



## עדכון רציף של חוקי סיווג לתמיכה בסביבת נתונים דינאמית

אנה חלמסקי

ביה"ס למנהל עסקים

אוניברסיטת בר-אילן

רועי גלברד

ביה"ס למנהל עסקים

אוניברסיטת בר-אילן

### תקציר

בתהליכי סגמנטציה דינאמיים רבים אנו מסווגים מקרים חדשים המתקבלים מזרם הנתונים על פי מודל שנבנה על בסיס מקרים קודמים. כל עוד המקרים החדשים "דומים מספיק" לסגמנטים הקודמים, הסיווג מתנהל באופן מהיר וחסכוני. עם זאת, כאשר מקרה חדש שונה מהותית מסגמנטים קיימים, נדרשת בחינה מחודשת של הסגמנטים שנוצרו בעבר. הבדיקה המחודשת עשויה לגרום ליצירת סגמנטים חדשים או לעדכון הקיימים. במחקר זה אנו מניחים שבסביבות נתונים דינאמיות של נתוני עתק לא ניתן לבחון מחדש את כל נתוני העבר ולכן אנו מציעים להשתמש במאגרי זיכרון קטנים (data buffers) המשמשים לאחסון של מקרים נבחרים כחלופה לשימוש בכל נתוני העבר. אנו מציגים מנגנון דינאמי אינקרמנטלי התומך בסגמנטציה וסיווג של זרם הנתונים, ללא שדה מטר, בזמן אמת. על מנת להפחית את המאמץ החישובי של תהליך סגמנטציה בסביבות דינאמיות ועוסקות נתונים, המודל המוצע Dynamic Classification Unit (DCU) מבצע עדכונים רק על סמך הנתונים במאגרי הזיכרון המצומצמים. הערכת מודל ה-DCU מוצגת באמצעות השוואה עם שתי גישות מקובלות לניהול זרם הנתונים ועדכון סגמנטים: גישה סטטית שאינה מאפשרת יצירת סגמנטים חדשים או מיזוג של הקודמים וגישה דינאמית שמאפשרת יצירת סגמנטים חדשים או מיזוג הקודמים, אך העדכון מתבצע על סמך כל נתוני העבר בסגמנט הרלוונטי. בסביבות נתונים דינאמיות, ויזואליזציה של תהליך סגמנטציה לאורך זמן לרוב אינה מאפשרת למשתמש לעקוב אחר מספר היבטים באופן סימולטני, כגון מעקב על רמת הקבוצה ורמת הפרט, בקרת הגרסאות או מעקב אחר קצב עדכון הסגמנטים. המחקר מציג שיטת ויזואליזציה מקיפה, המכונה להלן ExpanDrogram, שנועדה לתמוך במסווגים דינאמיים הפועלים בסביבת נתוני עתק בכפוף לשינויים במאפייני הנתונים. השיטה מאפשרת למשתמש לשלוט על מגוון רחב של פרמטרים על מנת למקסם את ההתאמה האישית של בעיית הסגמנטציה למשתמש. בנוסף לויזואליזציה עצמה, המשתמש יכול לבחור בתצוגה שכבות נוספות (layouts) שמדגישות היבטים ספציפיים של תהליך הסגמנטציה, כגון רובד של מגמות חדשות או רובד של ערכים חריגים.

**מילות מפתח:** סגמנטציה דינאמית אינקרמנטלית, סיווג, ניתוח אשכולות, ויזואליזציה של תהליך הסיווג

## מבוא

תהליכים של גילוי ידע וכריית נתונים הפכו לחלק בלתי נפרד מהניסיונות המתמידים של הארגונים להרחיב את נכסיהם הבלתי מוחשיים. הארגונים שואפים להשגת יתרון תחרותי באמצעות ניתוח רציף של הנתונים שזורמים ממקורות שונים בזמן אמת. מאגרי מידע הולכים וגדלים לממדים בהם כלים אנליטיים מסורתיים לניתוח וויזואליזציה אינם יכולים לעמוד בעומס הנתונים (Fayyad & Stolorz, 1997) וקיים צורך מתמיד בחיפוש אחר שיטות יעילות ויציבות יותר (Fan et al., 2014). הצמיחה הבלתי פוסקת בהיקף הנתונים הובילה להבנה ברורה שכלים אנליטיים צריכים להיות מותאמים יותר לסביבת נתונים דינאמית. בפרט, כלים אלה צריכים לאפשר פתרונות לניתוח, חיזוי וקבלת החלטות מהירים, חסכוניים ונגישים בכל רגע נתון (Park et al., 2001).

תהליכי פילוח (segmentation) וסיווג (classification) של האוכלוסיות הנחקרות וחיפוש אחר הדמיון בין הסגמנטים השונים (כגון לקוחות, מדידות, מוצרים או אירועים) הפך לאחד מתחומי העניין המרכזיים בכריית נתונים (Deza M.M. & Deza E., 2014). הליך זה כרוך בעלויות עיבוד גבוהות ובחוסר יציבות אלגוריתמית. הספרות בתחום למידת מכונה (machine learning) משתמשת במספר מונחים למשימות הסגמנטציה השונות, כגון supervised classification (שמהווה חלק מלמידת מכונה מונחית) (Milligan & Hirtle, 2012) או unsupervised clustering (שמהווה חלק מלמידת מכונה לא מונחית) (Gelbard et al., 2007). המונח הראשון משמש למקרים בהם שדה המטרה ידוע מראש. לדוגמה, רופא שמפענח את תוצאות הבדיקות של המטופלים אמור לסווג את המקרה לאחת מהמחלות הקיימות ובהתאם לזה להמליץ על

הטיפול. המונח השני משמש למקרים בהם שדה המטרה אינו ידוע מראש, כך שהתוצר הסופי של החלוקה לסגמנטים (אשכולות) דורש פירוש והסבר (interpretation) של מומחה ידע. לדוגמה, מרצה שמפתח שיטות הוראה היברידיות נעזר בניתוח אשכולות על מנת לאתר קבוצות הומוגניות של תלמידים, ומחפש הסבר לדמיון בין התלמידים בכל אשכול ולשוני בין האשכולות, במונחים של התאמה/אי התאמה של שיטת ההוראה ומאפייני התלמידים.

סביבות נתונים דינאמיות ומרובות נתונים מאופיינות ע"י זרימה מתמדת של מקרים חדשים שעשויים להיות בעלי תכונות קיימות או חדשות, וגם התכונות עצמן עשויות להשתנות כל הזמן. בסביבות כאלה אי אפשר לחשב מחדש את כל מדדי הדמיון בכל פעם שמופיע מקרה חדש - אפילו לא באמצעות חומרה ואלגוריתמיים משוכללים ביותר (Guha & Mishra, 2016). סביבות נתונים דינאמיות מאפיינות תחומים שונים ומגוונים, כגון: ניתוח רשתות חברתיות, בנקאות, שוק ההון, ייצור, רפואה ובטחון. הצורך ההולך וגובר במציאת כלים אנליטיים יעילים יותר הוביל להתפתחות התחום של **ניתוח נתונים אינקורמנטי** (Shah Siddharth et al., 2012).

בתהליכי סגמנטציה רבים, מקרים חדשים מסווגים לסגמנטים השונים בהתאם למודל שנבנה על בסיס מקרי העבר. כל עוד המקרים החדשים "דומים במידה מספקת" לסגמנטים הקיימים המכילים מקרים קודמים, תהליך הסיווג מתקדם בצורה מהירה וחסכונית. אולם, כאשר מקרה חדש שונה באופן מהותי מהסגמנטים הקיימים, נדרש לעדכן את החלוקה לסגמנטים, דבר שיכול להוביל ליצירת סגמנטים חדשים או לארגון מחדש של סגמנטים קיימים. הנחת היסוד המרכזית במחקר הנוכחי היא שבסביבות נתוני עתק Big Data שכוללות שינויים דינאמיים, לא ניתן

**במלואו.** ההשערה היא שהגישה הדינאמית-אינקרמנטלית תהיה מהירה יותר מהגישה הדינאמית, וככל שנפח הנתונים הולך וגדל, היתרון בשימוש ב buffers מוגבלים הופך ליותר ויותר משמעותי. כמו כן, הגישה הדינאמית-אינקרמנטלית צפויה להיות יעילה וגמישה יותר מהגישה הסטטית עקב האפשרות ליצירת סגמנטים חדשים.

השוואה בין שתי הגישות הדינאמיות, כאשר בגישה האינקרמנטלית גודל ה-buffer מוגבל, ובגישה ה-"לא אינקרמנטלית" גודל ה-buffer אינו מוגבל, מראה כי הגישה הדינאמית-אינקרמנטלית הנבחנת במחקר הנוכחי, מצליחה להשיג גמישות מירבית ויעילות חישובית על חשבון ירידה מסוימת באיכות התוצאה.

