מצגת 1-Introduction to data mining

חשיבות הכרית נתונים- מתוך כמות נתונים נפיק תועלת לארגון ע"י ניתוח של נתונים שלו מהנתונים מוצאים ידע, ברב המקרים נצפה מתוך נתונים קיימים.

למידה מתוך הנתונים- ML ללמוד מיתוך הנתונים.

הוספת נתונים, ניתוח, ולימוד מהנתונים.

מצגת 2- Exploratory Data Analysis

נתונים שאנחנו מקבלי מי שהוסיף אותם עושה טעויות והם לא תמיד שלמים הכוונה חלק מהם יכולים להיות קשורים למטרה שלנו וחלק לא, לכן צריך לנתח את הנתונים, מחפשים ממוצע חציון וכל מיני פרמטרים סטטיסטים וקורלציה בין הנתונים. Data set- Nורכב מתצפיות ומספר סידורי שלהם בכל תצפית יש משתנים של התצפית שהם העמודות, יכול להיות שקיים גם משתנה מטרה .

המשתנים קובעים את משנה המטרה Y.שהיא התשובה הסופית.

ישנם מס' דרכים לנתח את הנתונים- יהיה או לא יהיה, מה ההסתברות ומה הכמות.

בודקים לכל משתנה מהו – מהי המשמעות של כל משתנה , עד כמה ניתן לסמוך על אותו נתון.

ניתן ללמוד עבור כל משתנה האם ההתפלגות שלו היא סימטרית או לא. ניתן לראות לפי החציון האם ההתפלגות סימטרית. החציון והממוצע יהיו זהים כאשר ההתפלגות שווה.

בנוסף לומדים קורלציה – בעזרת גרף שבו נמצאים המשתנים – ומוצאים האם יש קשר ביניהם, אם יש קורלציה כל הנקודות יהיו סביב אותו קו. לומדים את אופי הקורלציה בין המשתנים.

מצגת 3 - Data Preparation

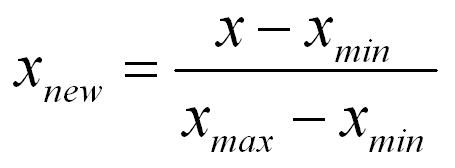
הכנה של הנתונים, לבדוק האם הנתונים שבידינו מתאים למשימה.

צריך לבדוק ולמנוע נתונים שחוזרים ולבדוק

מה נכון בנוגע להכנת נתונים?

נרמול לפי X ממוצע לא תמיד כדאי אלא אם כן ההתפלגות של הנתונים היא נורמלית

X-X   
--------  
D2

נרמול לפי 0-1 לא תמיד כדאי אלא אם המשתנים נתונים בהתפלגות אחידה

אם ההבדלים בין הנתונים הם מאוד גדולים קשה לנרמל את הנתונים לכן נהפוך אותם ללוגריתם.

מצגת 4-Learning from Data: Linear Regression

**רגרסיה לינארית**-ברגרסיה לינארית נשתמש כאשר המשתנה המוסבר, אותו אנו מבקשים לנבא הינו בסולם מדידה כמותי (כלומר כזה שניתן להמיר אותו לסקאלה מספרית בת 5 ערכים לפחות). בנוסף לכך, חשוב לבחון שהקשר בין כל אחד מהמשתנים המסבירים (בנפרד) לבין המשתנה המסביר יכול להיות מתואר באמצעות קו ישר (כלומר לינארי). דוגמא לכך יכולה להיות: ניבוי שביעות הרצון של עובד מהארגון (המשתנה המוסבר בסולם עולה של 0-100) באמצעות מספר הקשרים החברתיים בעבודה (המשתנה המסביר הראשון) והשכר שלו (המשתנה המסביר השני). בניית קו הרגרסיה לניבוי שביעות הרצון נעשה באמצעות בחינת[**קשרי פירסון**](http://www.pvalue.co.il/2011-09-08-20-50-09) בין המשתנים המסבירים לבין המשתנה המסביר. לרגרסיה הלינארית שני תוצרים עיקריים: הראשון הינו משוואת ניבוי מהצורה: Y=b0+b1x1+b2x2, כאשר Y מייצג את הציון המנובא (ולאו דווקא האמיתי) במשתנה המסביר, b0 מייצג את הקבוע של המשוואה (כלומר מהו ציון ה-Y המנובא בהינתן שערכי המשתנים המסבירים הינם אפס) ו-b1 מייצג את התרומה לערכים המנובאים של Y כאשר X1 גדל ביחידה אחת, וכל שאר המשתנים המסבירים קבועים במודל. בדוגמא שלנו, למשל אם קו הניבוי הינו Y=3+10\*friends+20\*salary. ניתן לראות כי הרמה ההתחלתית של שביעות הרצון במודל הינה 3, כאשר היא עולה ב-10 על כל חבר נוסף בעבודה ועולה ב-20 על כל שקל נוסף במשכורת. השני הינו טיב הניבוי של המודל כפי שמיוצג על ידי הגודל R², ומשמעותו איזה אחוז מ-Y מוסבר באמצעות שימוש במשתנים x1 ו-x2. ככל שאחוז השונות המוסברת גבוה יותר, וכן המודל מובהק (כלומר [**pvalue**](http://www.pvalue.co.il/articles1) נמוך מ-0.05) אזי ניתן לומר שמודל הרגרסיה שבנינו נוח להכללה לאוכלוסיה.

נחשב ל מודל פשוט ביותר, הולכים מהנחה שניתן לחשב y מx

אחרי שהנתונים עברו מיון סינון ונרמול לכל נתון מתקבל שגיאה

Ein -שגיאה לנתונים הקיימים.



