בית הספר
למנהל עסקים
הפקולטה למדעי החברה
אוניברסיטת בר־אילן

תאוריה ופרקטיקה בניהול
כתב עת בהוצאת בית הספר למנהל עסקים
יוני 2026, סיון התשפ״ו | גיליון מס׳ 11

כל מה שצריך לדעת על ניתוח אשכולות ומעולם לא שאלת

רועי גלברד
בית הספר למנהל עסקים
אוניברסיטת בר-אילן

תקציר

ניתוח אשכולות אלו טכניקות שעוסקות בארגון הנתונים בקבוצות בעלות היגיון מובחן. היגיון מסדר זה הוא אחד מאופני ההבנה והלמידה הבסיסיים במגוון רחב של תחומים כגון סיווג סימפטומים רפואיים או פילוח שוק הצרכנים. כיום, עם ההתפתחות הזמינות והידידותיות של כלים לניתוח נתונים, משתמש עם אוריינות טכנית בסיסית יכול בקלות להפעיל כלים של סיווג וניתוח אשכולות. באופן זה המשתמש לא זקוק עוד לידע אודות תהליך ומנגנוני העיבוד, אלא רק להבנה הפרשנית לגבי האופן בו יש לקרוא ולפרש את התוצאות שהתקבלו. הדבר מאוד יעיל ואמין כאשר אנחנו מפעילים טכניקות ״יציבות״. מנגד, הדבר מסוכן כאשר מדובר בטכניקות ״לא יציבות״ כדוגמת ניתוח אשכולות, שמפיקות תוצאות שונות עם כל שינוי בערכים של משתני ההפעלה. לכן, ראוי להסב את תשומת ליבו של המשתמש לרכיבים המשפיעים על החלוקה לקבוצות ועל קביעת ההיגיון המסדר. המאמר הנוכחי עומד על ליבת המשתנים המשפיעים על התהליך ומצביע על הנקודות בהן המשתמש חייב להיות בעל ידע והבנה מעמיקים, כדי שיוכל לבקר ולאמת את התוצר המתקבל.

מילות מפתח - ניתוח אשכולות, סיווג, ניתוח גורמים, מדדי דימיון, כריית נתונים.

מבוא

בשנת 2010 פורסם המאמר ״50 Years Beyond K-means״ לציון חמישים שנים לאלגוריתם K-means (Jain, 2010). שם האלגוריתם, K-means, הפך עם השנים לשם נרדם לתחום כולו, תחום ניתוח האשכולות (Cluster Analysis), שעניינו **ארגון נתונים בקבוצות בעלות היגיון מובחן**. טכניקות מעטות זוכות לחיי מדף ארוכים שכאלה, במיוחד כשמדובר בטכניקות שמשולבות בטכנולוגיות ובעיבודים החדשים ביותר. כמו תגליות וחידושים מדעיים נוספים, גם האלגוריתמיקה של K-means התפתחה במקביל בתחומי מדע שונים, כשחלק מהתגליות פורצות וזוהרות וחלקן נשארות בצל, כמו דרווין וואלאס, ניוטון ולייבניץ, אדיסון וטסלה; כמו יעקוב ועשיו במרוץ אחר הבכורה. באותו אופן גם ל-K-means היסטוריה עשירה שהחלה בשנת 1956 בפרסום של סטטיסטיקאי צרפתי בשם הוגו שטיינהאוס (Steinhouse, 1956), ובדוח טכני של סטיוארט לויד, ממעבדות בל, שנכתב בשנת 1957 אך פורסם רק 25 שנה לאחר מכן (Lloyd, 1982). בהמשך, בשנת 1962, שני מתודולוגים מתחום הביולוגיה, סניס וסוקל, מפרסמים את המאמר Numerical Taxonomy (Sneath & Sokal, 1962), ושנה לאחר מכן מפרסמים ספר אודות עקרונות הטקסונומיה הנומרית, שמהווה ריכוז סדור של מגוון הטכניקות בתחום (Sokal & Sneath, 1963). השם הרשמי K-means מופיע לראשונה רק בשנת 1967 במאמר של ג׳ימס מקווין (MacQueen, 1967), וקצת קודם לכן, בשנת 1965, אדוארד פורגי מפרסם טכניקה זהה בכתב עת בתחום הביולוגיה (Forgy, 1965), ואותה טכניקה מוצגת במקביל גם על ידי צמד חוקרים מסטנפורד, במסגרת דוח לקרן מחקר של ה-Naval Research (Ball & Hall, 1965).

ארגון נתונים בקבוצות בעלות היגיון מובחן הוא אחד מאופני ההבנה והלמידה הבסיסיים במגוון רחב של תחומים. הטקסונומיה הביולוגית מסווגת את האורגניזמים למסגרות חיים שונות, אנשי שיווק מסווגים את הצרכנים לסגמנטים שונים, רופאים מסווגים סימפטומים למחלות, אנשי מקצוע טכניים מסווגים סימפטומים לתקלות, ובאותו אופן כל תחום

והיישומים שמשמשים אותו. הקטגוריות השונות של הקבוצות בעלות ההיגיון המובחן נוצרות בתהליך המכונה בשם ניתוח אשכולות לא מונחה (Unsupervised cluster analysis). תהליך בו אין קביעה מראש לגבי מספר הקבוצות אליהן תחולק האוכלוסייה, אין הגבלה לגבי מספר החברים בכל קבוצה, וכמובן שאין תיוג של הפרטים. תהליך בו מסווגים פרטים על בסיס תיוג מוקדם נקרא סיווג (Classification), אך בניתוח אשכולות אין כל תיוג ואין כל מידע לגבי סיווג אפשרי של הפרטים. תהליך הסיווג הוא נדבך שנבנה על בסיס ניתוח האשכולות, לאור ההיגיון המבחין שנמצא בנתונים. היגיון שמתורגם לתכונות ולערכים בולטים, שבאמצעותם יופעלו תהליכי הסיווג (Classification).

ההיגיון המסדר, באמצעותו אנחנו מבדלים בין הקבוצות, מתבסס על יחסי דימיון-מרחק שאנחנו מעריכים כשאנו משווים בין הפרטים שבאוכלוסייה. הדימיון יכול להתייחס לדימיון בין שני פרטים, בין פרט וקבוצה, בין שתי קבוצות, כאשר המטרה לאורה התהליך מתבצע היא להגיע למצב של **מרחק מינימלי בין הפרטים הנכללים בכל אחת מהקבוצות, ומרחק מקסימלי בין הקבוצות עצמן**. הסיבוכיות המתמטית של התהליך הזה היא גבוהה מאוד ולכן חוקרים ניסו למצוא טכניקות שמעבר ליכולתן לייצר קבוצות בעלות היגיון מובחן, יהיו גם בעלות יעילות מתמטית גבוהה, ו-K-means היא הטכניקה המובילה בתחום זה כבר קרוב לשבעים שנים. יחד עם זאת, המשתמש חייב להבין שהיעילות המתמטית מושגת באמצעות הנחות ש- K-means מניח לגבי העיבוד, ובכלל זה הגרלה של מוקדים בהם מתחיל תהליך העיבוד, ועצירת העיבוד ברגע שמשיגים אופטימום מקומי. בניגוד ל--K means, אלגוריתמים היררכיים, כדוגמת Ward, מבצעים עיבוד "שלם" במסגרתו מחושבת טבלת מרחקים בין כל זוג פרטים, ומכאן שסיבוכיותם היא O^2 ומגבילה את השימוש בהם להיקף מצומצם של נתונים, שמותנה בפלטפורמה בה מבוצע העיבוד. לפיכך, למרות ש- K-means זו הטכניקה המובילה בזכות היעילות המתמטית שלה, הרי שבמקרים בהם היקף הנתונים מצומצם יחסים מומלץ לבחון במקביל גם אלגוריתמים אחרים, ובכלל זה אלגוריתמים היררכיים.

כיום, כל משתמש עם אוריינות טכנית בסיסית, יכול בלחיצת כפתור קלה להפעיל כלים של סיווג וניתוח אשכולות, בדיוק כפי שהתרגלנו בניתוחים סטטיסטיים כדוגמת רגרסיות ומתאמים, אותם ניתן לבצע במגוון תוכנות ובכלל זה גיליון אלקטרוני סטנדרטי. באופן זה המשתמש לא זקוק עוד לידע אודות התהליך ומנגנוני העיבוד, אלא רק להבנה הפרשנית לגבי האופן בו יש לקרוא ולפרש את התוצאות שהתקבלו. הדבר מאוד יעיל ואמין כאשר אנחנו מפעילים טכניקות "יציבות" שנותנות תוצאות אחידות בכל תנאי הפעלה (כגון, בדיקת מתאמים בין משתנים). מנגד, הדבר מסוכן כאשר מדובר בטכניקות "לא יציבות" כדוגמת ניתוח אשכולות, שמפיקות תוצאות שונות עם כל שינוי בערכים של משתני ההפעלה, כגון: מדד הדימיון בין פרטים, הדימיון בין קבוצות, מדד הערכת ההקבצה, אלגוריתם ההקבצה, וכדי. לכן, ראוי להסב את תשומת ליבו של המשתמש לרכיבים שמשפיעים על התהליך, ולהצביע על הנקודות בהן המשתמש חייב להיות בעל ידע והבנה מעמיקים, כדי שיוכל לבקר את התוצר המתקבל ולצלוח בשלום את הקרקע "הבוצית" של ניתוח האשכולות.

המאמר הנוכחי מציג דיון מתודולוגי בנקודות החולשה שכל משתמש, כל שכן מורה בתחום, צריך להיות מודע להן בזמן ההוראה והשימוש בכלים של ניתוח אשכולות ו/או ניתוח גורמים (Factor Analysis). ניתוח גורמים הוא תהליך זהה לחלוטין לניתוח אשכולות, אלא שהוא מופעל על המשתנים, להבדיל מניתוח אשכולות שמופעל על הרשומות (כפי שמוסבר בפרק ג'), ולפיכך השאלות שהמאמר מציג ביחס לניתוח אשכולות, רלוונטיות לתהליך של ניתוח גורמים. ספרות המחקר בתחום החינוך עושה שימוש גדול בטכניקות אלה לצורך הצגה ופירוש של ממצאים. לדוגמה, בכתב העת Computers & Education התפרסמו בשנתיים וחצי האחרונות, משנת 2023 ועד אמצע 2025, מאה ושמונה (108) מאמרים שעושים שימוש בניתוח אשכולות (כגון: Roski et al., 2024; Alvarez-Garc et al., 2024; Huang et al., 2025). יחד עם זאת, פרסומים מועטים מאוד עוסקים בהיבט המתודי של השימוש בטכניקות אלה, וגם הם לא באופן ישיר כי אם כנלווים להקשר אחר, כגון לניתוח נתוני עתק (Stojanov & Daniel, 2024), ללימוד של לוחות מחוונים אנליטיים, Analytics Dashboards (Paulsen & Lindsay, 2024), או להסבר תהליכים של כריית נתונים ומודלים של בינה מלאכותית

(Antonenko, 2012; Türkmen, 2025). לפיכך, המאמר הנוכחי ממקד את הדיון בהיבטים המתודיים של ניתוח האשכולות עצמו ובכלל זה, דיון במדדי דימיון-מרחק, דיון בייצוג המשתנים, בהערכת חשיבות ובולטות של משתנה, מימד הזמן והיבטים דינאמיים בניתוח אשכולות, ומדדים להערכת איכות ניתוח האשכולות.