בסביבות נתונים דינאמיות שתוארו לעיל, ויזואליזציה מקיפה של תהליך הסיווג או של תוצר הסיווג מצד אחד יכולה לשפר את תהליך הפרשנות עבור המשתמש ולאפשר לו להשתתף באופן פעיל בתהליך קבלת החלטות (Shneiderman et al., 2016). מצד שני, ויזואליזציה מקיפה לעיתים קרובות לא מאפשרת למשתמש לעקוב במקביל אחר היבטים שונים, כמו למשל לראות בו-זמנית את רמת הקבוצה ורמת הפרט, לבצע בקרה אחר גרסאות, לצפות ביצירת מגמות חדשות והופעת מקרים חריגים (D. A. Keim, 2002). במקרים רבים המשתמש נאלץ להסתפק בתמונה סטטית של נקודת מצב מסוימת ולא מצליח לעקוב אחרי שינויים דינאמיים בתהליך הסגמנטציה (D. Keim et al., 2013).

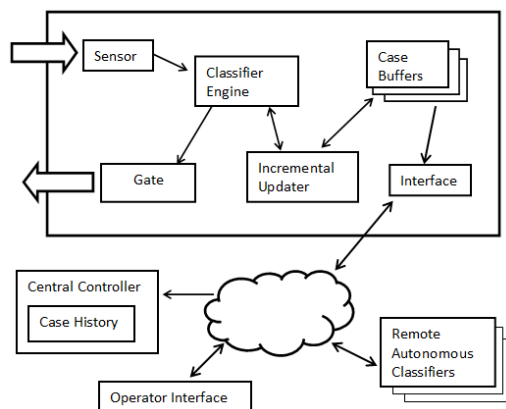
שיטת הויזואליזציה ExpanDrogram שפותחה במהלך המחקר, תומכת במסווגים דינאמיים הפועלים בסביבות נתונים גדולות תוך התחשבות באפשרות של שינוי במאפייני הנתונים. השיטה מציעה מגוון רחב של תכונות לצורך ייעול מיטבי והתאמה אישית של תהליך קבלת החלטות לכל תחום או לכל משתמש.

לבחון מחדש את כל נתוני העבר, ולכן יש צורך בפתרון אינקרמנטלי שבו רק חלק קטן מהנתונים נבדק, ועל סמך נתונים אלו מתבצע עדכון של הסגמנטים. במסגרת המחקר הנוכחי פותח מודל המאפשר הפעלה כסוכן אוטונומי. המודל מכונה בשם Dynamic Classification Unit (DCU) והוא מבוסס על שימוש במקרים נבחרים בלבד של נתונים, המאוחסנים במאגרי זיכרון קטנים (buffers) (Solt & Horovitz, 2012). כתוצאה מההפחתה המשמעותית בזמן העיבוד, ה-DCU יכול לשמש כמסווג הפועל בזמן אמת.

על מנת לאפשר התאמה של המודל ליישומים שונים של סיווג נתונים דינאמי, עומד לרשות המשתמש מגוון פרמטרים, כגון סוג המשתנים, רמת הרגישות, מספר סגמנטים התחלתי, מספר מינימלי או מקסימלי של מקרים שייכנסו ל buffers, האלגוריתם ליצירת אשכולות ועוד. תהליך קבלת ההחלטות תלוי בהעדפות המשתמש ו/או בסטנדרטים המקובלים בתחום. המשתמש יכול לקחת חלק פעיל בניחות ופרשנות של התוצר הסופי. לחילופין, במקרים של זרימת נתונים דרך חיישנים, קבלת החלטות עבור חלוקה לסגמנטים, זיהוי מקרים חריגים ומגמות חדשות יכול להתבצע באופן אוטומטי ללא התערבות המשתמש. המשתמש יכול לכייל את המסווג בהתאם לדרישות עבור רמת הרגישות והמשך התהליך יתבצע בהתאם לכך.

ההערכה של יעילות ה-DCU בוצעה על ידי השוואה עם שתי גישות נוספות לניהול וניתוח זרם הנתונים. הגישה הראשונה היא גישה סטטית בה מספר סגמנטים קבוע מראש ולא ניתן לשינוי, כך שכל מקרה חדש מתווסף לאחד הסגמנטים הקיימים. הגישה השנייה היא גישה דינאמית שמאפשרת פיצול או מיזוג של סגמנטים כחלק מתהליך העדכון, אך הופעתם של מקרים שעוברים את סף הרגישות גורמת לבחינה מחודשת של הסגמנט הרלוונטי

ארכיטקטורת המודל מוצגת בתרשים מספר 1. המונח "חיישן" Sensor מייצג את "המשפך" שדרכו זורם זרם הנתונים. הנתונים זורמים בזמן אמת מהחיישן ל"מסווג" Classifier Engine שתפקידו לסווג את המקרים החדשים לסגמנטים קיימים או ליצור סגמנטים חדשים בהתאם לכללי סיווג המבוססים על הגדרת הפרמטרים. תרשים זרימה, המוצג בתרשים 2, מציג את תהליך קבלת ההחלטות עבור כל מקרה חדש (ההסבר המפורט מוצג ב"שלב הריצה"). תוצאת ההחלטה אודות כל מקרה חדש נשלחת למרכיב בשם "שער" Gate. בהתאם להגדרות הפרמטרים (כגון, סף הרגישות, גודל ה- buffer וכו'), ה- DCU מעדכן באופן אינקרמנטלי את אוכלוסיית המקרים ששמורים ב- buffer מסוים. המאמר "Dynamic Classifier and Sensor Using Small Memory Buffers" מציג באופן מפורט את הארכיטקטורה של המודל (R. Gelbard & Khalemsky, 2018).



תרשים 1 – ארכיטקטורה של המודל DCU

### שלב האתחול ושלב הריצה

המסווג פועל בשני שלבים עיקריים: שלב אתחול, ושלב ריצה שבאמצעותו מתבצע סיווג של מקרה חדש שיכול לגרום לעדכון של הסגמנטים שנוצרו קודם. חשוב לציין שה- DCU מסוגל להתחיל לעבוד ללא שלב האתחול (שלב הלמידה), עם buffers ריקים וללא כללי סיווג מוכנים. שלב האתחול מיועד לסייע בקיצור זמן הלמידה או כאשר יש מידע מוקדם

המטרה העיקרית של שיטת ExpanDrogram היא לספק תמונה מקיפה ככל הניתן של תהליך הסגמנטציה על ידי שילוב של רמת הפרט ורמת הקבוצה, מתן "בקרת גרסאות" המאפשרת למשתמש לצפות בהיסטוריה של השינויים ועוד. בחירת הפרמטרים במודל ה- DCU שעומד מאחורי תהליך הסיווג עצמו, משפיעה על הרזולוציה של ה- ExpanDrogram. המשתמש יכול לשנות את ערכי הפרמטרים ולצפות בהשפעתם על התוצר הויזואלי, כגון לזהות עבור אילו רמות רגישות המסווג מצליח לזהות מגמות חדשות. קיימת אפשרות לייעל את הפרשנות של חלקים מסוימים מה- ExpanDrogram, כגון מיקוד רק בזיהוי של מגמות חדשות, רק במקרים חריגים, בתהליך ההיווצרות וההתפתחות של סגמנט ספציפי וכו'. המאמר "ExpanDrogram: A Method for Comprehensive Visualization of Segmentation over Time" מציג את מגוון ההיבטים המעורבים בכיול ובשימוש בויזואליזציה הנ"ל (A. Khalemsky & Gelbard, 2021).

### Dynamic Classification Unit

כאמור, מסווג ה- DCU מבוסס על מאגרי זיכרון קטנים (buffers) שמאחסנים מספר מצומצם של מקרים מייצגים מכל קבוצה, והוא מאפשר עדכון רציף של חוקי סיווג בזמן אמת ותומך בבעיות סגמנטציה בסביבות נתונים דינאמיות של נתוני עתק. המאמר "A Dynamic Classification Unit for Online Segmentation of Big Data via Small Data Buffers" מציג בפירוט את תיקוף המודל (Anna Khalemsky & Gelbard, 2019). והמאמר "Dynamic Classification for Materials-Informatics: Mining the Solar Cell Space" מדגים את השימוש במודל בתחום ידע ספציפי של הנדסת חומרים (Yosipof et al., 2020).

### ארכיטקטורת מודל ה- DCU

המקרה החדש לסגמנט הקרוב ביותר היה מעל סף הרגישות.

3. יצירת סגמנט חדש עבור מקרה חריג.

המצב מתאפשר אם המרחק בין המקרה החדש לסגמנט הקרוב ביותר גבוה מסף הרגישות, אך פיצול הסגמנט לתתי-סגמנטים עדיין לא אפשר את סיווגו או במקרים בהם לא היו מספיק מקרים בסגמנט הקרוב ביותר שהצדיקו את הפיצול.

4. איחוד מספר סגמנטים קיימים לסגמנט אחד.

המצב מתאפשר אם ישנם מספר סגמנטים קרובים שהמרחק ביניהם נמוך מסף מסוים, ואינו מצדיק הפרדה.

טבלה מספר 1 מבארת את משמעות הסימונים בהם נעשה שימוש בתרשים הזרימה המוצג בתרשים 2.

