שונות (Milstein et al., 2013).האלגוריתם שתואר במאמר מייצר סדרות נתונים עם רמות משתנות של רעש (החל מ 0%, כלומר סדרת נתונים מונוטונית לחלוטין, ועד ל % רעש בהתאם לקלט מהמשתמש). בנוסף, האלגוריתם מכיל פונקציה מונוטונית למשתנה התלוי לפי בחירת המשתמש. אולם, ככל הידוע לנו ומניסוי שקיימנו לא ניתן ליצור בעזרת אלגוריתם זה סדרות נתונים בהיקף של למעלה מכמה אלפים בודדים (האלגוריתם נכתב ב VBA), הממשק למשתמש חלקי וחסר, והנתונים המוגרלים לפיו מפולגים מהתפלגות אחידה בלבד, מה שבוודאי לא תמיד קורה בנתוני אמת. בהתבסס על מאמר זה, נעשה מחקר נוסף (Ben-David et al., 2014) שעשה שימוש באלגוריתם ליצירת Dataset של 1,000 דוגמאות, ובעזרתו בוצע ניסוי לבדיקת רגישות מודלים מדליית מידע לאותו רעש לא מונוטוני. הממצא המשמעותי ביותר מהניסוי הוא שאכן ישנם מודלים שרגישים יותר לרעש מאשר אחרים, ואין אף מודל שהוא ה"מוצלח" ביותר בכל טווח הרעש שנבדק.

3. מתודולוגיה

מטרת המחקר כאמור מתחלקת לשתיים: 1. לשפר אלגוריתם קיים מבחינת יעילות זמן ריצה, תוספת פונקציות שונות וממשק למשתמש. 2. לבצע ניסוי בהיקף נרחב לבדיקת רגישות מודלים לרעש בנתונים אורדינאליים. כדי להשיג את מטרת הפרויקט הראשונה, התחלתנו בלימוד JAVA ובמקביל בכתיבת גרסה בסיסית לתוכנה בסביבת Eclipse. בהמשך, לאחר שראינו כי הגרסה פועלת בצורה תקינה הוספנו פונקציות שונות, בחירת התפלגויות סטטיסטיות וממשק למשתמש. כדי להשיג את מטרת הפרויקט השנייה הפקנו עשר סדרות נתונים מלאכותיות של 10,000 דוגמאות לסדרה בעזרת האלגוריתם שיצרנו וביצענו עליהם ניתוח והשוואה של מודלים בעזרת תוכנת Weka לדליית מידע.

4. ממצאים

במחקר בדקנו 11 מודלים על 10 סדרות נתונים, כאשר בכל סדרת נתונים 10,000 דוגמאות, וזאת לאחר שכבר בוצע ניסוי דומה (Ben-David et al. 2014) אבל עם 1,000 דוגמאות, ואנו משערים שיש הבדלים בתוצאות הניסוי כאשר מייצרים יותר דוגמאות. ואכן, לפי תוצאות הניסוי מודלים שונים שדורגו בניסוי עם 1,000 דוגמאות במקומות גבוהים, דורגו בניסוי שלנו עם 10,000 דוגמאות במקומות נמוכים, ולהיפך. לדוגמא, מודל OCC/OSDL שדורג בניסוי שלנו במקום הראשון לפי מדד KAPPAב 0% רעש, דורג בניסוי עם 1,000 דוגמאות רק במקום ה- 4. כאשר רמת הרעש עלתה ל 10% ה OCC/OSDL דורג בניסוי שלנו במקום ה-3, בעוד בניסוי עם הגדרות זהות עם 1,000 דוגמאות דורג במקום האחרון. ממצא חשוב נוסף הוא ברגישות מודלים שונים לאחוז הרעש הלא מונוטוני בנתונים. למשל, בהתפלגות נתונים אחידה 10,000 דוגמאות ע"פ מדד KAPPA כאשר רמת הרעש עלתה במקצת בלבד, מ 0% (קבצי נתונים מונוטוניים לחלוטין) ל 1%, ביצועי מודל OCC/OLM פחתו מ 0.9 ל 0.354 (ירידה של 60.6%). להשוואה, לפי מדד KAPPA כאשר אחוז הרעש עלה גם כן מ 0% ל 1%, ביצועי מודל LOGISTIC Regression פחתו מ 0.983 ל 0.829 (ירידה של 15.6%), וביצועי מודל J48 פחתו 0.81 ל 0.783 (ירידה של 3.3% בלבד).

5. סיכום

יצאנו למחקר עם 2 מטרות מרכזיות שתוארו בשלב המתודולוגיה. הצלחנו ליצור אלגוריתם יעיל ב JAVAהמסוגל להפיק סדרות נתונים אורדינאליות עם רעש לטובת שימוש עתידי של חוקרים ב Data Mining. באשר למטרה השנייה בפרויקט איששנו את שכתוב בספרות, כי אכן ישנם מודלים שרגישים יותר לרעש בנתונים וביצועיהם דועכים בקצב מהיר יותר מאחרים. לא הצלחנו למצוא מודל יחיד שהוא ה"מוצלח" ביותר בכל רמות הרעש, ולפי תוצאות הניסוי שערכנו סביר כי אין עדיין כיום מודל כזה. כמו כן, ראינו כי הגדלת כמות הנתונים ל10,000 משפרת את איכות התוצאות ודיוק המודלים ב 9% בממוצע. בעתיד, ארגונים עסקיים יוכלו לעשות שימוש בתוצאות ניסוי זה בבחירת המודל המתאים לחיזויים בהתאם לכמות הרעש בנתונים שלהם.