א. הערכת איכות ניתוח האשכולות

ניתוח אשכולות מוגדר כתהליך לא מונחה, Unsupervised Cluster Analysis, שמספר הקבוצות אליהן תחולק האוכלוסייה לא מוכתב בו מראש, אך האלגוריתמים מחייבים שבכל הרצה המשתמש יגדיר את מספר הקבוצות לחלוקה. לפיכך, הניתוח מתבצע במספר סבבים-איטרציות, כשבכל סבב מחלקים את האוכלוסייה למספר הולך וגדל של קבוצות. לאחר מכן מפעילים מדדי הערכה שונים על סדרת התוצאות, והחלוקה שמקבלת את הציון הגבוה ביותר, על פי מדד ההערכה שיוגדר, היא זו שתיבחר ותעבור לשלב הבא של פרוש התוצאה – **Interpretation of results**. פירוש התוצאות הוא שלב ייחודי לניתוח אשכולות שאינו נדרש בבעיות סיווג. בבעיות של סיווג, כגון בבעיה של חיזוי נטישת לקוחות של חברת כבלים, ברור לנו שהכוונה היא לחלק את האוכלוסייה לשתי קבוצות: קבוצת הלקוחות הנאמנים וקבוצת הלקוחות הנוטשים, ולאתר בתוך כל קבוצה פרופילים של לקוחות. לדוגמה, בקרב הנוטשים נוכל למצוא פרופילים כגון: לקוח הנוטש בגלל מחיר, לקוח הנוטש בגלל מגוון התכנים, ולקוח שנוטש בגלל איכות השירות. אם בדוגמה הזו היינו מריצים מסווג של עץ החלטה, אזי באופן פשטני, היינו מצפים לקבל עץ עם ארבעה עלים: 1) עלה של הלקוחות הנאמנים, 2) עלה של הנוטשים בגלל מחיר, 3) הנוטשים בגלל מגוון תכנים, 4) הנוטשים בגלל איכות שירות. כלומר, למרות שמדובר בבעיה שמסווגת את האוכלוסייה לשתי קבוצות ברורות, של לקוח נאמן/נוטש, הרי שאנו מצפים לקבל מספר גדול יותר של עלים, כאשר כל עלה מייצג פרופיל ספציפי של נאמנות/נטישה. באותו אופן הדבר נכון לגבי ניתוח אשכולות. גם אם נרצה לחלק את האוכלוסייה לשני אשכולות, כגון שני סגמנטים שיווקיים, הרי שנרצה להבין את הגוונים השונים שיכולים להיות בכל סגמנט. לפיכך, **יש לראות אשכול באופן אנאלוגי לזה של עלה בעץ החלטה**, ובשלב הפרשנות ננסה להבין את הפרופיל שבא ליידי ביטוי בכל אשכול. אך כאמור, את שלב החקירה הפרשנית נעשה לאחר שנקבל חלוקה כזו שמדדי ההערכה יצביעו עליה שהיא הטובה ביותר מבין החלוקות האפשריות.

יש מגוון אלגוריתמים ומגוון מדדים שניתן להשתמש בהם לצורך החלוקה לקבוצות ולצורך הערכת איכות החלוקה שהתקבלה. מדדי ההערכה מנסים לאמוד עד כמה "אופטימלית" החלוקה שהתקבלה, כשה-"אופטימום" אמור לבטא מצב בו **המרחק בין החברים בתוך כל קבוצה הוא מינימלי** (WGD - Within Group Distance), **והמרחק בין הקבוצות הוא מקסימלי** (BGD - Between Groups Distance). הבעיה היא שיש המון דרכים לחישוב של דימיון/מרחק, וכל דרך צפויה לתת תוצאה אחרת כפי שמוסבר בפרק הבא - *"מדדי דימיון"*.

באופן טבעי, המרחק בין החברים בתוך כל קבוצה (WGD) הולך וקטן ככל שיש יותר קבוצות. אם באוכלוסייה מסוימת יש X פרטים, ונחלק אותם ל-X קבוצות, אזי בכל קבוצה יהיה פרט אחד בלבד וערך ה-WGD יהיה אפס, שזו התוצאה הטובה ביותר שהוא יכול לקבל. אבל במקביל, ככל שיש יותר קבוצות, המרחק בין הקבוצות (BGD) הולך וגדל - והמדדים השונים מחפשים נקודת "אופטימום" שמייצגת שילוב מיטבי של שני הממדים האלו, שבאופן טבעי נמצאים בניגוד זה עם זה.

כאמור, הכלים השונים מספקים מגוון מדדים לצורך הערכת איכות החלוקה המתקבלת, ועל המשתמש להבין מי הוא המדד המתאים ביותר לבעיה אותה הוא בוחן. הצורך בהבנה עמוקה של מגוון המדדים ממחיש את "חוסר היציבות" של ניתוח האשכולות, ואדגים זאת בתוצאות ביניים שקיבלנו בעבודה שעסקה בשימור לקוחות. לצורך ביצוע ניתוח האשכולות השתמשנו בחבילת R-NbClust, חבילה שהכילה בתוכה שלושים מדדים להערכת איכות החלוקה המתקבלת. **טבלה 1** מציגה חלוקה של האוכלוסייה למספר הולך וגדל של אשכולות, החל מחלוקה לשני אשכולות (בצד שמאל של הטבלה) ועד חלוקה ל-15 אשכולות (בצד ימין של הטבלה), כפי שכתוב בשורה העליונה. מתחת לשורת הכותרת, שורה אחת מתייחסת לניתוחי אשכולות שהשתמשו באלגוריתם בשם Ward, ובשורה מתחתיה התייחסות לניתוחי אשכולות שהשתמשו באלגוריתם K-Means. בכל אחת מההרצות האלגוריתמים התבקשו, כאמור,

לחלק את האוכלוסייה למספר הולך וגדל של קבוצות (החל מ-2 ועד 15), ובתוך כל תא רשום מספר שמייצג את מספר המדדים (מתוך 30 המדדים שהיו בחבילת R-NbClust) שציינו שהחלוקה שהתקבלה זו החלוקה הטובה ביותר האפשרית. לדוגמה, 4 מדדים (מתוך ה-30) הצביעו על כך שהחלוקה של Ward לשלושה אשכולות היא החלוקה הטובה ביותר, ו-6 מדדים (מתוך ה-30) הצביעו על כך שהחלוקה של K-means לשלושה אשכולות היא החלוקה הטובה ביותר.

Clusters / Algorithm	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Measures pointing optimal result
WARD	2	4	4	0	1	1	0	2	1	0	0	1	4	1	21
KMEANS	3	6	2	0	1	1	0	0	1	1	0	2	3	2	22

טבלה 1. מספר המדדים שציינו שהחלוקה שהתקבלה היא הטובה ביותר

המשמעות העולה מהטבלה המוצגת היא שניתן יהיה לשכנע משתמש ולקוח נאיביים, שכמעט כל חלוקה שתתקבל היא החלוקה הטובה ביותר. בדוגמה הספציפית לא נוכל להגיד שחלוקה ל-5 אשכולות, 8 אשכולות ו-12 אשכולות הן חלוקות טובות, אך לגבי כל שאר האפשרויות, ניתן יהיה ל-"הוכיח" על ידי מדד כלשהו שהן הטובות ביותר. יותר מכך, יש אלגוריתמים רבים לניתוח אשכולות, לא רק Ward ו-K-means, כך שבהחלט ייתכנו אלגוריתמים שהחלוקה שלהם ל-5, 8 או 12 אשכולות כן יקבלו את הערכת האיכות הגבוהה ביותר על ידי מדד כלשהו. כפי שהחלוקה של Ward ל-11 אשכולות לא נתמכה על ידי אף מדד ההערכה, ואילו החלוקה ל-11 אשכולות של K-means נתמכת על ידי מדד אחד. ובכיוון ההפוך, החלוקה של K-means ל-9 אשכולות לא נתמכה על ידי אף מדד, ואילו החלוקה ל-9 של Ward נתמכת על ידי שני מדדים. מעבר לכך, שינוי פרמטרים בכל אחד מהאלגוריתמים יניב גם כן תוצאות שונות ובעקבות זאת גם ערכים שונים של מדדי ההערכה, מה שממחיש לנו את חוסר העקביות וחוסר היציבות של תוצאות ניתוח האשכולות המתקבל.

בשנת 1954 יצא הספר How to lie with statistics (Huff, 1954), והנה כעת, ניתוח האשכולות פותח בפנינו "אופקים" רחבים חדשים ל-"שקרים" נוספים. הדוגמה שהוצגה ממחישה עד כמה חשוב שהמשתמש יבין את "קו המחשבה" של **האלגוריתם** בו בחר להשתמש, וכן את "קו המחשבה" על פיו מחושב **מדד ההערכה** באמצעותו בחר להעריך את איכות החלוקה לאשכולות שהתקבלה. כמו כן, מומלץ למשתמש לנתח באופן עצמאי את השינוי בערכי WCD ו-BCD, המהווים את הבסיסי להערכת איכות החלוקה, במהלך החלוקות השונות. מקובל לחשב הן את השינוי והן את נגזרת השינוי כדי לוודא שהחלוקה מייצגת נקודת אופטימום מקומי, כך שכל תזוזה ממנה מוליכה לתוצאה שהיא פחות טובה בהתייחס לשילוב של WGD ו- BGD.

ב. מדדי דימיון-מרחק

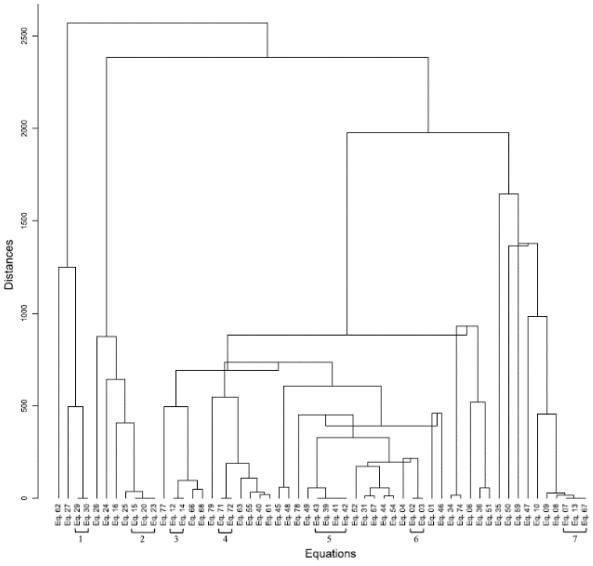
כל המערכות העוסקות בתהליכים של חלוקה לקבוצות ובתהליכים של סיווג, מתבססות על פעולות חישוב של מרחקים ודימיון. אם נייצג את טווח הערכים של הדימיון על סקאלה בין 0 ל-1, כאשר 1 מייצג זהות מוחלטת, ו-0 מייצג שונות מוחלטת. אזי המרחק יהיה על סקאלה הפוכה בה 1 מייצג מרחק מוחלט ו-0 מייצג מיקום זהה. כך שדימיון ומרחק מציגים את אותה תמונה, רק מזווית מבט הפוכה.