תרשים 2 מציג את תהליך הסיווג של מקרה חדש שמתקבל בזרם הנתונים הדינאמי.

אודות סגמנטים קיימים. במקרים אלו כל סגמנט מכיל מאגר קטן של נתונים מייצגים וכל החישובים של מדדי דמיון מתנהלים מול מאגרי נתונים מצומצמים אלו. גודל ה-buffers מנוהל באמצעות אחד הפרמטרים.

שלב הריצה הוא הסיווג של מקרה חדש. שלב זה יכול להסתיים באחד מארבעת המצבים הבאים:

1. סיווג של מקרה חדש לאחד הסגמנטים

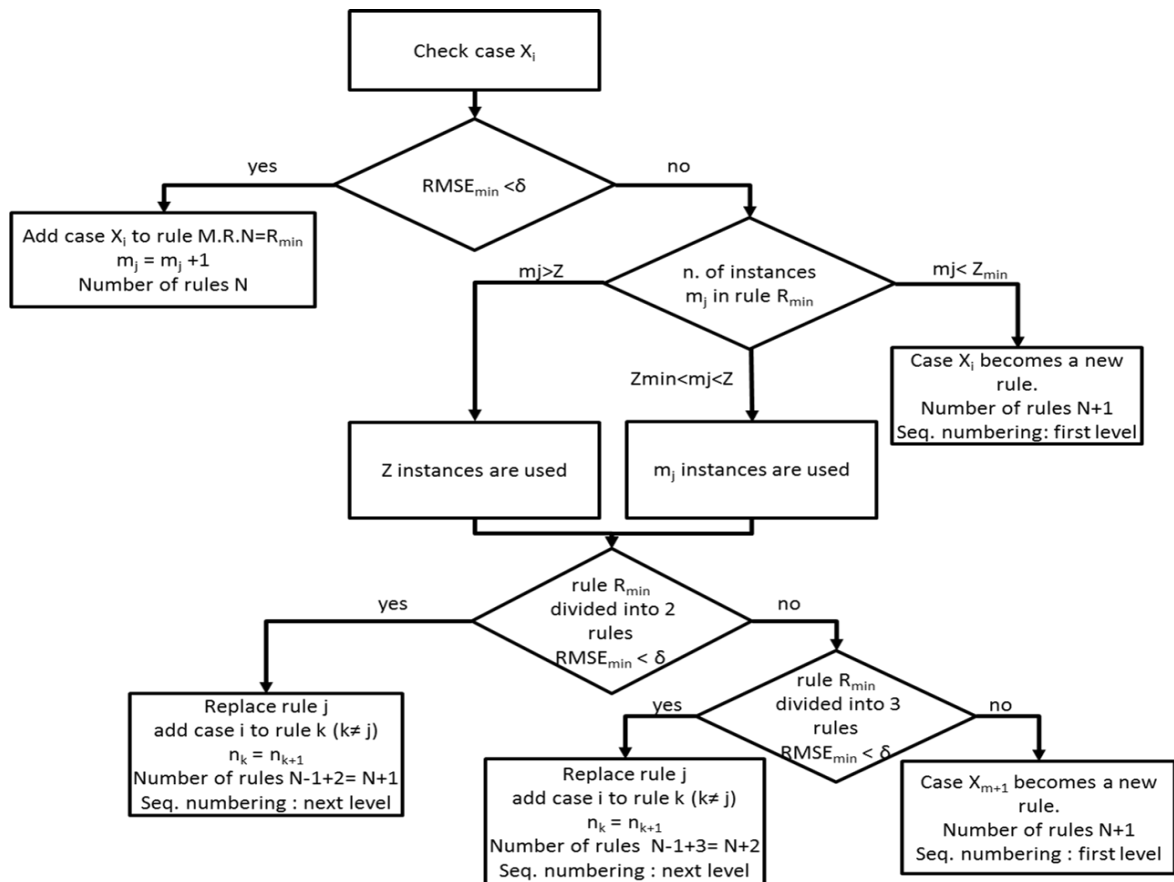
הקיימים. המצב מתאפשר במקרים בהם המרחק בין המקרה החדש לבין הסגמנט הקרוב ביותר קטן או שווה לערך סף הרגישות שנבחר מראש. מכאן והלאה, המרחק נמדד במונחים של מדדי דמיון ומרחק.

2. סיווג של מקרה חדש לאחד מתתי-

הסגמנטים ההומוגניים שנוצרו אחריו פיצולו של הסגמנט הקרוב ביותר. המצב מתאפשר במקרים בהם המרחק בין

| סימון                        | תיאור  |
|------------------------------|--|
| $X_i$                        | מקרה חדש   |
| $N$                          | מספר סגמנטים במועד הסיווג של המקרה החדש  |
| $RMSE_j \quad j = 1 \dots N$ | ערך של מדד מרחק בין המקרה החדש לבין מרכז הסגמנט  |
| $RMSE_{min}$                 | המרחק לסגמנט הקרוב ביותר למקרה החדש  |
| $R_{min}$                    | הסגמנט הקרוב ביותר למקרה החדש  |
| $M. R. N$                    | $RMSE_{min} = RMSE(X_i, R_{min}) = \min(RMSE_j)$<br>מספר סידורי של הסגמנט הקרוב ביותר. המספור מתבצע באופן שמאפשר בקרה על הגרסאות.          |
| $m_j \quad j = 1, \dots, N$  | מספר מקרים בכל אחד מהסגמנטים הקיימים   |
| $m_{min}$                    | מספר מקרים בסגמנט הקרוב ביותר  |
| $\delta$                     | רמת הרגישות (במונחים של מדד מרחק RMSE) שקובעת את מדיניות הסיווג (סיווג מיידי לסגמנט הקרוב ביותר או פיצול של הסגמנט לתתי-סגמנטים הומוגניים) |
| $Z_{min}$                    | מספר מקרים מינימלי ב buffer שמצדיק פיצול   |
| $Z$                          | גודל ה buffer  |

טבלה 1. ביאור הסימונים המוצגים בתרשים הזרימה המוצג בתרשים-2.



תרשים 2 – שלב ריצה וסיווג של מקרה חדש הפרמטרים במודל

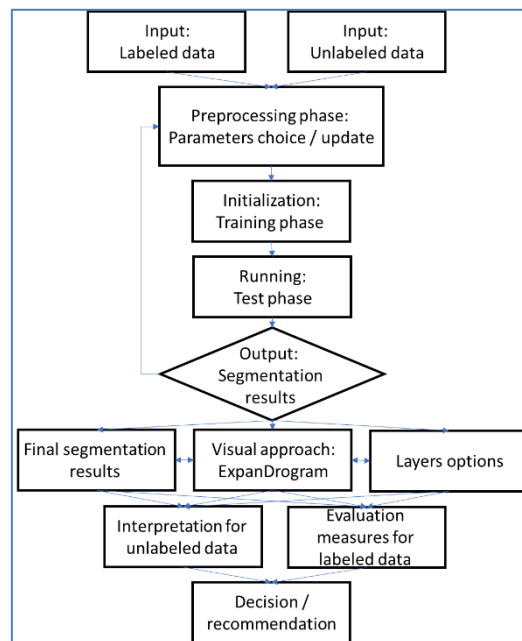
מתייחסים למספר היבטים: רמת הקבוצה (מאפיינים של המקרים המשתייכים לאותו סגמנט), רמת הפרט (מאפיינים של מקרים חריגים), אבחנה בין מקרים חריגים שגרמו ליצירת סגמנטים חדשים לבין מקרים חריגים שהתניעו יצירה של מגמה חדשה, בקרת גרסאות וכו'. ייצוג ויזואלי של תהליך סגמנטציה דינאמית שמבוסס על הפעלת ה-DCU מאפשר למשתמש להשתתף באופן פעיל בתהליך קבלת החלטות. תרשים 3 מציג את שילוב הניתוח הויזואלי בתהליך הסגמנטציה והאינטרפרטציה. מסווג ה-DCU, שבאמצעותו נוצרת ומנוהלת הסגמנטציה של זרם הנתונים, מספק בזמן אמת את "תמונת המצב", והמשתמש יכול לצפות בתהליך כמו גם בתוצרים שלו. באמצעות שינוי של ערכי פרמטרים המשתמש יכול לשנות רזולוציה וכן לקבל תמונה אודות רבדים שונים (layers),

על מנת להביא להתאמה אופטימלית של המודל למאפיינים של תהליכי סגמנטציה שונים, המשתמש יכול לשלוט במגוון פרמטרים ובאופן זה להתאים את המודל לצרכיו. חלק מהפרמטרים הכרחיים לתפקוד בסיסי של המודל (כגון מספר סגמנטים התחלתי, סף הרגישות לסיווג מיידי של המקרה החדש, חלוקה ל- training set / test set). חלק אחר הוא אופציונלי והשימוש בהם נועד לסייע בהתאמה של המודל לדרישות ספציפיות של המשתמש, כגון בחירת משקל שונה לתכונות שונות, מספר סגמנטים מירבי, שיטת מילוי ה-buffer.

