

תקציר הצעת המחקר

תחום כריית מידע (Data Mining) הינו תחום מתפתח הן בעולם האקדמי והן בעולם העסקי. טכניקות כמו הקבצות (Clustering) וסיווג (Classification) של אובייקטים שהמידע אודות תכונותיהם שמור במסדי נתונים הינן נפוצות בתחום זה. טכניקות אלו מבוצעות ע"י חישוב דמיון בין אובייקטים שונים וקיבוץ האובייקטים בהתאם לערכי הדמיון שחושבו. שיפור טכניקות אלו, מעסיק חוקרים רבים. במסגרת זו פותחו אלגוריתמים שונים לביצוע קיבוץ האובייקטים. ניתן לחלק אלגוריתמים אלו בדרכים שונות, כגון, על פי אופן ביצוע ההקבצות - שיטות היררכיות (Hierarchical) לעומת שיטות חלוקה (Partitional), על פי הפלט המתקבל - אובייקט יכול להימצא בו זמנית בקבוצה אחת בלבד או במספר קבוצות (Fuzzy Clustering), ועוד. המשותף לכל האלגוריתמים לקיבוץ אובייקטים הוא הצורך בחישוב דמיון (או מרחק) בין האובייקטים השונים. גם כאן קיימות שיטות שונות לחישוב מידת הדמיון בין האובייקטים. חישוב זה נקבע בין היתר על פי צורת ייצוג המידע המתאר את תכונות האובייקטים וכן על פי סוג התכונות - תכונות עם ערכים רציפים כגון גובה ומשקל, לעומת תכונות עם ערכים בדידים כגון צבע עיניים ומצב משפחתי.

צורת ייצוג המידע המקובלת כיום הינה מוגבלת בהיבטים הבאים :

במסדי נתונים רגילים, חישוב דמיון בין תכונות בעלות ערכים רציפים נחשב לפשוט ויכול להיעשות ע"י שיטות כמו חישוב המרחק האוקלידי בין הערכים המשווים. באופן זה, קל למצוא כי גובה 184 דומה יותר לגובה 186 מאשר לגובה 190. לעומת זאת מציאת דמיון בין תכונות בעלות ערכים בדידים אינה פשוטה כלל. לא קיימות שיטות סטנדרטיות למצוא כי צבע עיניים ירוק דומה יותר לכחול מאשר לשחור, או כי רווק (מצב משפחתי) דומה יותר לגרוש מאשר לנשוי. במקרים אלו קיים ידע אודות הדמיון בין הערכים הבדידים שאינו מצוי בייצוג המידע אודות ערכי התכונות. חסרוננו של ידע זה מונע מציאת דמיון טובה בין האובייקטים השונים ועל כן חוסם קיבוץ טוב של האובייקטים השונים.

במסדי נתונים בינאריים (Binary Databases) ערכי התכונות "נפרסים", וקיום ערך של תכונה מאופיין ע"י ביט דולק (1) ואי קיום הערך מאופיין ע"י ביט כבוי (0). בצורת ייצוג מידע כזו,

אם מפעילים מדדי דמיון סטנדרטיים לחישוב דמיון בין רצפים בינאריים, כגון כמות ה"אחדים" שמופיעים במקומות זהים ביחס לגודל הרצף, לא ניתן למצוא דמיון בין ערכים שונים של אותה תכונה. לדוגמא, אין אינדיקציה כי הגובה 178 דומה יותר לערך 179 מאשר לערך 185, ועל כן לא ניתן לבצע קיבוץ טוב של אובייקטים.

במסדי נתונים עמומים (Fuzzy Databases) שם ערך תכונה יכול להיות חסר, לא מדויק או עמום, כגון: גובה = "נמוך" או גובה = "בערך 180", לא ניתן לבצע קיבוץ אובייקטים באמצעות כלי ניתוח סטנדרטיים כמו SAS ו-SPSS. שיטות ייצוג המידע הקיימות היום במסדי נתונים עמומים אינם יכולים להוות קלט עבור כלי ניתוח סטנדרטיים כך שיופעלו עליהם טכניקות מקובלות של כריית מידע, ועל כן לא ניתן לבצע קיבוץ טוב של אובייקטים בכלים אלו.

על מנת לפתור את הבעיות הנ"ל, המחקר הנידון מציע:

- פיתוח שיטת ייצוג מידע חדשה.
- התאמת שיטת הייצוג לכל סוגי מסדי הנתונים.
- שכלול טכניקות חישוב דמיון על בסיס שיטת הייצוג הנ"ל.
- פיתוח מודל לכריית מידע המבוסס על הטכניקות הנ"ל.

מאפייני שיטת הייצוג

שיטת ייצוג המידע המוצעת במחקר, מתאימה לכל סוגי תכונות אובייקטים - תכונות עם ערכים רציפים, כגון: גובה ומשקל, ותכונות עם ערכים בדידים, כגון: צבע עיניים ומצב משפחתי. בנוסף, שיטת הייצוג טובה לערכים מפורטים (Crisp) כגון גובה = "178", וכן לערכים עמומים (Fuzzy) כגון גובה = "בין 170 ל-180".

מעבר לייצוג המידע המקורי אודות ערכי תכונות האובייקטים, כפי שקיים בשיטות הרגילות כגון, רשומות במסדי נתונים יחסיים או feature vectors, שיטת הייצוג הזו משלבת בתוכה גם ידע אודות הדמיון בין ערכי התכונות המקוריים לבין שאר הערכים האפשריים של כל תכונה. ידע זה הוא המאפשר שיפור בטכניקות של קיבוץ אובייקטים.