מדד דימיון הוא כל מערך חוקים שמקיים את ארבעת התנאים הבאים: א). המרחק בין A ו-B הוא ערך בין 0 ו-N (את ערכו של N ניתן לנרמל ולעבור כך לטווח של 0-1). ב). אם A זהה ל-B אז המרחק ביניהם הוא 0. ג). המרחק בין A ו-B זהה למרחק בין B ו-A (סימטריה שבתחומים רבים, כגון זמן טיסה, לא מתקיים). ד). אם C היא נקודה כלשהי בין A ו-B, אז המרחק בין A ו-B קטן או שווה לסכום המרחקים מ-A ל-C ומ-C ל-B.

כאמור, כל מערך חוקים שמקיים את ארבעת התנאים הנ"ל תקף לשמש כמדד דימיון, ומכאן ברור שניתן לייצר מגוון רחב מאוד של מדדי דימיון. את עושר המדדים ניתן לראות בצורה מרוכזת במאמרי סקירה המציגים גם מגוון של מדדי דימיון וגם את יחסי הדימיון בין מדדי הדימיון השונים. סקירה אחת מציגה ארבעים וחמישה (45) מדדי דימיון למשתנים מספריים (Cha, 2007), וסקירה אחרת מציגה שבעים ותשעה (79) מדדי דימיון שונים המיועדים רק עבור משתנים בינאריים, המיוצגים כווקטור של ביטים, שכל אחד מהביטים יכול לקבל ערך של "0" או של "1" (Wijaya, 2016). על מאה עשרים וארבעה (124) המדדים האלה ניתן בקלות להוסיף מדדים נוספים, הן

למשתנים בינאריים והן למשתנים רציפים או קטגוריאליים.

מדדי דימיון שונים יכולים להפיק הערכות דומות אבל גם הערכות שונות לחלוטין. מאמרי הסיכום מציגים תרשימי דנדרוגרמה הממחישים באופן ויזואלי את הדימיון והמרחק בין המדדים. **איור1** מציג את הדימיון בין שבעים ותשעה (79) המדדים עבור משתנים בינאריים המוצגים במאמר של ויז'איה (Wijaya, 2016). ניתן לראות שם כי שלושת המדדים בצד ימין, המיוצגים על ידי הספרור שניתן למשוואות החישוב שלהם (Eq.07, Eq.13, Eq.67), כמעט זהים זה לזה, אבל שלושתם מאוד רחוקים ממדדים Eq.15 , 20 Eq. ו-Eq.23 , והכי רחוקים ממדד Eq.62 שבפינה השמאלית של התרשים.



איור 1. דנדרוגרמה המציגה את הדימיון-מרחק בין 79 מדדים למשתנים בינאריים (מתוך Wijaya - 2016)

נמחיש את הדימיון והשוני בערכים המתקבלים ממדדי הערכה שונים על ידי דוגמא המשתמשת בייצוג בינארי בה כל פרט מיוצג כווקטור של אפסים ואחדים. **טבלה 2** מציגה שבעה זוגות של וקטורים בינאריים. העמודה הימנית מציגה את ערך הדימיון בין כל זוג וקטורים על פי מדד **דימיון פוזיטיבי** (Positive Atom Distance - PAD), שעושה את החישוב אך ורק על בסיס ביטים החיוביים (הביטים הדולקים שערכם "1"), כדוגמת מדד Dice. והעמודה משמאלה מציגה את ערך הדימיון על פי מדד שעושה את החישוב על בסיס כל הביטים בכל וקטור, כדוגמת מדד Hamming.

החישוב של מדד Hamming הוא הפעלה של האופרטור XOR (Exclusive OR) על שני הווקטורים, וחלוקה באורך הווקטור. נוסחת החישוב של מדד Dice היא: 2Nab /(Na+Nb) כאשר: Na מייצג את מספר הביטים הדולקים (שערכם "1") בווקטור a, Nb מייצג את מספר הביטים הדולקים בווקטור b, ו- Nab מייצג את מספר הביטים המשותפים שדולקים בשני הווקטורים.

צורות החישוב השונות מוליכות לקבלת ערכי דימיון שונים לחלוטין. לדוגמה, בשורה הראשונה (בטבלה 2) הדימיון על פי Dice הוא אפס מוחלט, ואילו על פי Hamming הדימיון גבוה מאוד ועומד על שיעור של 83.3%. לעומת זאת, בשורה החמישית ערך הדימיון בשתי השיטות קרוב מאוד ועומד על שיעור של כ-80%, ובשורה השישית בשתי השיטות ערך הדימיון נמוך מאוד.

#	Sequences	HD	PAD
1	100 000 000 000	0.833	0.000
2	111 000 100 000	0.666	0.500
3	110 000 000 001	0.500	0.000
4	111 110 000 000	0.166	0.000
5	110 000 111 000	0.833	0.800
6	111 100 000 010	0.166	0.000
7	111 100 100 000	0.500	0.400

טבלה 2. ערכי דימיון המתקבלים משני מדדים שונים המופעלים על זוגות של וקטורים בינאריים

לפערים אלו בהערכת הדימיון-מרחק בין שני פרטים, בין אם הם מיוצגים כווקטורים בינאריים, או כווקטורים של מספרים רציפים, או בכל צורת ייצוג אחרת; לפערים אלו בערכי הדימיון-מרחק בין הפרטים יש השלכות מהותיות על תוצר ניתוח האשכולות המתקבל (Gelbard et al., 2000; Erlich et al., 2002; Erlich et al., 2003). לפיכך, על המשתמש לדייק בבחירת מדד הדימיון-מרחק שיופעל על ידי אלגוריתם ניתוח האשכולות בו הוא משתמש.

מדד הדימיון-מרחק הוא פרמטר שהמשתמש יכול לשנות בכל הרצה, ובאופן זה לראות את התוצאות השונות המתקבלות. בחירת המדד היא בעלת משמעות קריטית, והיא צריכה להתאים למאפייני הבעיה. מחקרים רבים בתחום מדעי האדם עושים שימוש במשתנים המכונים בשם "**Dummy Variables**", היוצרים וקטור של ביטים, שבכל קבוצה של ביטים אחד הביטים דולק (ערכו "1") והביטים האחרים הרלוונטיים אליו כבויים (ערכם "0"). בתחום המחשוב צורת ייצוג זו נקראת בשם גישת Bit Vector with One-Hot Encoding. **בצורת ייצוג שכזו חובה להשתמש במדד דימיון פוזיטיבי** (PAD) כגון Dice, ולא במדד דימיון המתייחס באופן שווה לכל הביטים (הדולקים והכבויים), כדוגמת מדד Hamming. אך לא בטוח שכל מי שמשתמש בצורת הייצוג הזו אכן מודע לסוגי המדדים שיש להפעיל כדי לבטא נכון את הדימיון-מרחק בין הרשומות.

אם נחזור למדדי הדימיון-מרחק שנסקרים בשני המאמרים שצוינו (Cha, 2007; Wijaya, 2016), הרי שכל מאה עשרים וארבעה (124) המדדים האלו מיועדים לחישוב דימיון-מרחק בין שני פרטים. אך תהליך החלוקה לקבוצות בוחן את הדימיון-מרחק לא רק בין זוגות של פרטים, אלא גם בין פרט וקבוצה, ובין קבוצה לקבוצה. בכל סוג של השוואה (בין פרטים, פרט וקבוצה, ובין קבוצות) יש מגוון טכניקות חישוב שמספקות ערכים שונים שמשפיעים על תוצאת החלוקה לאשכולות. **טבלה 3** מציגה כדוגמה שישה מדדים לחישוב דימיון-מרחק בין פרט וקבוצה. כל ששת המדדים עושים שימוש בנוסחה זהה אך מציבים ערכים שונים במקדמים (Ax, Ay, B, C), וכתוצאה מכך מספקים ערכי דימיון שונים המוליכים לתוצאות קיבוץ שונות (Gelbard & Spiegler, 2000).

$$D(\{xy\}, z) = Ax*D(x, z) + Ay*D(y, z) + B*D(x, y) + C*|D(x, z) - D(y, z)|$$

Technique	Ax	Ay	B	C
Nearest neighbor	0.5	0.5	0	− 0.5
Farthest neighbor	0.5	0.5	0	0.5
Median	0.5	0.5	− 0.25	0
Centroid	$Nx/(Nx + Ny)$	$Ny/(Nx + Ny)$	$-Ax*Ay$	0
Group average	$Nx/(Nx + Ny)$	$Ny/(Nx + Ny)$	C	0
Ward's method	$(Nz + Nx)/(Nx + Ny + Nz)$	$(Nz + Ny)/(Nx + Ny + Nz)$	$-Nz/(Nx + Ny + Nz)$	0

טבלה 3. פירוט מקדמים של שישה מדדי דימיון המחשבים דימיון בין פרט וקבוצה (Gelbard & Spiegler, 2000)

באותו אופן ניתן להציג מגוון מדדים להערכת הדימיון-מרחק בין שתי קבוצות, ובכך לתת את המענה השלם להערכת הדימיון : דימיון בין פרטים, דימיון בין פרט וקבוצה, ודימיון בין קבוצות. על בסיס ערכי דימיון-מרחק אלו מחשבים את הדימיון בין החברים בתוך כל אחת מהקבוצות ואת המרחק בין הקבוצות, ובאופן זה מעריכים את שילוב הערכים של WGD ו-BGD, ומקבלים את ההחלטה לגבי החלוקה לאשכולות שנאמץ. גם במדדי הדימיון הבינאריים ישנה התייחסות לחישוב הדימיון בין פרטים ובין קבוצות, ובמקביל לחישוב של PAD - Positive Atom Distance, מוגדר גם חישוב של PGD - Positive Group Distance (Gelbard & Spiegler, 2000).

מהדוגמאות שהוצגו עולה הצורך בהבנה של המשתמש לא רק לגבי ההשלכות הנובעות מהמורכבות והיעילות החישובית של האלגוריתם שיופעל, K-means, Ward, או כל אלגוריתם אחר. אלא להכיר גם את מגוון האפשרויות והתנאים המתאימים להפעלה של כל אחד **ממדדי הדימיון-מרחק**, וכן הבנה של **מדדי ההערכה** בהם ראוי להשתמש לצורך הערכת תוצאת ניתוח האשכולות שהתקבלה.