### ויזואליזציה Expander

במהלך ניתוחי רגישות רבים עלה צורך בייצוג ויזואלי מקיף של תהליך הסגמנטציה. התוצרים השונים של השימוש במודל

וכך לשפר את קבלת ההחלטות והאינטרפרטציה אודות סיווג הנתונים לסגמנטים.



תרשים 3 – שילוב של תהליך סגמנטציה, ויזואליזציה ואינטרפרטציה של תוצאות

### מרכיבי ה-ExpanDrogram

ה-ExpanDrogram מספקת ויזואליזציה מורחבת ומקיפה שנקודת המוצא שלה היא עץ ה-Dendrogram אליו הוספו מאפיינים וחיוויים חדשים. הדיאגרמה מציגה התפתחות של תהליך הסגמנטציה עם דגש על השלבים הקריטיים (פיצול של סגמנט קיים לתתי-סגמנטים הומוגניים, יצירת סגמנט חדש מהסוג "מקרה חריג" או סגמנט חדש מהסוג "מגמה חדשה". נקודות מפתח מוצגות על ציר הזמן, כאשר לחילופין, ניתן להשתמש בציר המציג את סדר זרימת הנתונים. הציר מוצג במקביל לדיאגרמה עצמה ומציג "רק את מה שחשוב", לדוגמה מועד פיצול הקבוצה או מועד הופעת מגמה חדשה. הציר "מכווץ" באופן מכוון על מנת לאפשר למשתמש לצפות בכל התהליך של הסגמנטציה בבת אחת, גם אם נקודות המפתח נמצאות במרחקים שלא ניתנים להשוואה (לדוגמה, כאשר פיצול של קבוצה אחת התרחש ב-1 לינואר, קבוצה

חדשה הופיעה ב-2 לינואר ומגמה חדשה הופיעה ב-31 לאוקטובר). הסגמנטים המקוריים או סגמנטים חדשים שנוצרו במהלך הסגמנטציה יכולים לכלול מספר מאוד שונה של מקרים, החל ממאות אלפים בסגמנטים המקוריים שרק הולכים ומתמלאים במהלך תהליך הסיווג, ועד סגמנטים עם מקרים בודדים. החשיבות של כל סוג של סגמנט איננה פונקציה של מספר המקרים בו. לדוגמה, בתהליכי סיווג מהסוג גילוי הונאה (fraud detection) גילוי של מקרה חריג (שמייצג הונאה פוטנציאלית) חשוב הרבה יותר מאשר סגמנט עם מיליוני מקרים שמייצגים עסקאות חוקיות. מהסיבה הזאת ה"קופסאות" שמייצגות סגמנטים ב ExpanDrogram הן בגודל זהה. קיימת אפשרות להשתמש בצבעים שונים על מנת לסמן סגמנטים מסוגים שונים.

### מרכיבי תרשימים Layers \ Zoom-In

קיימת אפשרות לחדד נקודות ספציפיות בתרשים ExpanDrogram באמצעות ייצוג ויזואלי של רבדים (layers) שונים (לדוגמה, מיקוד ברובד של מגמות חדשות או מיקוד ברובד של מקרים חריגים). כמו כן קיימת אפשרות לשנות את הרזולוציה של ExpanDrogram עבור סגמנט מסוים ולבחון אותו מנקודת מבט שונה: לצפות בשינויים בקצב הצבירה של מקרים חדשים לסגמנט הנבחר (ניתן להשוות את קצב הצבירה בשני סגמנטים קרובים). הפעולה הזאת יכולה להיות מועילה בלמידה של מגמות חדשות: אם קצב הצבירה עולה אחרי היווצרות המגמה, אך בשלב מסוים הקצב יורד באופן משמעותי ורק מקרים בודדים מסווגים לסגמנט בהמשך, ניתן להסיק שהמגמה תמה. אפשרות נוספת קשורה לגילוי מקרים חריגים (anomaly detection). ניתן לראות את מועד הופעת המקרה שמוגדר כחריג, את פירוט וקטור הנתונים שלו ואת וקטור הממוצעים

בתהליך הסיווג אוחדו לקבוצות. עבור כל קבוצה חושב centroid, ז.א. וקטור הממוצעים של כל התכונות.

שלב 3: יצירת מספר זהה של אשכולות אלגוריתם k-means, שאינו מתייחס לשדה מטרה, הופעל תוך הכוונה לייצר מספר אשכולות זהה למספר העלים שנוצרו בשלב 1. ושוב חושב וקטור הממוצעים עבור כל אשכול שהתקבל.

שלב 4: אינטרפרטציה של תוצאות חושב מדד מרחק RMSE (root mean square error) בין כל וקטור הממוצעים באשכול ובין כל אחד מהווקטורים שהתקבלו בשלב 2. ערך נמוך יותר של RMSE קבע לאיזה ערך של שדה המטרה ישתייך כל אשכול.

שלב 5: השוואה בין דיוק הסיווג של עץ ההחלטה ובין דיוק הסיווג באמצעות ניתוח האשכולות

בשלב זה בוצעה השוואה של שדה המטרה המקורי לתוצאות של שתי גישות הסיווג. מטרת הבדיקה היא להראות שניתוח האשכולות, ללא שימוש בשדה המטרה, מסוגל לספק תוצאה לא פחות טובה מאלגוריתם סיווג שמשמש בשדה המטרה.

### **הערכה באמצעות שלושה בסיסי נתונים**

הערכת מודל ה-DCU בוצעה באמצעות שלושה בסיסי נתונים ובאמצעות מספר ניתוחי רגישות. בסיס נתונים נוסף נבחר כדי להדגים את שיטת הויזואליזציה החדשה של ה-ExpanDrogram. תיאור מפורט של בסיסי הנתונים מוצג בפרק "כלים ונתונים". להלן תיאור של שלושת המטרות העיקריות בהערכת המודל:

המטרה הראשונה היא לבחון את השפעת הפרמטר של רמת הרגישות על התוצרים השונים המתקבלים בתהליך עדכון דינאמי אינקרמנטלי של סגמנטים: (1) מספר

של כל המשתנים של כלל הנתונים. במידה ורק תכונה מסוימת מתוך הווקטור הינה בעלת ערך חריג, ייתכן שמדובר בטעות מדידה. לעומת זאת, אם שילוב של תכונות הסתכם למדד מרחק גבוה במיוחד, ייתכן שמדובר על תופעה ייחודית.

### **שיטת המחקר**

#### **מבחן מקדים - השוואה לתוצרים של עצי החלטה וניתוח אשכולות**

כאמור, יש שני כיוונים עקרוניים בתחום הסגמנטציה: בעיות סיווג (classification tasks) בהן שדה המטרה ידוע ונתון מראש, ובעיות ניתוח אשכולות (clustering tasks) בהן שדה המטרה אינו ידוע כלל.

כבדיקה מקדימה לצורך המחקר, בדקנו האם אלגוריתם של אשכול מסוגל לספק תוצאות קרובות לתוצאות של עצי החלטות. מחקרים קודמים של ברק וגלבארד (Barak & Gelbard, 2011) מראים שניתן להתייחס למסלול מהקודקוד של עץ החלטה לעלה כחוק סיווג בדומה לאשכול המתקבל בניתוח אשכולות. רעיון זה נתן בסיס לבדיקת היתכנות המודל. ההשוואה בין עצי החלטה לאלגוריתם מקובל בניתוח אשכולות k-means נעשתה בהתאם לשלבים הבאים:

שלב 1: הרצה של עץ החלטות תוך שימוש בשדה מטרה ידוע

לשם כך נעשה שימוש בשלושה מודלים של עצי החלטות – J48, REPTree, Random Forest. האלגוריתם יצר מספר מסוים של עלים, וכן נרשמו מדדי איכות הסיווג. חשוב לציין כי השימוש בעצי החלטה נועד אך ורק למטרת הוכחת ייתכנות מתודית. שכן, המודל מיועד להתמודד עם נתונים לא מתויגים (ללא שדה מטרה).

שלב 2: קיצוץ של כל העלים המשתייכים לאותו ערך בשדה המטרה לקבוצה אחת בהתאם למספר הערכים האפשריים בשדה המטרה, כל המקרים שהשתייכו לאותו ערך



### כלים ונתונים

טבלה מספר 2 מציגה את בסיסי הנתונים. שלושת הראשונים שימשו לצורך הערכת ביצועי המודל. בסיס הנתונים הרביעי שימש לצורך המחשת היוזואליזציה.

| סוג הבעיה | N       | מספר תכונות | שם בסיס הנתונים שימוש במחקר |
|-----------|---------|-------------|-----------------------------|
| סיווג     | 1,000   | 4           | LEV                         |
| סיווג     | 20,560  | 5           | OCCUPANCY                   |
| סיווג     | 100,000 | 10          | DEEPSCAPULA                 |
| אשכול     | 420,550 | 14          | JENA CLIMATE                |

טבלה 2 – בסיסי נתונים

בסיס הנתונים LEV נתרם ע"י פרופ' בן דוד (Ben-David, 1992), הוא מכיל נתונים של הערכת איכות ההוראה שניתנו ע"י תלמידי תואר שני במנהל עסקים. התלמידים נתנו משוב לארבעה קריטריונים של איכות ההוראה ובסוף, ציון סופי לכל קורס. בסיס הנתונים Occupancy Detection הורד ממאגר נתונים (UCI Machine Learning Repository: UCI Data Sets, n.d.). בסיס נתונים זה כולל סדרה עיתית של מדידות שנעשו במשרדים, כגון טמפרטורה ולחות. שדה המטרה היה התפוסה של המשרד. בסיס הנתונים deepScapula הורד ממאגר הנתונים (Kaggle: Your Kaggle Home for Data Science, n.d.) רחב של מאפיינים אורטופדיים, כאשר שדה המטרה הינו סוג ההטיה. בסיס הנתונים Jena Climate הינו סדרה עיתית של מאפייני מזג האוויר. בסיס הנתונים הורד ממאגר הנתונים Kaggle.