השיטה משלבת רעיונות מתורת הלוגיקה העמומה (Fuzzy Logic) וכן רעיונות הלקוחים מבסיסי נתונים בינאריים ומבסיסי נתונים עמומים. מבסיסי נתונים בינאריים מאמצים את הרעיון להפוך

כל תכונה לרצף של נתונים בהתאם לכלל הערכים האפשריים עבור התכונה המסוימת ("פריסה" של ערכי התכונה), ומבסיסי נתונים עמומים מאמצים רעיונות של מודלים לייצוג וחישוב דמיון.

שיטת ייצוג זו תשווה עם שיטות ייצוג אחרות המתייחסות לעמימות כגון Probabilistic feature vectors או Possibility distribution.

ייצוג המידע נבנה במספר שלבים :

בשלב הראשון, נקבעת המסגרת בה יאוחסן המידע. מסגרת זו היא למעשה מטריצה בה שורות מייצגות אובייקטים שונים ועמודות מייצגות ערכי התכונות השונות האפשריים עבור האובייקטים השונים.

בשלב השני, עבור כל תכונה של האובייקטים, מחושבת מטריצת דמיון בין הערכים השונים של התכונה המסוימת. עבור תכונות עם ערכים רציפים, מטריצת דמיון זו יכולה להיווצר באופן אוטומטי ע"י חישוב פונקציות דמיון מתאימות, ועבור תכונות עם ערכים בדידים, מטריצת דמיון זו מחושבת בעזרת מומחה לתחום הבעיה הנידון. הערכים במטריצות הדמיון הם בין 0 ל-1, כאשר 0 מציין חוסר דמיון מוחלט, 1 – מציין דמיון מוחלט, דהיינו זהות, ומספרים בין 0 ל-1 מציינים דמיון חלקי.

בשלב האחרון, עבור כל אובייקט, הערך המקורי של כל תכונה שלו, מוחלף ברצף ערכים המייצגים על פי מיקומם את מידת הדמיון של ערך מסוים לערך המקורי של התכונה.

התאמת שיטת הייצוג לכל סוגי מסדי הנתונים.

המחקר מפרט את התאמת שיטת הייצוג למסדי נתונים רגילים, למסדי נתונים בינאריים ולמסדי נתונים עמומים, ומשווה בין שיטת הייצוג החדשה לבין שיטות ייצוג אחרות הקיימות בבסיסי הנתונים השונים.

במסדי נתונים רגילים, ערך כל תכונה מוחלף ברצף מספרים המתארים דמיון בין ערך אפשרי לערך התכונה המקורי. מציאת דמיון בין האובייקטים השונים יכול להתבצע בשיטות מקובלות כמו חישוב מרחק אוקלידי. הוספת הידע אודות הדמיון של ערכי תכונות שונים לערך המקורי של התכונה מאפשר קיבוץ אובייקטים טוב יותר עבור כל סוגי התכונות.

במסדי נתונים בינאריים, כל ביט במסד הנתונים הבינארי, מוחלף ברצף ביטים המתארים את הדמיון לערך המקורי של התכונה. רצף של "אחדים" מציין דמיון מוחלט, רצף של "אפסים" מציין חוסר דמיון מוחלט, ורצף מעורב של אפסים ואחדים מציין דמיון חלקי. המחקר מתאר את הדרך לייצור רצפים אלו כך שיהא ניתן להפעיל עליהם מדדי דמיון המקובלים במסדי נתונים בינאריים וע"י כך לקבל קיבוץ טוב של האובייקטים.

במסדי נתונים עמומים, ערכים עמומים, כמו "גבוה" כשעוסקים בתכונה גובה, מוחלפים ברצף מספרים המתארים דמיון של ערכי התכונות האפשריים לתיאור העמום המקורי שהתקבל. באופן זה, יהא ניתן להפעיל מדדי דמיון מקובלים גם על בסיסי נתונים עמומים ויהא ניתן לבצע קיבוץ טוב של אובייקטים בעזרת כלי ניתוח סטנדרטיים.

שכלול טכניקות חישוב דמיון על בסיס שיטת הייצוג

המחקר מציג דרך ייחודית לחישוב דמיון המותאמת לשיטת הייצוג. שיטת חישוב זו מושווית עם שיטות אחרות הנהוגות בבסיסי הנתונים השונים. שיטה זו מאפשרת חישובי דמיון טובים יותר מאשר אלו שמתקבלים בשיטות האחרות. השילוב של שיטת ייצוג המידע עם שיטת חישוב הדמיון המותאמת לה, מבטיח ביצוע קיבוץ אובייקטים טוב יותר מאשר בשיטות הקיימות היום.

פיתוח מודל לכריית מידע המבוסס על הטכניקות הנ"ל

המחקר מיישם את מודל הייצוג המוכלל באב-טיפוס שיאפשר בחינה של שיטות הקבצה שונות. המודל יאומת ויעומת עם נתוני מבחן המקובלים בתחום ההקבצה, ויושווה למודלים מתחרים כגון Gaussian uncertainty model ו-Possibility distribution model. המחקר יוכיח את השיפור המתקבל בקיבוץ אובייקטים, ע"י השוואת תוצאות הקבצה שבוצעו בעזרת שיטת הייצוג החדשה, עם תוצאות המתקבלות בשיטות ייצוג מידע אחרות. מודל מוכלל זה יאפשר שיכלול של מגוון טכניקות הקבצה וסיווג, דבר שיתרום לתחום כריית המידע על יישומיו השונים.