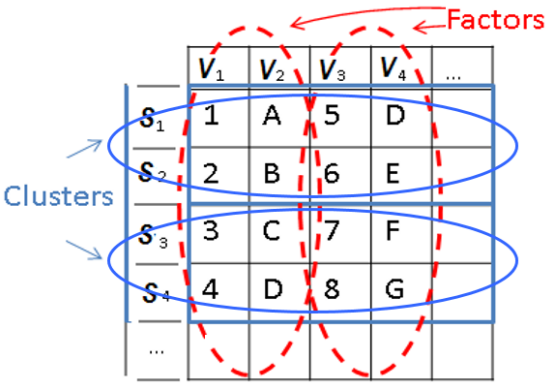
ג. ניתוח גורמים, ייצוג ובולטות של תכונות

ראינו שהמאפיינים של האלגוריתם ושל מדד הדימיון משפיעים על תוצאת החלוקה לקבוצות. כמו כן ראינו שכמעט לכל תוצאה אפשרית של חלוקה לקבוצות ניתן להציג מדד הערכה ש-"יצהיר" שהתוצאה שהתקבלה היא הטובה ביותר האפשרית. מצב זה, בו כל תוצאה, מוטה ככל שתהיה, יכולה לקבל הערכת איכות מושלמת ; מצב בלתי נסבל זה של חוסר יציבות, הוא מסוכן ומחייב את המשתמש להבנה נרחבת כדי להבטיח רמה מקצועית ואתית גבוהים.

חוקרים שעוסקים בהיבט המתודי של ניתוח אשכולות מודעים למצב לא יציב זה ופועלים בהתאם. אך אם נבחן את הספרות בתחום של ניתוח גורמים (Factor Analysis), נמצא שם התייחסות מועטה לנושאים האלה, למרות שבניתוח גורמים מבוצעת פעולה זהה לזו שמתבצעת בניתוח אשכולות, אלא שהיא מופעלת על המשתנים (Variables) ולא על הרשומות (Samples). שכן, **ניתוח גורמים הוא עיבוד אורתוגונלי לניתוח אשכולות**, והמשתמש הנאיבי שרוי בתחושה שהתוצאות המתקבלות הן המיטביות ושלא יכולות היו להתקבל תוצאות אחרות. זו כמובן טעות מוחלטת, כיוון שניתוח גורמים מפעיל טכניקה זהה של ניתוח

אשכולות, על אותו סט נתונים בדיוק, תוך שהוא מפעיל אותה לצורך קיבוץ המשתנים לאשכולות.

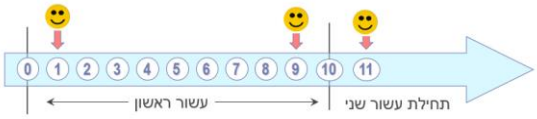
איור 2 ממחיש את האורתוגונליות וההקבלה בין ניתוח אשכולות וניתוח גורמים. המטריצה מציגה ארבע רשומות - Samples (S1, S2, S3, S4), וארבעה משתנים (V1, V2, V3, V4), ואת ערכו של כל משתנה בכל רשומה. את ניתוח האשכולות מייצגות האליפסות הכחולות שכל אחת מהן מייצגת אשכול (Cluster), ואת ניתוח הגורמים מייצגות האליפסות האדומות שכל אחת מהן מייצגת גורם (Factor). ניתוח האשכולות וניתוח הגורמים משתמשים כאמור באותו סט של נתונים. אם נסובב את מטריצת הנתונים בתשעים מעלות שמאלה ונפעיל אליה את כלי ניתוח האשכולות הרי שביצענו ניתוח גורמים, אלא שכעת אנחנו מודעים לחוסר היציבות של התוצר שקיבלנו ולקושי בהערכת איכותו. לפיכך, כל חוקר העושה שימוש בניתוח גורמים ראוי שיהיה מודע למגבלות הנידונות במאמר זה ולזהירות בה יש לנהוג.



איור 2. התהליכים השקולים של ניתוח אשכולות וניתוח גורמים

הסבת המבט אל התכונות-המאפיינים של הפרטים, אל המשתנים (Variables), מעלה את נושא החלופות לייצוג המשתנים, ואת השפעת הייצוג על הערכת הדימיון-מרחק ועל ההקבצה שתתקבל. כל משתנה ניתן לייצוג במגוון צורות. אפשר לייצג כמספר רציף, כמספר בדיד, כמשתנה קטגורי, או בייצוג בינארי. לדוגמה, מידה של מכנסיים ניתן לייצג כמספר רציף של היקף המותניים בסנטימטרים, כמספר רציף של היקף המותניים חלקי שניים (כפי שמקובל במידות-UK/US), כמשתנה נומינלי (XS, S, M, L, XL), כמשתנה בינארי, בו הביט של המידה המתאימה דולק "1" וכל שאר הביטים כבויים "0" (Bit Vector with One-Hot Encoding), או כמספר בדיד כלשהו (1,2,3 וכד') שיצרן הבגד מחליט עליו בצירוף טבלת המרה. אנחנו יכולים גם ל"נסר" את טווח הערכים בכל אופן אחר שנראה לנו כמתאים לבעיה ולייצג את הטווחים בצורה מספרית או נומינאלית כלשהי.

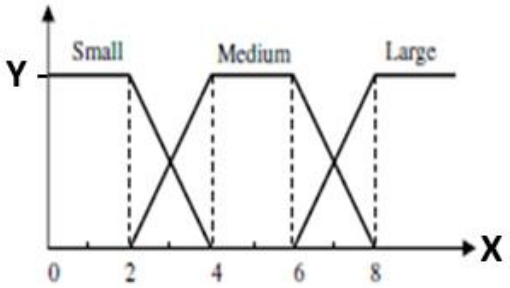
איור 3 ממחיש את השפעת צורת הייצוג על הערכת הדימיון-מרחק בין הפרטים. נניח שאנחנו רוצים לייצג גיל של ילד. נסתכל על שתי צורות ייצוג. באחת, הייצוג הוא כמספר טבעי של גיל הילד בשנים, והשני ישקף את גיל הילד בעשורים. נבחן כעת את הערכת הדימיון שנקבל לגבי שלושה ילדים: ילד בן שנה, ילד בן 9 שנים וילד בן 11 שנים. ייצוג על פי גיל הילד בשנים יוביל לכך שילד בן 9 יוגדר כדומה יותר לילד בן 11 מאשר לילד בן שנה. לעומת זאת, ייצוג על פי עשורים יקבע שילד בן 9 דומה לילד בן שנה, וכי שניהם אינם דומים כלל לילד בן 11.



איור 3. השפעת צורת הייצוג על הערכת הדימיון-מרחק בין הפרטים

לשאלה: "למה לייצג גיל של ילד באופן שאינו מספר השנים?" ניתן להשיב על ידי דוגמאות כגון משימה שמטרתה לקבץ משפחות על פי העומס המוטל על היחידה המשפחתית. במקרה שכזה מומחה התוכן יכול להחליט שהנטל בגידול ילד בשנת חיין הראשונה (גיל 0-1) שונה מנטל הגידול בשנתיים שלאחר מכן (מגיל שנה ועד 3), ושונה מנטל הגידול בחמש השנים שלאחר מכן (עד גיל 8), וכד'. כלומר, מומחה התוכן יכול להחליט "לנסר" את סקלת הגיל בכל דרך שנראית לו מתאימה לבעיה איתה הוא מתמודד, והאיש הטכני אמור לוודא שצורת הייצוג ומדד הדימיון בו הוא ישתמש, יפיקו ערכי דימיון שתואמים לרציונל של הבעיה. דוגמה אחרת יכולה להיות ייצוג של מקומות (עיר, מדינה וכד'). ייצוג מקום על ידי קוד מספרי ייצור דימיון בין מקומות על בסיס הקרבה בין המספרים – דימיון שאינו קיים במציאות. לחילופין, ייצוג מקומות באופן נומינלי או בגישה הבינארית יוביל למצב בו אין כל דימיון בין מקומות, וכל מקום דומה אך ורק לעצמו.

צורת ייצוג נוספת היא הייצוג העמום, שעושה שימוש בפונקציות של משולשים וטרפזים כדי לבטא את סיווג טווחי הערכים האפשריים של המשתנה. **איור 4** מדגים מצב בו טווח הערכים 0-2 מוגדר כ-small, טווח הערכים 4-6 מוגדר כ- medium וערכים של 8 ומעלה מוגדרים כ-Large. יחד עם זאת, הוא מגדיר פונקציות לינאריות שמבטאות את המעבר בין small ל-medium, ובין medium ל-large. בדוגמה המוצגת באיור 4, הערך שלוש יוגדר כחצי small וחצי medium. מכאן, שכדי לבטא את המאפיין-התכונה של גודל, כפי שהוא מוצג באיור, לא ניתן להסתפק במשתנה אחד ויש צורך להשתמש בשלושה משתנים נפרדים, אחד ל-small, אחד ל-medium ואחד ל-large, וכל אחד מהם יכול לקבל ערך מספרי בטווח של 0-1. באופן זה, הייצוג של המידה 3 תבוטא על ידי שלושה ערכים מספריים בהם: small=0.5, medium=0.5, large=0. השימוש בשלושה משתנים כדי לייצג את תכונת הגודל יכולה להיראות מוזרה, אבל יש לה יתרון כאשר אנחנו רוצים לבטא תכונות שיכולה להיות בהן עמימות. אם לדוגמה יבקשו מאיתנו לתאר אדם נרצה שתהיה לנו אפשרות להעריך את הגובה שלו כגובה שהוא "בין בינוני לגבוה". צורת הייצוג העמומה, שהודגמה באיור 4, מתאימה למטרה זו וניתן להתאים לה מגוון רחב של מדדי דימיון (Meged & Gelbard, 2011).



איור 4. שימוש בפונקציות של טרפזים לייצוג עמום של תכונות

בהתייחסות למדדי דימיון (בפרק הקודם - ב'), צוינו שני מאמרי סקירה, האחד הציג ארבעים וחמישה (45) מדדי דימיון למשתנים מספריים (Cha, 2007), והשני הציג שבעים ותשעה (79) מדדי דימיון עבור משתנים בינאריים (Wijaya, 2016). תכונות נומינליות נוח מאוד לייצג כווקטור בינארי, ובמחקרים רבים ממירים אותן ל- Dummy Variables כדי לאפשר לבודד בין הערכים השונים ובאופן זה לבדוק את מידת ההשפעה על המודל. בניגוד לכך, משתנים מספריים ומשתנים אורדינליים לא ניתן לכאורה לייצג באופן בינארי, כיוון שבייצוג שכזה אובד יחס הסדר בין משתנים אורדינליים, וכן אובד היחס הכמותי בין משתנים מספריים (היחס של פי כמה). ניתן להתגבר על מגבלה זו על ידי שימוש בטכניקה של "ריפוד" הביטים הדולקים (Gelbard, 2013), ובאופן זה להשתמש במדדי הדימיון הבינאריים גם לצורך חישוב דימיון-מרחק ברשומות המכילות משתנים אורדינליים ומשתנים מספריים.