כל אחד מבסיסי הנתונים נבחר בקפידה על מנת להציג בעיה שונה בתחום הסגמנטציה: בסיס נתונים LEV מציג נתונים אורדינליים ונחשב ל"קשה" מבחינת איכות הסיווג. בסיס הנתונים

סגמנטים סופי, (2) ממוצע המרחקים של כל המקרים שסווגו לסגמנט מסוים ממרכז הסגמנט (במונחים של RMSE), (3) סטית תקן של המרחקים, (4) סוגי הסגמנטים שהתקבלו (סגמנטים מקוריים שלא פוצלו, סגמנטים מקוריים שפוצלו, מהו מספר המגמות החדשות ומהו מספר המקרים החריגים).

המטרה השנייה היא לבחון את אפקט ההתכנסות של תהליך העדכון. ככל שתהליך העדכון של סגמנטים מתקדם, כך גדלים המרווחים בין עדכון לעדכון. במלים אחרות, המודל יוצר סגמנטים ייצוגיים ורמת ה"ייצוגיות" רק הולכת ומשתפרת עם הזמן, כך שהמודל מצליח לסווג כמעט כל מקרה חדש לסגמנטים שנוצרו, ללא צורך בעדכונים נוספים.

המטרה השלישית היא הערכה של יעילות הגישה האינקרמנטלית בהשוואה לגישות הקיימות: הגישה הסטטית (שלא מאפשרת פיצול או מיזוג של סגמנטים) והגישה הדינאמית (שמאפשרת פיצול או מיזוג של סגמנטים, אך דורשת בחינה מחודשת של כל המקרים שהצטברו בסגמנט המסוים). ההשוואה בין שלושת הגישות נעשתה באמצעות מדידת זמני הריצה של המודל. המצב הסטטי בא לידי ביטוי ע"י קביעת הפרמטרים "מספר קבוצות התחלתיות" ו"מספר קבוצות סופיות" כערך קבוע מראש, וללא הגבלה של גודל ה buffer. הגישה הדינאמית באה לידי ביטוי ע"י אי הגבלה של ה buffer וללא הגבלה של הפרמטר "מספר קבוצות סופיות". הגישה הדינאמית-אינקרמנטלית באה לידי ביטוי ע"י הגבלה של גודל ה buffer למספר מקרים קטן יחסית בהשוואה לסך המקרים בתוך הסגמנטים וללא הגבלה של הפרמטר "מספר קבוצות סופיות". יתר הפרמטרים היו זהים לשלושת המצבים.

### היתכנות המודל - דיון בתוצאות:

כל הניסויים הראו תמונה דומה בה ניתוח אשכולות האמצעות k-means סיפק דיוק מעט פחות טוב, אך עקבי מאוד בהשוואה לכל שלושת האלגוריתמים של עצי החלטה. לדוגמה, בבסיס הנתונים "הבעייתי" LEV בו שלושת האלגוריתמים של עצי החלטות סיפקו רמות דיוק נמוכות בין 60.1% לבין 62.8% תוך שימוש בשדה המטרה; k-means סיפק רמות דיוק בין 47.6% לבין 52.3%, ללא שימוש בשדה המטרה.

תוצאות בדיקת ההיתכנות מראות כי בהיעדר שדה המטרה אנו עדיין יכולים לסמוך על אלגוריתמים של אשכול כדי לספק תוצאות מדויקות. בנוסף, עובדה זו מאששת את החלטתנו להשתמש באלגוריתם האשכולות במודל שלנו. התוצאות מאשרות גם את האנלוגיה בין אשכול לנתיב בעץ החלטה.

### התכנסות הסגמנטציה וזיהוי חריגים

תרשים 4 מציג השוואה בין הדינמיקה של שלב הריצה בשלוש רמות רגישות עבור בסיס הנתונים DeepScapula. הגדרת הפרמטרים בהרצות כללה שלוש רמות סף שונות: בחירת 70% מהנתונים עבור training set, קביעת 10 קבוצות בתחילת תהליך הסגמנטציה, מספר קבוצות סופי לא מוגבל, 100 מקרים ב buffer). התרשים ממחיש את ההתכנסות של התהליך. ניתן לראות כי בתחילת התהליך יש מספר רב של עדכונים ובהמשך מספר העדכונים הולך ופוחת. ברמות גבוהות יותר של רגישות, תהליך העדכון הוא מתון יחסית, והעדכונים פחות תכופים, ברמות רגישות נמוכות יותר התהליך פעיל מאוד בהתחלה, אך מאט בהמשך.

Occupancy Detection הציג דוגמה של סדרה עיתית בינונית בגודלה עם שדה מטרה נתון. בסיס הנתונים DeepScapula מהווה דוגמה לבסיס נתונים גדול יחסית עם מספר רב של מאפיינים ועם שדה מטרה נתון. בסיס הנתונים Jena Climate מציג בעיית אשכול של סדרה עיתית ללא שדה מטרה והיקף רשומות גדול יחסית.

הכלים בהם השתמשנו:

המודל נכתב כקוד בסביבת Python<sup>1</sup> על מנת לאפשר אוטומטיזציה של התהליך וניתוחי רגישות. האלגוריתמים לסיווג וניתוח אשכולות הופעלו באמצעות Weka.

(Weka - Browse /Weka-3-7-Windows-X64 at SourceForge.Net, n.d.)

### תוצאות ודיון

#### היתכנות המודל - השוואה בין תוצרי עצי החלטות לבין ניתוח אשכולות

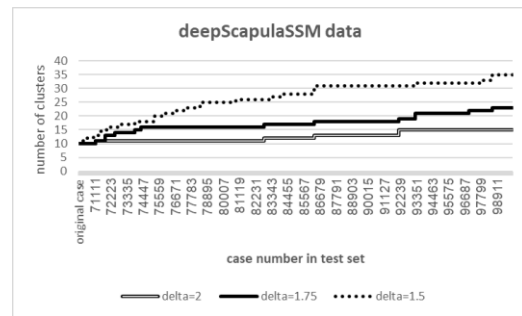
כאמור, תוצאות האשכול הושו לתוצאות עצי החלטה על מנת להעריך את יכולתם של אלגוריתמים של ניתוח אשכולות להשיג תוצאות טובות, למרות שאלגוריתמים אלו אינם משתמשים במידע על שדה המטרה. נבדקו שלושת האלגוריתמים של עצי החלטה הזמינים בתוכנת Weka 3.7.11: J48, Random Forest, REPTree. תוצאות ההשוואה מוצגות בטבלה 3. מדד ההשוואה לאיכות הסיווג בין האלגוריתמים הוא אחוז המקרים שהאלגוריתם סיווג נכון. אחוז הדיוק של עצי החלטה ניתן בפלט עבור כל הרצה של העץ, אחוז הדיוק של אלגוריתם k-means חושב לפי השלבים שהוצגו בפרק "מבחן מקדים להשוואה של תוצרי עצי החלטה וניתוח אשכולות".

<sup>1</sup> המחקר נתמך על ידי מענק MAGNET של רשות החדשנות הישראלית, שגם מימנה את רישום הפטנט.

היעילות של ה-DCU נמדדת במונחים של זמן הריצה. על מנת להראות את היתרונות של הגישה דינאמית אינקרמנטלית, בוצעה השוואה עם זמן הריצה במקרה של מספר סגמנטים קבוע ללא מגבלה על גודל ה buffer (הגישה הסטטית) וזמן הריצה במקרה של מספר סגמנטים בלתי מוגבל עם אפשרות לפיצול, אך עם גודל buffer בלתי מוגבל (הגישה הדינאמית). התוצאות מוצגות בטבלה מספר 5.