חישוב הדימיון מושפע מאוד מטווח הערכים שבכל משתנה. אם במשתנים A ו-B, טווח הערכים הוא 1-10, ובמשתנה אחר C, טווח הערכים הוא 1-1,000 אז ערכי המרחק שיתקבלו במשתנה C יהיו גדולים הרבה יותר מאשר ערכי המרחק שיתקבלו בשני המשתנים A ו-B יחד, ויגרמו להטיה שתושפע כמעט אך ורק מהערכים בתכונה C. כדי להתגבר על מגבלה זו יש להמיר את כל הערכים המספריים לערכים מנורמלים או לערכים מתוקננים.

נירמול (Normalization) פירושו הסבת הערכים המספריים לטווח של 0-1, על פי הנוסחה:
Xnorm = (X-Xmin)/(Xmax-Xmin) כאשר:
Xmin = הערך המינימלי האפשרי בטווח ו-Xmax = הערך המקסימלי האפשרי בטווח.

תיקנון (Standardization) פירושו הסבת הערכים המספריים כך שתתקבל התפלגות בה הממוצע=0, וסטיית תקן=1 (μ=0, σ =1), על פי הנוסחה:
Xstand = (X-μ) / σ כאשר: μ = ממוצע ערכי X, σ = סטיית התקן של ערכי X.

כאשר טווח הערכים של משתנה מוגדר באופן ברור ולא צפויים בו ערכי קיצון אז מקובל להשתמש **בנירמול**. לדוגמה, המשתנה "גיל הסטודנט בשנים" לא יכול לקבל ערכי קיצון. אם במקרה תתקבל רשומה ובתכונת הגיל יהיה רשום 4 או 140, אז יהיה לנו ברור שמדובר בטעות ונפעל כדי לוודא את גילו האמיתי של הסטודנט. לעומת זאת, כאשר משתנה צפוי לקבל ערכי קיצון, כגון בתחומים בהם מתקיים מצב של "זנב ארוך", אז מקובל להשתמש בטכניקה של **תיקנון**.

הפעולות שתוארו עד כה, בחירת אופן הייצוג של משתנה, אפשרות המעבר בין צורות ייצוג שונות (כגון

מעבר ממשתנים מספריים ואורדינליים לייצוג בצורה בינארית), נירמול ותקנון הערכים של המשתנה בהתאם לאפשרות ההופעה של ערכי קיצון - **פעולות אלו הן הכרחיות כדי לצמצם את ההטיה שעלולה להתקבל ממדדי הדימיון-מרחק**.

נקודה חשובה נוספת היא האפשרות לתת משקל שונה לתכונות. כעיקרון לא היינו רוצים לתת מראש משקלות לתכונות, אלא לקבל התייחסות לחשיבות, לבולטות של כל תכונה (saliency), בסוף תהליך ניתוח האשכולות. כלים שונים נותנים תמונה של בולטות התכונות אך ברובם ההתייחסות היא לבולטות התכונה בכלל האוכלוסייה, ולא לבולטות התכונות באשכולות השונים. לדוגמה, אם בכלל האוכלוסייה אחוז הגברים דומה לאחוז הנשים, האם במקרה שכזה נגדיר את התכונה "מגדר" כתכונה בולטת? באופן טבעי, כשפיזור הערכים הוא אחיד (אחוז דומה של גברים ונשים) הנטייה תהיה להגיד שהתכונה אינה "בולטת". יחד עם זאת אם האוכלוסייה הייתה מחולקת לשלושה אשכולות, שבאחד אחוז הגברים היה זהה לאחוז הנשים, ובשני הגברים הם 97%, ובשלישי הנשים הן 97%. במקרה שכזה היינו אומרים שהתכונה "מגדר" היא תכונה בולטת, והיינו מתייחסים אליה גם במסגרת תיאור האשכולות שהתקבלו, וגם לצורך הגדרת כללי סיווג אפשריים. משמעות הדבר היא שבולטות של תכונה מושפעת מפיזור הערכים באשכולות השונים, ולא רק באוכלוסייה הכללית. כמו כן, ניתן להעריך את מידת הבולטות של תכונה לא רק במשתנים נומינליים (כדוגמת "מגדר") אלא גם לגבי משתנים מספריים ואורדינליים (Barak & Gelbard, 2012; 2016).

ד. אנסמבל אלגוריתמים וניתוח פרטים שיש ספק לגביהם

חוסר היציבות מתבטא גם בכך שכל אלגוריתם צפוי לתת תוצאה שונה, גם אם כולם ישתמשו באותם מדדי דימיון ובאותן צורות לייצוג המשתנים. כמו כן, אין אלגוריתם שמניב באופן עקבי תוצאות טובות יותר מאלגוריתמים אחרים כאשר בודקים אותם על נתונים ממגוון רחב של בעיות. לפיכך, אין קיצורי דרך ובכל בעיה יש לבחון מחדש את מרחב האפשרויות לניתוח הנתונים במטרה לאתר את השילוב שיפיק את התוצאה הטובה ביותר (Gelbard et al., 2007). כפועל יוצא מזה כלי הניתוח מאפשרים הפעלה אוטומטית של מגוון אלגוריתמים כשעבור כל אלגוריתם ניתן להגדיר טווח ערכים לכל אחד מהפרמטרים, והמערכת מפיקה סיכום מרוכז של התוצאות שהתקבלו. לחילופין, על המשתמש לעשות זאת באופן ידני, או לשלב קוד ייעודי שיפעיל את רצף הבדיקות ויסכם אותן. דבר זה דורש מהחוקר מאמץ רב, אך כאמור אין קיצורי דרך ואין ארוחות חינם, גם לא בהקשר של אלגוריתמיקה (Wolpert & Macready, 1997).

לאחר שהחוקר מבצע סידרה של הרצות המשלבות אלגוריתמים, מדדי דימיון, ומדדי הערכה מגוונים, מתגבש מספר מצומצם של חלופות טובות. בשלב זה החוקר מנסה להבין, מעבר למדדים הכללים (WGD ו-BGD), גם היבטים מפורטים אודות האשכולות, כדוגמת התכונות הבולטות בכל אשכול (כפי שתואר בפרק הקודם). אבל כאן החוקר נתקל בבעיה נוספת והיא המגבלה ביכולת להשוות בין האשכולות הספציפיים שהתקבלו. שכן, אין כל אחידות בסימון ובהגדרה של האשכולות שהתקבלו בכל אחת מההרצות, גם אם האוכלוסייה חולקה למספר זהה של אשכולות.

איור 5 ממחיש את הבעיה על ידי הצגת דוגמה סכמתית בה שני אלגוריתמים נתבקשו לסווג אוכלוסייה שיש בה שמונה פרטים, לשלושה אשכולות. כל אחת משלוש השורות בטבלה מציגה את הפרטים ששויכו לכל אחד משלושת האשכולות, כשכל אשכול מיוצג בשורה נפרדת. לדוגמה, אלגוריתם A סיווג לאשכול שקיבל את התגית C1 את הפרטים 1,2,3. ואילו אלגוריתם B סיווג לאשכול שקיבל את התגית C1 את הפרטים 1,4,5. כתוצאה משיוך אוסף שונה של פרטים לאשכול שמתויג כ- C1, הרי שהמאפיינים של האשכול שהתקבל על ידי אלגוריתם A יהיו שונים מאלו שהתקבלו באלגוריתם B, ולמשתמש ברור כי למרות התיוג הזהה כ- C1 מדובר באשכולות שונים. כדי להתמודד עם הבעיה הזו האיור מציג סימונים באדום, ירוק וכחול, הממחישים את התהליך שהחוקר היה עושה בשלב זה. החוקר היה בוחן את הדימיון בין האשכולות על בסיס החברים המשתייכים לכל אשכול ומגיע למסקנה: שאשכול C1 באלגוריתם A דומה לאשכול C3 באלגוריתם B. שאשכול C2 באלגוריתם A דומה לאשכול C1 באלגוריתם B, ושאשכול C3 באלגוריתם A דומה לאשכול C2 באלגוריתם B. לאחר מציאת ההתאמה הזו החוקר יכול לתייג את

האשכולות באופן "סטנדרטי" באמצעות התיוגים: "אשכול אדום", "אשכול ירוק" ו-"אשכול כחול".

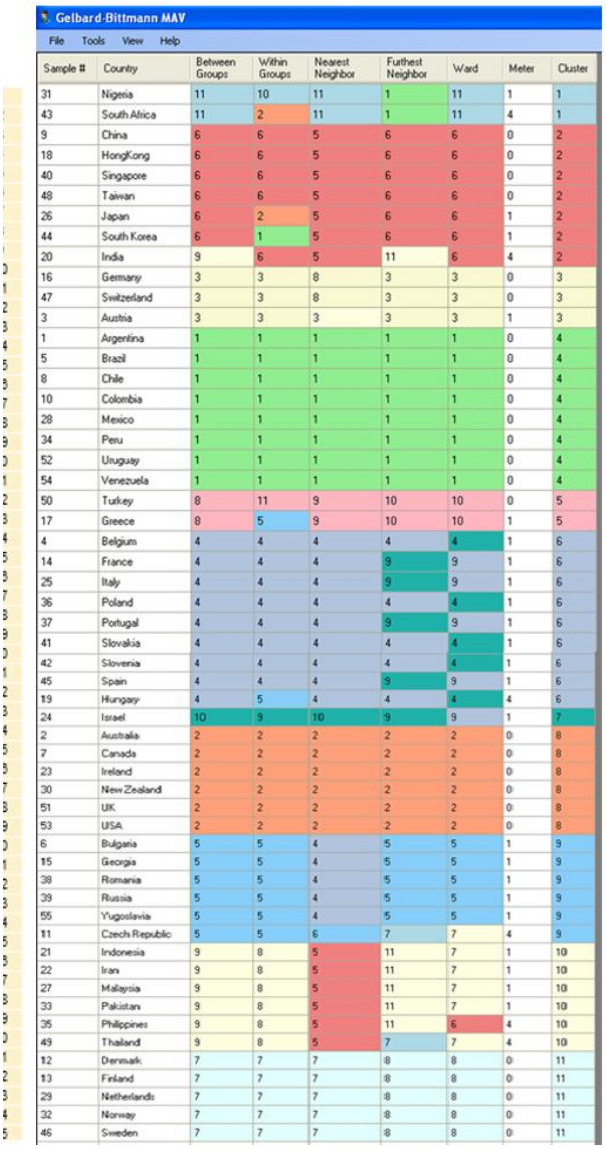
Cluster	Algorithm A	Algorithm B
C1	1, 2, 3	1, 4, 5
C2	4, 5	7, 8
C3	6, 7, 8	2, 3, 6

איור 5. מציאת "מכנה משותף" לאשכולות המתקבלים בהרצות שונות

פעולה זו לא רק מאפשרת סטנדרטיזציה של תיוג האשכולות, אלא גם מאפשרת להעמיק את הניתוח לרזולוציה של רשומה בודדת ולשפר את ההבנה אודות הנתונים, בזכות האפשרות לשאול שאלות כגון: לאיזה אשכול (אדום, ירוק או כחול) משויכת הרשומה על ידי כל אחד מהאלגוריתמים? מה מידת ההתאמה של החלוקות שהתקבלו על ידי האלגוריתמים השונים? לגבי איזה רשומות יש חוסר הסכמה בין האלגוריתמים ולגבי איזה רשומות יש הסכמה רחבה?