### בדיקת היעילות - דיון בתוצאות

ניתן לראות כי זמן הריצה בגישה הסטטית הוא הנמוך ביותר מבין שלושת הגישות (56 שניות). הסיבה לכך ברורה: לא נדרש שום מאמץ חישובי בארגון ועדכון של הסגמנטים, כל מקרה חדש מסווג לאחד הסגמנטים הקיימים, גם אם הוא יהיה רחוק מאוד (במונחים של מדדי מרחק) ממרכז הסגמנט. מאותה סיבה מדדי איכות הסיווג (ממוצע וסטית התקן של RMSE) הם הגרועים ביותר בין שלושת הגישות. אפשר לסכם שהגישה הסטטית מספקת עיבוד מהיר מאוד, אך לא מדויק. הגישה הדינאמית מספקת את מדדי איכות הסיווג הטובים ביותר משלושת הגישות. ההסבר לתופעה זו קשור ליכולת של האלגוריתם להתייעל במהלך תהליך העדכון ולבנות סגמנטים מאוד מייצגים, במיוחד לאור העובדה שקיימת אפשרות לקחת בחשבון את כלל הנתונים בסגמנט מסוים. לעומת זאת, אותו מאפיין שהוביל לאיכות הסיווג הטובה ביותר, הוביל גם לזמן הריצה הגבוה ביותר (154 שניות). הגישה הדינאמית-אינקרמנטלית, המוצגת במחקר הנוכחי, מאפשרת לקצר את זמן הריצה בהשוואה לגישה הדינאמית (138 שניות) ולשפר את מדדי איכות הסיווג בהשוואה לגישה הסטטית. ככל שהגודל של בסיסי הנתונים גדול יותר, כך ההבדלים בין הגישות צפויים להיות בולטים יותר.



תרשים 4 – התכנסות תהליך עדכון הסגמנטים

### התכנסות הסגמנטציה - דיון בתוצאות

למרות השימוש במאגרי נתונים מצומצמים (data buffers), המודל עדיין הצליח לזהות את השינויים בזרם הנתונים ובמיוחד ואת המקרים החדשים שגרמו לשינויים אלה. תהליך העדכון התכנס בכל בסיסי הנתונים, למרות השימוש בנתונים חלקיים בלבד.

ממצא חשוב נוסף מתייחס לזיהוי מקרים חריגים (outliers). תוצאות ההרצות של המודל DCU מראות שככל שרמת הרגישות עולה, המודל מצליח לזהות מקרים שלא ניתן לסווג לאף סגמנט אחר והם יוצרים סגמנטים חדשים. טבלה 4 מציגה תוצאות של הרצות של שלושה בסיסי נתונים עם שלוש רמות של רגישות כל אחד ואת מספר הקבוצות הסופי.

### זיהוי חריגים - דיון בתוצאות

ככל שסף הרגישות עולה ( $\delta$  נמוכה יותר), יותר מקרים צפויים לחרוג מעבר לסף ולגרום לעדכון של סגמנטים. המודל מצליח לזהות סגמנטים "מיוחדים" יותר שלא דומים לסגמנטים הקיימים, אפילו לסגמנטים הומוגניים שנוצרו כתוצאה מהפיצול של סגמנט קיים. זה מגדיל את סך מספר הסגמנטים בסוף ההרצה של המודל. חלק גדול מהם הם סגמנטים חדשים שמוגדרים כ outliers עם מקרה יחיד בתוכם או מספר קטן מאוד של מקרים, החלק השני מצביעים על סגמנטים שמציגים מגמות חדשות כי המודל המשיך לסווג את המקרים החדשים לסגמנטים שנוצרו.

### בדיקת יעילות הגישה האינקרמנטלית

### בהשוואה לסטטית ולדינאמית

| <i>Data</i>                   | <i>Tree algorithm</i> | <i>Tree accuracy</i> | <i>Number of segments</i> | <i>k-means accuracy</i> |
|-------------------------------|-----------------------|----------------------|---------------------------|-------------------------|
| <i>LEV</i>                    | J 48                  | 60.4%                | 65                        | 49.5%                   |
|                               | REPTree               | 60.1%                | 30                        | 52.3%                   |
|                               | Random Forest         | 62.8%                | 90                        | 47.6%                   |
| <i>Occupancy normalized</i>   | J 48                  | 99.2%                | 42                        | 94.44%                  |
|                               | REPTree               | 99.01%               | 40                        | 94.43%                  |
|                               | Random Forest         | 99.2%                | 186                       | 93.53%                  |
| <i>Occupancy standardized</i> | J 48                  | 99.22%               | 42                        | 86.43%                  |
|                               | REPTree               | 99.09%               | 40                        | 86.37%                  |
|                               | Random Forest         | 99.16%               | 183                       | 87.04%                  |
| <i>deepScapula normalized</i> | J 48                  | 88.279%              | 117                       | 72.11%                  |

טבלה 3 – השוואה של תוצרי הסיווג של עצי החלטה וניתוח אשכולות

| <i>Database</i>   | $\delta =$<br><i>threshold</i> | <i>Original</i> | <i>Splits</i> | <i>New segments</i> |                  | <i>Final number of segments</i> |
|---|--------------------------------|-----------------|---------------|---------------------|------------------|---------------------------------|
|   |                                |                 |               | <i>Outliers</i>     | <i>New trend</i> |                                 |
| <i>LEV</i><br><i>Initial number of segments = 10</i>            | 0.6                            | 41              | 217           | 23                  | 19               | 37                              |
|   | 0.7                            | 70              | 230           | --                  | --               | 24                              |
|   | 0.8                            | 138             | 162           | --                  | --               | 16                              |
| <i>Occupancy</i><br><i>Initial number of segments = 10</i>      | 0.85                           | 4007            | 3772          | 11                  | 770              | 22                              |
|   | 0.9                            | 3637            | 4923          | --                  | --               | 20                              |
|   | 1                              | 5232            | 3328          | --                  | --               | 15                              |
| <i>deepScapulaSSM</i><br><i>Initial number of segments = 10</i> | 1.5                            | 9433            | 20398         | 3                   | 3                | 35                              |
|   | 1.75                           | 19612           | 10388         | --                  | --               | 23                              |
|   | 2                              | 24771           | 5229          | --                  | --               | 15                              |

טבלה 4 – תוצאות סגמנטציה

| <i>Evaluation measures</i>      | <i>Static state</i> | <i>Dynamic state:<br/>Non-restricted buffer</i> | <i>Incremental dynamic state:<br/>Buffer size=100</i> |
|---------------------------------|---------------------|---|---|
| <i>Final number of segments</i> | 10                  | 18  | 16  |
| <i>Average RMSE</i>             | 0.5117              | 0.4205  | 0.41  |
| <i>Std. dev. (RMSE)</i>         | 0.2929              | 0.1491  | 0.2042  |
| <i>Running time</i>             | 56 sec              | 154 sec   | 138 sec   |

טבלה 5 – השוואה של שלוש גישות

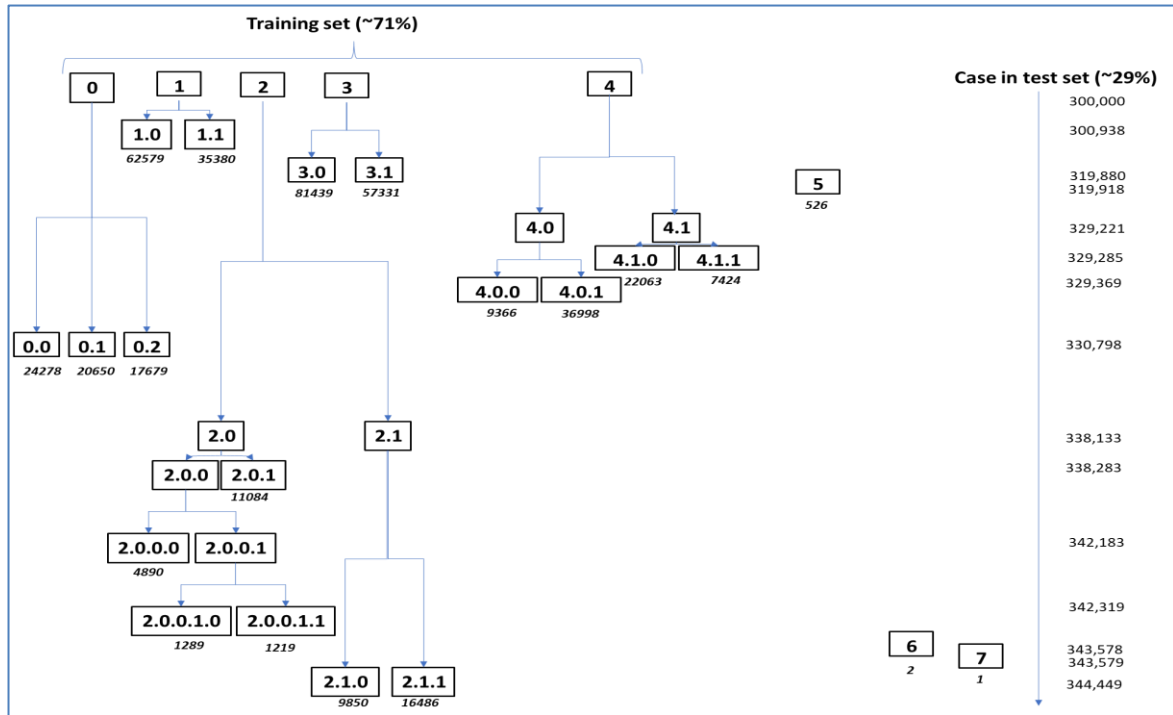
באופן מקיף את תהליך הסגמנטציה כמנגנון לתמיכה בתהליך האינטרפרטציה. זאת בניגוד למטרת הניסויים הקודמים שהתמקדה במשך זמן העיבוד על מנת לאפשר סגמנטציה בזמן-אמת. ניתן לזהות סגמנטים שהתפצלו לתתי-סגמנטים, מגמות חדשות ומקרים חריגים. בבסיס הנתונים Jena Climate נמצאו שני סגמנטים שבלטו עבור כמעט כל רמת רגישות

### ויזואליזציה ExpanDrogram

תרשים 5 מציג את תמונת ה-ExpanDrogram המתקבלת בניחות נתוני הסדרה העיתית Jena Climate ללא שימוש בשדה המטרה. נבחרו מספר פרמטרים עבור DCU, מתוכם סף הרגישות ומספר קבוצות התחלתי. ה- buffer לא הוגבל כיוון שמטרת הניסוי הייתה להציג

לקבל הפרטים בתרשים זה מאפשרים לבדוק באיזו נקודת זמן התגלו המקרים החריגים, מהם ערכי וקטור התכונות שלו לעומת וקטור הממוצעים של כלל המקרים וכו'.

(סגמנטים 6 ו-7 בתרשים 5). במצב כזה המשתמש יכול להיות מעוניין לבדוק את הסגמנטים לעומק ולחפש סיבה אפשרית לקיומם. תרשים 6 מציג layer מסוג "זיהוי מקרים חריגים", שהוא תרשים נוסף שניתן



תרשים 5 – ExpanDrogram שמציג את תהליך הסגמנטציה של בסיס הנתונים Jena Climate

| attribute   | All    | Id 343578 |
|-------------|--------|-----------|
| p(mbar)     | 989.21 | 990.52    |
| T(degC)     | 9.45   | 16.98     |
| Tpot(K)     | 283.49 | 290.93    |
| ...         | ...    | ...       |
| max.wv(m/c) | 3.53   | -9999     |
| wd(deg)     | 174.73 | 289.6     |

2 cases group 6

Id: 343578

13/7/2015 9:00

group 7 18 cases

Id: 343579

Id: 343580

Id: 343581

Id: 343582

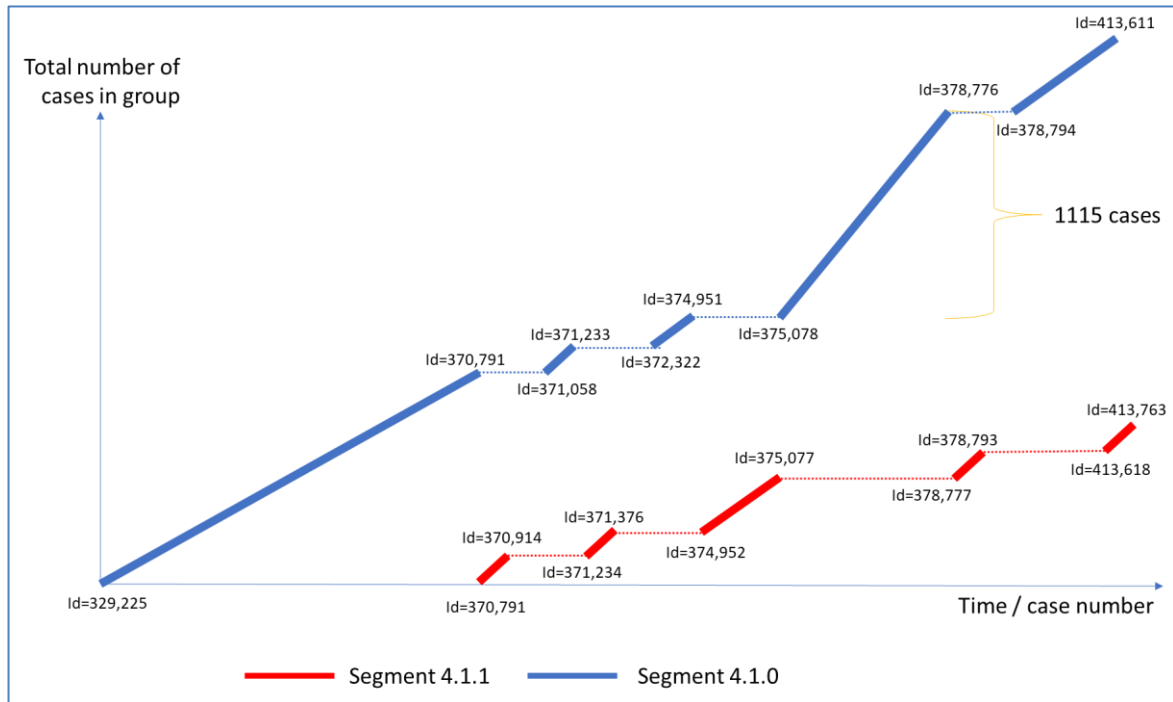
Id: 343583

⋮

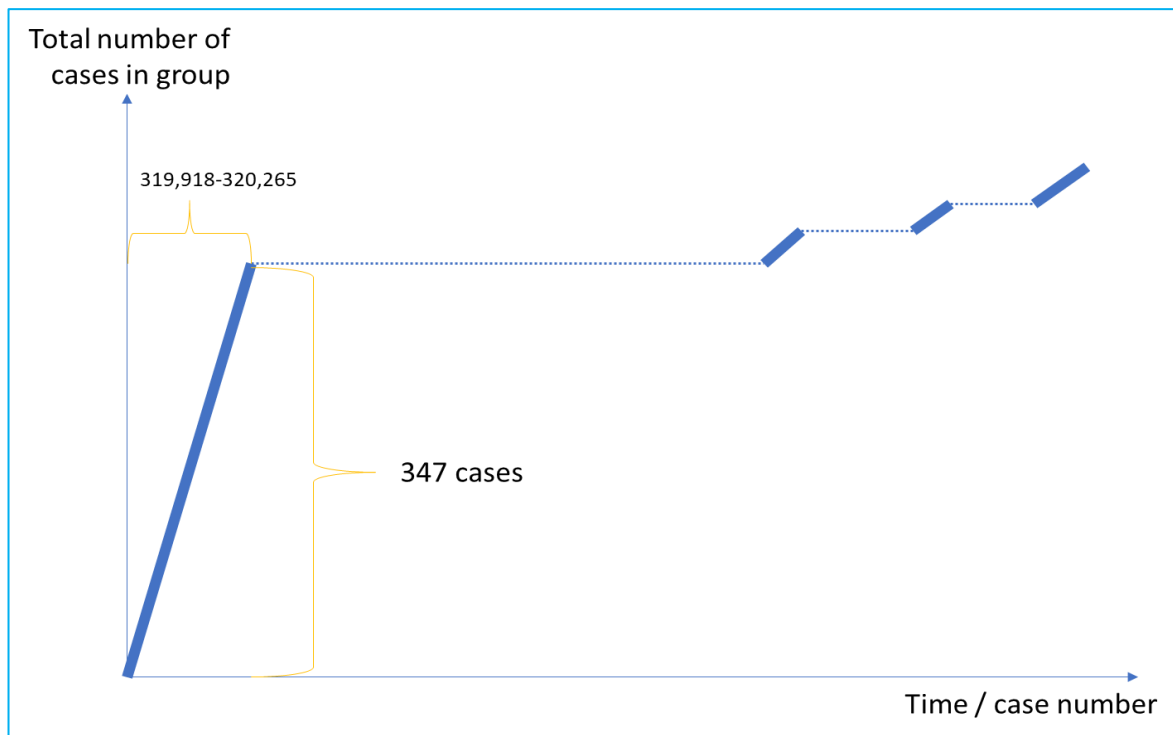
Id: 343596

Id: 343597

תרשים 6 - זיהוי מקרים חריגים – סגמנטים 6,7 שהתקבלו בתרשים 5



תרשים 7 – סגמנטים 4.1.0, 4.1.1 – השוואה של קצב המילוי



תרשים 8 – מילוי של סגמנט 5 – מגמה חדשה שנעצרה

4.1. שינויים בקצב המילוי יכולים להצביע על מגמות מסוימות שניתנות לפיענוח מצד המשתמש. תרשים 8 מציג את תהליך המילוי

התרשימים 7 ו-8 מציגים שתי אפשרויות נוספות לצורך בחינה של היבטים ספציפיים. תרשים 7 מציג את השוואה של תהליך המילוי של סגמנטים 4.1.0 ו-4.1.1 שהתפצלו מסגמנט

ההנחה של המחקר הנוכחי היא שבסביבה של נתוני עתק, אין אפשרות לערוך חישובים על בסיס כל נתוני העבר, במטרה לעדכן ולארגן מחדש את הסגמנטים הקיימים או ליצירת חדשים.