בחינת נתונים תוך שימוש במספר אלגוריתמים והשוואת התוצאות המתקבלות על ידי כל אחד מהם נקראת Ensemble learning. גישה שבאמצעות מכלול של אלגוריתמים (Ensemble algorithms) וניצול נקודות החוזקה של כל אחד מהם, מאפשרת לצמצם שגיאות ולשפר את תהליך הלמידה וההבנה (Dong et al., 2020). אך הפעולה הזו היא מאוד מורכבת, במיוחד כשהיינו מעוניינים לחזור עליה בכל חלוקה למספר שונה של אשכולות, ובלי הגבלה על מספר האלגוריתמים בהם נרצה להשתמש.

את התוצרים הנ"ל ניתן להפיק בטכניקה המכונה בשם Multi Algorithm Voting (Bittmann & Gelbard, 2007). MAV מאפשרת השוואה של תוצרי ניתוח אשכולות המתקבלים על ידי מגוון אלגוריתמים, סטנדרטיזציה של תיוג האשכולות, ובדיקת השיוך של כל רשומה בחלוקה שהתקבלה על ידי כל אלגוריתם. בנוסף השיטה מציעה דרך נוספת לקביעה של מספר האשכולות המיטבי אליו כדאי לחלק את האוכלוסייה, מספר שנקבע על ידי החלוקה בה תהיה הומוגניות מירבית של האשכולות שיתקבלו על ידי האלגוריתמים השונים. כמו כן, MAV נותנת תמונה ויזואלית אחודה של המידע הנ"ל. ניתן לראות שימוש ב-MAV במחקרים בפסיכולוגיה ובהתנהגות ארגונית (Gelbard et al., 2009; Ronen & Shenkar, 2013; Kempen et al. 2019). **איור 6** מדגים פלט של MAV בניתוח בעיה מתחום ההתנהגות הארגונית בה החוקרים, רונן ושנקר, ערכו מחקר השוואתי לניתוח תרבות ניהולית במגוון רחב של מדינות (Ronen & Shenkar, 2013). צילום המסך מציג טבלה שכל שורה בה מייצגת מדינה שנכללה במחקר, וכל עמודה מייצגת אלגוריתם. האלגוריתמים שנבחרו הם אלגוריתמים שהשיגו תוצאות טובות במחקרים דומים, ושמות האלגוריתמים מיוצגים על ידי המדד שניתנה לו העדפה באותו אלגוריתם: Between Groups - נתן משקל יחסית גבוה למרחק בים האשכולות. Within Groups - נתן משקל גבוה יחסית לקרבה של הפרטים בכל אשכול. Nearest neighbor – העריך את המרחק בין האשכולות על בסיס זוג השכנים הקרוב ביותר. Furthest neighbor – העריך את המרחק בין האשכולות על בסיס זוג השכנים הרחוק ביותר. ו-Ward שמעריך את המרחק על בסיס מוקדי התכונות (Centroids) בכל אחד מהאשכולות.


Gelbard-Bittmann MAV

File Tools View Help

Sample #	Country	Between Groups	Within Groups	Nearest Neighbor	Furthest Neighbor	Ward	Meter	Cluster
31	Nigeria	11	10	11	1	11	1	1
43	South Africa	11	2	11	1	11	4	1
9	China	6	6	5	6	6	0	2
18	HongKong	6	6	5	6	6	0	2
40	Singapore	6	6	5	6	6	0	2
48	Taiwan	6	6	5	6	6	0	2
26	Japan	6	2	5	6	6	1	2
44	South Korea	6	1	5	6	6	1	2
20	India	9	6	5	11	6	4	2
16	Germany	3	3	8	3	3	0	3
47	Switzerland	3	3	8	3	3	0	3
3	Austria	3	3	3	3	3	1	3
1	Argentina	1	1	1	1	1	0	4
5	Brazil	1	1	1	1	1	0	4
8	Chile	1	1	1	1	1	0	4
10	Colombia	1	1	1	1	1	0	4
28	Mexico	1	1	1	1	1	0	4
34	Peru	1	1	1	1	1	0	4
52	Uruguay	1	1	1	1	1	0	4
54	Venezuela	1	1	1	1	1	0	4
50	Turkey	8	11	9	10	10	0	5
17	Greece	8	5	9	10	10	1	5
4	Belgium	4	4	4	4	4	1	6
14	France	4	4	4	9	9	1	6
25	Italy	4	4	4	9	9	1	6
36	Poland	4	4	4	4	4	1	6
37	Portugal	4	4	4	9	9	1	6
41	Slovakia	4	4	4	4	4	1	6
42	Slovenia	4	4	4	4	4	1	6
45	Spain	4	4	4	9	9	1	6
19	Hungary	4	5	4	4	4	4	6
24	Israel	10	9	10	9	9	1	7
2	Australia	2	2	2	2	2	0	8
7	Canada	2	2	2	2	2	0	8
23	Ireland	2	2	2	2	2	0	8
30	New Zealand	2	2	2	2	2	0	8
51	UK	2	2	2	2	2	0	8
53	USA	2	2	2	2	2	0	8
6	Bulgaria	5	5	4	5	5	1	9
15	Georgia	5	5	4	5	5	1	9
38	Romania	5	5	4	5	5	1	9
39	Russia	5	5	4	5	5	1	9
55	Yugoslavia	5	5	4	5	5	1	9
11	Czech Republic	5	5	6	7	7	4	9
21	Indonesia	9	8	5	11	7	1	10
22	Iran	9	8	5	11	7	1	10
27	Malaysia	9	8	5	11	7	1	10
33	Pakistan	9	8	5	11	7	1	10
35	Philippines	9	8	5	11	6	4	10
49	Thailand	9	8	5	7	7	4	10
12	Denmark	7	7	7	8	8	0	11
13	Finland	7	7	7	8	8	0	11
29	Netherlands	7	7	7	8	8	0	11
32	Norway	7	7	7	8	8	0	11
46	Sweden	7	7	7	8	8	0	11

איור 6. הדגמת MAV על נתוני תרבות ניהולית (Ronen & Shenkar, 2013)

הפלט מתייחס להרצה בה כל האלגוריתמים התבקשו לחלק את האוכלוסייה ל-11 אשכולות. בהרצות אחרות האלגוריתמים נתבקשו, בכל הרצה, לחלק את האוכלוסייה למספר אחר של אשכולות, אך בדיקה של מידת ההסכמה-הומוגניות של התוצאות הראתה שהחלוקה ל-11 אשכולות קיבלה את ההסכמה המירבית, ועדיין יש חוסר הסכמה לגבי מספר מקרים, כפי שיוסבר. כל אחד מהאשכולות נצבע בצבע שונה, כפי שתואר בהסבר לאיור 5, והעמודה הימנית מציגה את הצבעים של 11 האשכולות. המספר שרשום בכל תא (למעט בשתי העמודות הימניות) מייצג את התיוג שניתן לאשכול על ידי כל אלגוריתם. הדבר בא להמחיש את העובדה שלא ניתן להשוות את הסיווג של הפרטים על ידי קוד האשכול שניתן על ידי כל אלגוריתם ולכן יש צורך בפעולת ההשוואה והסימון בצבעים נבדלים.

נסתכל בשתי השורות הראשונות הצבועות בכחול-אפרפר. האלגוריתם המכונה Furthest neighbor חושב שאת שתי השורות האלה יש לשבץ לאשכול הירוק (קבוצת הצבע הרביעית שבמסך), והאלגוריתם המכונה Within Groups חושב שאת השורה השנייה היה צריך לשייך לאשכול הכתום (קבוצת הצבע השמינית שבמסך). למרות זאת, הגדרת הרוב בהצבעה שנערכה בין האלגוריתמים קבעה ששניהם יהיו אשכול עצמאי. אם נסתכל על קבוצת הצבע העשירית (צהוב בהיר)נראה שם שהאלגוריתם המכונה Nearest neighbor ממליץ לשייך את כל החברים באשכול לאשכול האדום (קבוצת הצבע השנייה). נוכל לראות גם שיש חילוקי דעות גדולים לגבי הפרט האחרון בקבוצה (שורה 50 בסרגל שבצד שמאל של המסך) - Nearest neighbor ממליץ לשייך אותו לקבוצה האדומה, ו- Furthest neighbor ממליץ לשייך אותו לקבוצת הצבע השישית.

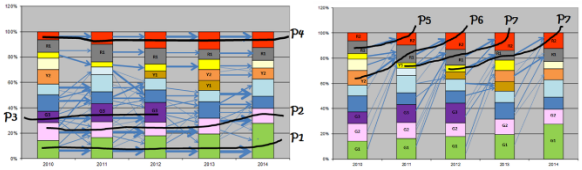
המידע הזה אודות פרטים שיש ספק לגבי אפשרות השיוך שלהם לאשכולות שונים (כגון שורה 2, שורה 9 ושורה 50), והמידע אודות האופן השונה בו האלגוריתמים היו משייכים את הפרטים בקבוצת הצבע השישית (שורות 32-24), הם בעלי חשיבות רבה לחוקר וניתן להגיע אליהם אך ורק בטכניקה של Ensemble learning, כפי שעושה MAV תוך שהוא מוסיף את הייצוג הוויזואלי שמפשט את הבנת התמונה, ומספק דרך נוספת לקביעה של מספר האשכולות המיטבי אליו כדאי לחלק את האוכלוסייה.

ה. ניתוח אשכולות דינאמי וניתוח על פני זמן

כאשר ליאו מסי עבר, בשנת 2021, מקבוצת ברצלונה לפריז סן-ג׳רמן (PSG). ביום הראשון למכירת חולצת המדים החדשה של מסי, נמכרו כ-832,000 חולצות שלו בתוך 30 דקות, בסכום של כ-10.5 מיליון דולר (Aaron, 2021). האם רוכשי החולצה החדשה היו כולם אוהדי PSG? או שיש ביניהם גם אוהדי ברצלונה שהמירו כעת את אהדתם ל-PSG? האם ה-״אשכול״ של קבוצת כדורגל כלשהי הוא קבוע, או שהוא יכול להשתנות על פני זמן בהתאם לאוסף החברים הנכלל

באשכול? את השאלה הזו אפשר לשאול בהקשרים שונים כשהתפיסה של אשכול היא תפיסה דינאמי שיכולים להיות בה חלקים קבועים, כמו יליד פריז שאוהד את הקבוצה כל חייו, אך גם חלקים משתנים, כמו אוהד של מסי ש-"נודד" יחד איתו אל כל קבוצה שישחק בה. בתחום אחר, כגון בתחום הפיננסים, ניתן חלק אגרות חוב קונצרניות לאשכולות ולבחון את השינויים בחלוקה לאשכולות על פני זמן, של אותן אגרות חוב. אפשר לבחון את האשכולות המתקבלים במונחים של רמות סיכון, תוך שהגדרת הסיכון יכולה להשתנות במהלך השנים. לדוגמה, רמת סיכון גבוהה יכולה להיות בעלת מאפיינים שונים בשנה של פריחה כלכלית, ביחס למאפיינים שהיו לה בשנה של שפל כלכלי. במצב שכזה אפשר לראות באשכול אנאלוגי למצב (State), ולנדידה ולשינויי ההשתייכות של הפרטים השונים בנקודות הזמן השונות כמעבר בין מצבים, State Transitions, באופן דומה למקובל בעולם המודלים הדטרמיניסטיים (Harel, 1987). היבטים אלו, של תנודתיות אפשרית בשיוך של פרט לאשכול, ואפשרות ההופעה או ההיעדרות של אשכול בנקודת זמן מסוימת, מבטאים "חוסר יציבות" נוסף שהחוקר צריך להיות מודע אליה. שכן, בנקודת זמן שונות לא רק שפרט יכול להשתייך לאשכולות שונים, אלא שייתכן שאשכול יופיע, יעלם בנקודת זמן אחרת, ויופיע שוב לאחר מכן. וכך, ניתוח שיעשה בנקודת זמן אחת ייתן תמונה שונה מזו שתתקבל בנקודת זמן אחרת.