המחקר הנוכחי מציג את ה-DCU, מסווג דינמי אינקרמנטלי המסוגל להתמודד עם סביבות נתונים דינמיות. החידוש העיקרי של ה-DCU הוא שימוש במאגרי זיכרון קטנים buffers לצורך אחסון מקרים ייצוגיים בכל סגמנט קיים, והתבססות עליהם לצורך ביצוע החישובים בכל נקודת זמן. עם הופעת מקרה חדש ששונה באופן משמעותי מהקודמים, המסווג מעדכן רק את ה- buffer הרלוונטי. ניתוחי רגישות רבים בוצעו על מנת להעריך את המודל. להלן סיכום הממצאים והמסקנות העולות מהם:

(1) השוואה בין איכות הסיווג של עצי החלטה לבין איכות הסיווג של k-means הראתה אנלוגיה בין אשכול לבין התיב בעץ החלטה והוכיחה שניתוח אשכולות מסוגל לספק סיווג מספיק טוב בהיעדר שדה מטרה ידוע.

(2) תהליך הריצה הראה התכנסות בכל בסיסי הנתונים שנבדקו. עובדה זו מצביעה על כך שהשימוש במאגרי נתונים קטנים אינו מונע איתור של סגמנטים הומוגניים.

(3) כיוול של פרמטרים שונים, במיוחד רמת הרגישות, מאפשר לגלות מגמות חדשות ומקרים חריגים.

(4) הגישה הדינאמית-אינקרמנטלית מדגימה את היתרון המשמעותי בהשוואה לגישה סטטית (בעיקר בהקשר של דיוק) ודינמית (בהקשר של חיסכון בזמן העיבוד).

(5) המודל פועל בצורה מדויקת גם ללא שלב אתחול ולמידה (training). בדרך זו ניתן לראות שהגישה האינקרמנטלית משיגה גם גמישות וגם יעילות חישובית מרבית.

של סגמנט 5 שצבר מספר גדול של מקרים בהתחלה, אך התהליך כמעט נעצר בהמשך.

### ויזואליזציה - דיון בתוצאות

ה-ExpanDrogram הוא תרשים דמוי עץ, אך בניגוד ל- Dendrogram הוא מאוד קריא עקב המיקוד במספר מובחן של פרטים. כדי להתאים את מיקוד הפרטים ורזולוצית הסיווג המשתמש יכול לשנות את ערכי הפרמטרים שהוגדרו ב-DCU ולקבל תמונה יותר או פחות מפורטת. אחת המטרות העיקריות של ה-ExpanDrogram היא להנגיש את הניתוח הוויזואלי של תהליך הסגמנטציה למשתמש. באופן זה המשתמש יכול לקבל את במידע במבט חטוף אחד וכן להחליט איזה חלק בתרשים או איזו תכונה חשובה יותר ולהשתמש באפשרות המתאימה כדי לקבל תמונה הממוקדת יותר בפרטים ספציפיים אלו.

תרשים 6 מציג מקרה ספציפי בו המשתמש יכול לגלות, באופן פשוט, שסגמנטים 6 ו-7 נוצרו כתוצאה מטעות בקליטת הנתונים: חלק מהערכים נרשמו כקודים ולא כערכים אמיתיים של המשתתפים (9999- אינו יכול להיות ערך קיים). גם בדיקת ערכים כזאת שמגלה טעות ולא תופעה מעניינת חדשה, יכולה לשמש את המשתמש בתהליך האינטרפרטציה.

### סיכום ומסקנות

בסביבה דינמית ועתירת נתונים, תהליך הסגמנטציה של זרם האירועים החדשים לקבוצות הומוגניות הוא אתגר משמעותי. זרם הנתונים אינו סטטי, ישנם שינויים דינמיים במאפייני הנתונים, מגמות ישנות נעלמות ומגמות חדשות מופיעות, וסגמנטים קיימים יכולים להתפצל או להתמזג. מודלים הבנויים על סמך נתוני העבר מאפשרים סגמנטציה מהירה כל עוד המקרים החדשים "דומים" מספיק לקיימים. במקרים בהם מתקבל נתון חדש ששונה באופן מהותי מהסגמנטים הקיימים, נדרש לארגן מחדש את הסגמנטים. תהליך זה מצריך משאבים אנליטיים רבים.

הבודד, לעקוב אחר "בקרת הגרסאות", לזהות מגמות חדשות ומקרים חריגים. הרזולוציה מותאמת באמצעות מגוון פרמטרים שניתן לשנות אותה בקלות. שכבות התצוגה השונות ויכולת ה-zoom-in מאפשרות למשתמש לזהות ולעקוב אחר מגוון רחב של פרטים ומצבים.

לבסוף, כדי לשפר את תהליך הפרשנות וקבלת ההחלטות הנדרשים במצבים בהם אין שדה מטרה, פותחה ויזואליזציה המספקת תמונה מקיפה ובו בעת ממוקדת, ה-ExpanDrogram. הוויזואליזציה מאפשרת למשתמש לראות בו זמנית גם את רמת הסגמנט וגם את רמת הפרט



**רשימה ביבליוגרפית**

- Barak, A., & Gelbard, R. (2011). Classification by clustering decision tree-like classifier based on adjusted clusters. *Expert Systems with Applications*, 38(7), 8220–8228. <https://doi.org/10.1016/j.eswa.2011.01.001>
- Ben-David, A. (1992). Automatic Generation of Symbolic Multiattribute Ordinal Knowledge—Based DSS's: Methodology and Applications. *Decision Sciences*, 1357–1372.
- Deza M.M. & Deza E. (2014). *Encyclopedia of Distances* (3rd ed.). Springer.
- Fan, J., Han, F., & Han, L. (2014). Challenges of Big Data analysis. *National Science Review*, 1, 293–314.
- Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. *Future Generation Computer Systems*, 13(2), 99–115. [https://doi.org/10.1016/S0167-739X\(97\)00015-0](https://doi.org/10.1016/S0167-739X(97)00015-0)
- Gelbard, R., & Khalemsky, A. (2018). Dynamic Classifier and Sensor Using Small Memory Buffers. In P. Perner (Ed.), *Advances in Data Mining. Applications and Theoretical Aspects* (pp. 173–182). Springer International Publishing.
- Gelbard, Roy, Goldman, O., & Spiegel, I. (2007). Investigating diversity of clustering methods: An empirical comparison. *Data and Knowledge Engineering*, 63, 155–166.
- Guha, S., & Mishra, N. (2016). Clustering Data Streams. In *Data Stream Management* (pp. 169–187). Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-540-28608-0\\_8](https://doi.org/10.1007/978-3-540-28608-0_8)
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- Kaggle: Your Home for Data Science*. (n.d.). Retrieved February 25, 2019, from <https://www.kaggle.com/>
- Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 1–8. <https://doi.org/10.1109/2945.981847>
- Keim, D., Qu, H., & Ma, K. (2013). Big-Data Visualization. *IEEE Computer Graphics and Applications*, 33(4), 20–21. <https://doi.org/10.1109/MCG.2013.54>

- Khalemsky, A., & Gelbard, R. (2021). ExpanDrogram: Dynamic Visualization of Big Data Segmentation over Time. *Journal of Data and Information Quality*, 13(2), 11:1-11:27. <https://doi.org/10.1145/3434778>
- Khalemsky, Anna, & Gelbard, R. (2019). A dynamic classification unit for online segmentation of big data via small data buffers. *Decision Support Systems*, 113157. <https://doi.org/10.1016/j.dss.2019.113157>
- Milligan, G. W., & Hirtle, S. C. (2012). Clustering and Classification Methods. In *Handbook of Psychology, Second Edition*. American Cancer Society. <https://doi.org/10.1002/9781118133880.hop202007>
- Park, S. C., Piramuthu, S., & Shaw, M. J. (2001). Dynamic rule refinement in knowledge-based data mining systems. *Decision Support Systems*, 31(2), 205–222. [https://doi.org/10.1016/S0167-9236\(00\)00132-9](https://doi.org/10.1016/S0167-9236(00)00132-9)
- Shah Siddharth, Chauhan N.C., & Bhandery S.D. (2012). Incremental Mining of Association Rules: A Survey. *International Journal of Computer Science and Information Technologies*, 3(3), 4041–4074.
- Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th ed.). Pearson.
- Solt, Y., & Horovitz, S. (2012). *Buffer management architecture* (Patent No. US8176291 B1). <http://www.google.com/patents/US8176291>
- UCI Machine Learning Repository: Data Sets*. (n.d.). Retrieved December 10, 2016, from <https://archive.ics.uci.edu/ml/datasets.html>
- Weka—Browse /weka-3-7-windows-x64 at SourceForge.net*. (n.d.). Retrieved October 15, 2017, from <https://sourceforge.net/projects/weka/files/weka-3-7-windows-x64/>
- Yosipof, A., Khalemsky, A., Gelbard, R., & Senderowitz, H. (n.d.). Dynamic Classification for Materials-Informatics: Mining the Solar Cell Space. *Molecular Informatics*, n/a(n/a). <https://doi.org/10.1002/minf.202000173>