איור 7 ממחיש ניתוח אשכולות על פני זמן וזיהוי של תבניות נדידה. הדוגמה מציגה מעקב לאורך חמש השנים (בין השנים 2010-2014) אחר ההתנהגות של אגרות חוב קונצרניות. הצבעים מייצגים את האשכולות שזוהו בכל שנה, והחיצים מייצגים את תנועת אגרות החוב בין האשכולות במהלך חמש השנים. ניתן לראות שהאשכול הסגול שהופיע במהלך שלוש השנים הראשונות (2010-2012), נעלם ולא מופיע ב-2013 וב-2014. באותו אופן, האשכול הכתום מופיע בשנת 2010, נעלם בשנתיים שלאחר מכן, ומופיע שוב ב-2013 וב-2014. . הקווים השחורים המודגשים בגרף הימני מייצגים תבניות של התדרדרות ערך אגרות החוב עד לכלל חדלות פירעון (P5, P6, P7). הצבע האדום מייצג את האשכול של חדלות הפירעון וניתן לראות שהניתוח מאפשר לזהות תבנית נדידה שנמשכת שלוש שנים במהלכן אגרת החוב עוברת מהאשכול הצהוב אל האשכול האפור ולאחר מכן אל האשכול האדום של חדלות הפירעון.

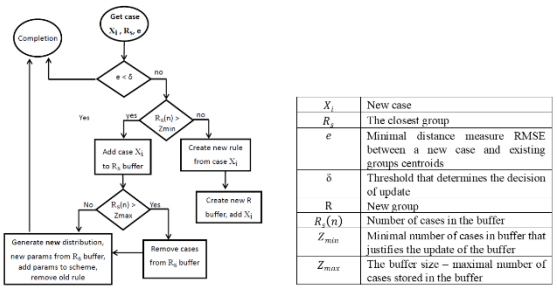


איור 7. תבניות נדידה בין אשכולות (Ramon-Gonen & Gelbard, 2017)

הדוגמה ממחישה היבט דינאמי שיש להביא בחשבון במסגרת התהליך של ניתוח אשכולות. כלומר, גם אם מהות האשכולות נשארת ללא שינוי על פני זמן, הרי שייתכנו שינויים במאפיינים של האשכול, כגון המאפיינים של אשכול אג"ח בסיכון גבוה שיכולים להשתנות בתקופות של גאות ושפל כלכליים. כמו כן הדוגמה ממחישה היבט אונטולוגי, לפיו ייתכן ואשכול מסוים, כגון האשכול הסגול, יופיע בתקופת זמן מסוימת ולא יופיע בתקופות אחרות. נקודה שלישית שהדוגמה ממחישה היא שניתוח הפרטים הנכללים בכל אשכול בנקודות הזמן השונות, יכול להעיד על דפוסים של נדידה, כך שיש לראות את הרכבו של אשכול, כמו גם את שיוכו של פרט לאשכול מסוים, כתופעות דינאמיות שיכולות ואפילו צפויות להשתנות לאורך הזמן.

המודעות להיבט הדינאמי של התופעות, ולאופן השונה בו הן ישתקפו בניתוח האשכולות שנבצע בנקודות הזמן השונות, מעלה את הצורך להתייחס אל תהליך ניתוח האשכולות כאל תהליך שאמור להיות פתוח כל הזמן לשינויים. באופן זה יש לאפשר יצירה של אשכולות חדשים באופן דינאמי, שינוי תכונות של אשכול, מיזוג אשכולות לאורך זמן, ביטול אשכולות, ואפילו פתיחת אשכול ייעודי לכל מקרה חריג שיופיע (Gelbard, 2019; Khalemsky & Gelbard 2020).

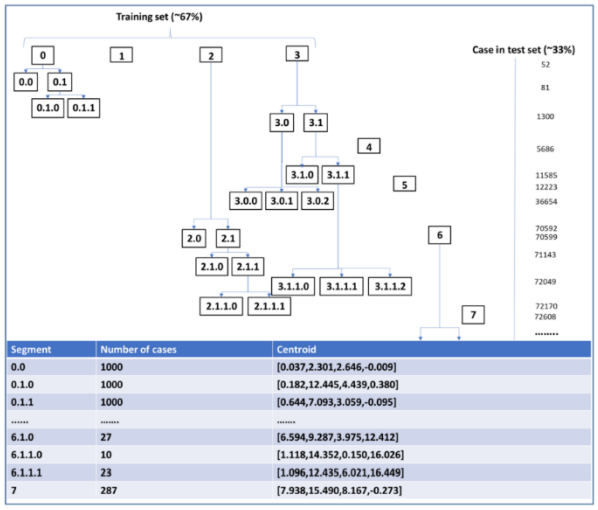
איור 8 מציג תרשים זרימה של לוגיקה ליישום הדינאמיות שתוארה, ומתאר מנגנון להגדרה דינאמית של אשכולות שמשתנים כל העת בהתאם לפרטים המופיעים בפניו. מנגנון זה יכול לעבוד כסוכן עצמאי תוך שהוא מתקשר עם סוכנים מקבילים לו הפועלים בסביבות אחרות, מייצרים אשכולות סיווג אחרים, ומעדכנים אחד את השני כל העת.



איור 8. תרשים זרימה של מסווג דינאמי (Gelbard, 2019)

במסגרת התפיסה הדינאמית של תהליך ניתוח האשכולות, יש כאמור לאפשר יצירה של אשכול חדש ונבדל עם ההופעה של כל פרט חריג. אנשי שיווק, מסיבות אסטרטגיות, מעדיפים שבחלוקה לאשכולות לא יהיו אשכולות קטנים, כיוון שזה יכול לחייב פעולות ייעודיות כגון, פיתוח של גרסאות מוצר שונות לכל אשכול. יחד עם זאת, צריך להבין שפרט חריג יכול להיות שהוא מקרה חריג (outlier) או אפילו טעות (בטעות שויכו לו ערכים חריגים, כגון רשומה של סטודנט בה גיל הסטודנט הוא 4 או 140), אך באותה מידה יכול להיות שזו הסנונית הראשונה שמבשרת את תחילתה של תופעה חדשה, ואנחנו מאוד רוצים שתהיה לנו אפשרות לזהות את המגמות החדשות מוקדם ככל האפשר. האשכול החדש שנייצר יכול להתמלא עם הזמן בפרטים נוספים, ובאמת לייצג תופעה שמתחילה להתרחב. אך באותה מידה יכול להיות שהאשכול החדש ימשיך להכיל את הפרט הבודד שיסתבר כמקרה חריג (outlier), או מספר מצומצם של מקרים שבהיבט האסטרטגי של השיווק לא ניתן יהיה לתת להם את הטיפול הספציפי הנדרש.

הייצוג הוויזואלי של דינאמיות זו הוא מורכב ולא ניתן לבטא אותו באמצעות תרשים מסוג דנדרוגרמה (דוגמה לדנדרוגרמה מוצגת באיור 1). לצורך ייצוג ויזואלי שיאפשר בו זמנית לבטא גם את מימד הזמן וגם את האפשרות לשינוי דינאמי בתכולה ובהרכב הקבוצות, לצורך זה ניתן להשתמש ב- ExpanDrogram המבטא הרחבה והוספת היכולות הדינאמיות לוויזואליזציה של הדנדרוגרמה (Khalemsky & Gelbard 2021). **איור 9** מדגים את מאפייני ה- ExpanDrogram ואת תהליך התפתחותה עם ההופעה של פרטים חדשים. העץ ממוספר בשיטה היררכית כמקובל בספרור של תוכן עניינים, ובשיטות כגון WBS (Work Breakdown Structure). כל פרק יכול להתפצל לעוד ועוד תת פרקים, שכל אחד מהם משרשר את הקוד של מי שהיה לפניו. לדוגמה, פרק 3 יכול להתפצל לתת פרקים 0 ו-1 שסיפרורם יהיה 3.0 ו-3.1. ותת פרק 3.0 יכול להתפצל לתת פרקים נוספים שסיפרורם יהיה 3.0.0, 3.0.1 ו- 3.0.2. באותו אופן פרק יכול שלא יהיו לו תת פרקים. ובשונה מהמקובל בתוכן עניינים, תת פרקים יכולים גם להתאחד בהמשך ואז יקבלו ספרור כפול. כל ספרור שכזה מייצג אשכול, וציר Y מייצג את מימד הזמן הכרונולוגי, כך שנקודת הופעתו של האשכול מייצגת את הנקודה בזמן בה הוא נוצר. אשכולות 1, 4, 5 ו-7 שבאיור, הם אשכולות שהתכונות שלהם לא השתנו במהלך כל משך הזמן המוצג באיור. להבדיל מהאשכולות האחרים שהחלו להתפצל ככל שהופיעו פרטים חדשים. הסקאלה בצד ימין של המסך מציגה את מספר הפרטים שתהליך ניתוח האשכולות בחן לאורך הזמן (72,608 פרטים), ובתחתית המסך מוצג מידע נוסף אודות כל אשכול. מידע המציין את מספר הפרטים בכל אשכול וכן וקטור עם הערכי המוקדים (Centroids) של כל אחת מהתכונות שמודדים.



Segment	Number of cases	Centroid
0.0	1000	[0.037,2.301,2.646,-0.009]
0.1.0	1000	[0.182,12.445,4.439,0.380]
0.1.1	1000	[0.644,7.093,3.059,-0.095]
.....	.....	.....
6.1.0	27	[6.594,9.287,3.975,12.412]
6.1.1.0	10	[1.118,14.352,0.150,16.026]
6.1.1.1	23	[1.096,12.435,6.021,16.449]
7	287	[7.938,15.490,8.167,-0.273]

איור 9. ExpanDrogram – הוספת מימד דינאמי לדנדרוגרמה (Khalemsky & Gelbard 2021)

האבחנה בקיומו של מימד דינאמי בתופעה הנחקרת, חשובה מאוד לצורך גיבוש האופן המתודי בו ראוי לנהל את ניתוח האשכולות. חשוב שלחוקר יהיה ברור אם ניתוח האשכולות שהוא מבצע אמור לשקף תמונת מצב סטאטית, או שמדובר בתופעה דינאמית שיכולה להשתנות במהלך הזמן. האם בעוד פרק זמן יכולים להופיע אשכולות חדשים ? האם אשכולות קיימים יוכלו להיעלם ? האם שיוכו של פרט לאשכול הוא שיוך לצמיתות ? או שפרט יכול לנוע בין אשכולות ?

ו. סיכום

מטרת מאמר זה היא לשפוך אור על הידע המתודי, להבדיל מהטכני, שאמור להיות בידי חוקר המשתמש בשיטות של ניתוח אשכולות וניתוח גורמים, כדי שיוכל לבקר באופן מושכל תוצר ניתוח שיוצג בפניו. המחשוב והבינה המלאכותית מספקים כיום כלים איכותיים וידידותיים, המאפשרים לכל משתמש, עם אוריינות טכנית בסיסית, להפעיל כלים של סיווג ושל ניתוח אשכולות. הדבר מאוד יעיל ואמין כשמדובר בטכניקות "יציבות" שמפיקות תוצאות אחידות בתנאי הפעלה שונים (כגון בדיקת מתאמים בין משתנים), אך הדבר מסוכן כאשר מדובר בטכניקות "לא יציבות" שנותנות תוצאות שונות בתנאי הפעלה שונים, ועל המשתמש להיות בעל הבנה שתאפשר לו לבקר באופן מקצועי את תוצר הניתוח שיתקבל בלחיצת כפתור. המאמר מדגים את חוסר היציבות האינהרנטי שקיים בתהליך ניתוח האשכולות, ומסב את תשומת ליבו של הקורא לחמש נקודות עקרוניות שעליו להיות מודע אליהן במהלך הניתוח ופירוש התוצאות.

א. ***היבט מדדי ההערכה***, שמאפשר כמעט לכל חלוקה למצוא מדד הערכה שיכריז עליה כעל החלוקה הטובה ביותר.

ב. ***היבט מדדי הדימיון-מרחק***, שמשפיעים באופן מהותי על ההיגיון המסדר של קיבוץ הפרטים לאשכולות. המאמר הציג הפניות ל-124 מדדים להערכת דימיון-מרחק בין פרטים. על מדדים אלו מתווספים מדדים להערכת דימיון-מרחק בין פרט וקבוצה, ולהערכת דימיון-מרחק בין קבוצות. על החוקר להיות מודע גם למהותם של מדדי הדימיון שנבחרו, גם לאופן בו ישוקלל השילוב של WGD ו-BGD, וגם לאופן בו תיבחר נקודת ה-"אופטימום" כך שכל תזוזה ממנה תייצג שילוב פחות טוב של WGD ו-BGD.

ג. ***היבט ייצוג המשתנים***, וההשלכות של צורת הייצוג על הדימיון שיחושב בין הפרטים. פרק ג מתייחס גם לנושא של ניתוח גורמים (Factor Analysis), ומסב את תשומת ליבו של החוקר לכך שניתוח גורמים הוא תהליך של ניתוח אשכולות המופעל על המשתנים (ולא על הרשומות), ולפיכך יש להיות מודעים לחוסר היציבות שיש גם בו.

ד. ***היבט האלגוריתם***, והתוצאה השונה שתתקבל מכל אלגוריתם גם אם כולם יופעלו על סט זהה לחלוטין של נתונים, תוך שימוש באותם מדדי דימיון. מכאן הצורך להיות מודעים לאופן הפעולה של כל אלגוריתם, ובו בזמן לשאול את השאלה האם לא נכון יותר להפעיל מגוון אלגוריתמים בגישת Ensemble learning, כפי שמציגה שיטת MAV, שבנוסף מאפשרת לבחון את חילוקי הדעות בין האלגוריתמים לגבי כל רשומה, וכן להמליץ על החלוקה בה יש הסכמה מירבית בין האלגוריתמים.

ה. ***דינאמיות/סטאטיות התופעה הנחקרת***, וגיבוש מתודולוגיה לניתוח האשכולות בהתאם. על החוקר לשאול את עצמו האם בעוד פרק זמן יכולים להופיע אשכולות חדשים? האם אשכולות קיימים יוכלו להיעלם או להתמזג עם אשכול אחר? האם שיוכו של פרט לאשכול הוא שיוך לצמיתות או שפרט יכול לנוע בין אשכולות?

תודות - לכל תלמידי המחקר שחקרו איתי את הסוגיות המעורבות בנושא: רן ביטמן, אביעד ברק, אנה חלמסקי, אביחי מגד, ורוני רמון גונן - הייתה לי הזכות וההנאה הגדולה לחקור ולפתח רעיונות חדשים יחד איתכם.

מקורות

Aaron, C.A. (12.8.2021). Lionel Messi's PSG Jersey Sold Out in Just 30 Minutes. *Hypebeast Kr*. Retrieved on 23.7.2025. https://hypebeast.com/2021/8/lionel-messis-psg-jersey-sold-out-30-minutes-info

Alvarez-Garcia, M., Arenas-Parra, M., & Ibar-Alonso, R. (2024). Uncovering student profiles. An explainable cluster analysis approach to PISA 2022. *Computers & Education*, 223, 105166.

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development*, 60(3), 383-398.

Ball, G., Hall, D., 1965. ISODATA, a novel method of data analysis and pattern classification. Technical report NTIS AD 699616. *Stanford Research Institute*, Stanford, CA.

Barak, A., & Gelbard, R. (2012). A decision support method, based on bounded rationality concepts, to reveal feature saliency in clustering problems. *Decision Support Systems*, *54*(1), 292-303.

Barak, A., & Gelbard, R. (2016). Classification by clustering using an extended saliency measure. *Expert Systems*, *33*(1), 46-59.

Bittmann, R. M., & Gelbard, R. M. (2007). Decision-making method using a visual approach for cluster analysis problems; indicative classification algorithms and grouping scope. *Expert Systems*, *24*(3), 171-187.

Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2), 1.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14(2), 241-258.

Erlich, Z., Gelbard, R., & Spiegler, I. (2002). Data mining by means of binary representation: a model for similarity and clustering. *Information Systems Frontiers*, 4, 187-197.

Erlich, Z., Gelbard, R., & Spiegler, I. (2003). Evaluating a positive attribute clustering model for data mining. *Journal of Computer Information Systems*, 43(3), 100-108.

Forgy, E.W., 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics* 21, 768–769.

Gelbard, R. (2013). Method and system for extended bitmap indexing. U.S. Patent No. 8,346,779. 1.1.2013.

Gelbard, R. (2013). "Padding" bitmaps to support similarity and mining. *Information Systems Frontiers*, *15*(1), 99-110.

Gelbard, R. (2019). Method and system for dynamic updating of classifier parameters based on dynamic buffers. U.S. Patent No. 10,268,923. 23 Apr. 2019.

Gelbard, R., Carmeli, A., Bittmann, R. M., & Ronen, S. S. (2009). Cluster analysis using multi-algorithm voting in cross-cultural studies. *Expert Systems with Applications*, *36*(7), 10438-10446.

Gelbard, R., Goldman, O., & Spiegler, I. (2007). Investigating diversity of clustering methods: An empirical comparison. *Data & Knowledge Engineering*, *63*(1), 155-166.

Gelbard, R., & Spiegler, I. (2000). Hempel's raven paradox: A positive approach to cluster analysis. *Computers & Operations Research*, *27*(4), 305-320.

Harel, D. (1987). Statecharts: A visual formalism for complex systems. *Science of computer programming*, 8(3), 231-274

Huang, L., Zhan, Y., & Ba, S. (2025). Modeling student teachers' self-regulated learning of complex professional knowledge: A sequential and clustering analysis with think-aloud protocols. *Computers & Education*, 233, 105310

Huff, D. (1954). *How to Lie with Statistics*. Norton & Company Inc., New York. https://www.horace.org/blog/wp-content/uploads/2012/05/How-to-Lie-With-Statistics-1954-Huff.pdf

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.

Kempen, R., Meier, A., Hasche, J., & Mueller, K. (2019). Optimized multi-algorithm voting: increasing objectivity in clustering. *Expert Systems with Applications*, 118, 217-230.

Khalemsky, A., & Gelbard, R. (2020). A dynamic classification unit for online segmentation of big data via small data buffers. *Decision support systems*, *128*, 113157.

Khalemsky, A., & Gelbard, R. (2021). ExpanDrogram: Dynamic Visualization of Big Data Segmentation over Time. ACM Journal of *Data and Information Quality*, 13(2), 1-27.

Lloyd, S., 1982. Least squares quantization, in *PCM. IEEE Trans. On Information Theory* 28, 129–137. Originally as an unpublished Bell laboratories Technical Note (1957).

MacQueen, J. (1967). Multivariate observations. *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281-297)

Meged, A., & Gelbard, R. (2011). Adjusting Fuzzy Similarity Functions for use with standard data mining tools. *Journal of Systems and Software*, *84*(12), 2374-2383.

Paulsen, L., & Lindsay, E. (2024). Learning analytics dashboards are increasingly becoming about learning and not just analytics-A systematic review. *Education and Information Technologies*, 29(11), 14279-14308.

Ramon-Gonen, R., & Gelbard, R. (2017). Cluster evolution analysis: Identification and detection of similar clusters and migration patterns. *Expert Systems with Applications*, *83*, 363-378.

Ronen, S., & Shenkar, O. (2013). Mapping world cultures: Cluster formation, sources and implications. *Journal of International Business Studies*, 44(9), 867-897.

Roski, M., Sebastian, R., Ewerth, R., Hoppe, A., & Nehring, A. (2024). Learning analytics and the Universal Design for Learning (UDL): A clustering approach. *Computers & Education*, 214, 105028

Sneath, P. H., & Sokal, R. R. (1962). Numerical taxonomy. *Nature*, 193, 855-860.

Sokal, R. R., & Sneath, P. H. (1963). Principles of numerical taxonomy.

Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci*, 1(804), 801.

Stojanov, A., & Daniel, B. K. (2024). A decade of research into the application of big data and analytics in higher education: A systematic review of the literature. *Education and information technologies*, 29(5), 5807-5831.

Türkmen, G. (2025). The review of studies on explainable artificial intelligence in educational research. *Journal of Educational Computing Research*, 63(2), 277-310.

Wijaya, S. H., Afendi, F. M., Batubara, I., Darusman, L. K., Altaf-Ul-Amin, M., & Kanaya, S. (2016). Finding an appropriate equation to measure similarity between binary vectors: case studies on Indonesian and Japanese herbal medicines. *BMC bioinformatics*, 17, 1-19.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1), 67-82.