Perceptron- סיווג

- הפרספטרון הוא [אלגוריתם](http://he.wikipedia.org/wiki/%D7%90%D7%9C%D7%92%D7%95%D7%A8%D7%99%D7%AA%D7%9D) ממשפחת ה[אלגוריתמים הלומדים](http://he.wikipedia.org/wiki/%D7%9C%D7%9E%D7%99%D7%93%D7%94_%D7%97%D7%99%D7%A9%D7%95%D7%91%D7%99%D7%AA). זהו אלגוריתם למידה "און ליין", משמע הוא לומד תוך כדי ריצה מדגימות ש[אבחן](http://he.wikipedia.org/wiki/%D7%90%D7%91%D7%97%D7%95%D7%9F) בזמן פעולתו. הפרספטרון הוא אלגוריתם סיווג, היינו, מטרתו היא להבדיל בין סוגים שונים של [דגימות](http://he.wikipedia.org/wiki/%D7%93%D7%92%D7%99%D7%9E%D7%94_(%D7%A1%D7%98%D7%98%D7%99%D7%A1%D7%98%D7%99%D7%A7%D7%94)) אותן הוא מקבל. לדוגמה, לאבחן בין [גברים](http://he.wikipedia.org/wiki/%D7%92%D7%91%D7%A8) ו[נשים](http://he.wikipedia.org/wiki/%D7%90%D7%99%D7%A9%D7%94) על פי מידת ה[נעל](http://he.wikipedia.org/wiki/%D7%A0%D7%A2%D7%9C) ו[אורך](http://he.wikipedia.org/wiki/%D7%90%D7%95%D7%A8%D7%9A) ה[שיער](http://he.wikipedia.org/wiki/%D7%A9%D7%99%D7%A2%D7%A8).

לכל נקודה יש שגיאה.

E1 = [yi – h (xi)]\*2

Learning אלגוריתם- אלגוריתם לשינוי משקלים כדי לקבל ein מינימלי.

מצגת 5- Machine Learning Theory

**הנחה ראשונה** קיימת פונקציה שבעזרת x אפשר לחשב y, ואם אין לנו x בטבלת הdata אז אין מה ללמוד.

אנו רוצים לבנות פונקציה שתהיה מתאימה לכל הx שקיימים.

באמצעות איזשהו אלגוריתם אנחנו בוחרים משקלות שאנחנו בוחרים את הפונקציה שלנו.

**הנחה שנייה** שבתוך הhypothesis set יש פונקציה שנותנת לחשב את הy בצורה איכותית

אם ההנחות האלו נכונות אנחנו יכולים לחשב את ה-f

0עושים התאמה של הw באמצעות השגיאה

EOUT- שגיאה בנקודות שעדיין אין לי בdata set (שגיאות שיהיו)

זה בעיה שלא יודעים EOUT ויודעים רק את הEIN.

במידה וn מספיק גדול אז eout יהיה קרוב לein

האם אם Ein יהיה מספיק קטן גם Eout?

תלוי בכמות הפרמטרים והנקודות – אומרים שכמות הנקדות צריכה להיות יותר גדולה מכמות הפרמטרים אומרים שכדאי בכמות הנקדות יהי ה פי 10 מכמות בפרמטרים.

Dvc-מורכבות המודל, כשהמודל יותר מורכב הdvc יותר גדול (כמות הפרמטרים)

Dvc / n = ההפרש בין ein ל eout

האיזון של מורכבות המודל:

.

**קיימות 2 סוגי שגיאות:**

Bias **1**– ניתן להקטין את השגיאה הזו משום שהיא נובעת מבחירת סוג המודל. את השגיאה של bias ניתן להקטין ע"י כך שנבחר מודל יותר מתאים.

Variance **2** – מעיד על רעש בנקודות EOUT את השגיאה הזו לאניתן לתקן

מצד אחד ישנה שגיאה שנובעת מכך שבחרנו מודל פשוט ולא נכון ושגיאה שנייה נובעת מכך שיש רעש ונקודות שלא מתאימות למודל האמיתי



מצגת 6- Linear Models

Nonlinear transforms

אם אין מספיק נקודות ביחס לנתונים לא כדאי להשתמש במודל זה

**רגרסיה לוגיסטית-** כאמור, אחד התנאים לרגרסיה לינארית הינו שהמשתנה המוסבר הינו בסולם כמותי. אולם מה נעשה אם המשתנה המוסבר הינו בסולם קטגוריאלי בעל שני ערכים בלבד? נשתמש במודל של רגרסיה לוגיסטית. למשל, אנו רוצים לנבא כוונת רכישה של iPad (ירכוש [1]/לא ירכוש [0] - המשתנה המוסבר) על פי גיל הלקוח (המשתנה המסביר הראשון) ועל פי מין הלקוח (גבר [1] /אישה [0] – המשתנה המסביר השני). מכיוון שבמודל זה לא מתקיימת הנחת הלינאריות בקשר מכיוון שהמשתנה המוסבר איננו כמותי, **מודל הרגרסיה הלוגיסטית נשען על חישוב סיכויים לערך 1 של המשתנה המוסבר**, סיכויים המכונים Odds Ratios או בקיצור OR. OR משמש כמקדמי המשתנים המסבירים ברגרסיה הלוגיסטית, כלומר לכל אחד מהמשתנים המסבירים מוצמד OR שונה שהפרשנות נעשית בהשוואה לערך 1. למה הכוונה? אם למשל OR של גיל הרכישה הינו 0.5, פירוש הדבר שכל שנה נוספת בגיל הלקוח מגדילה את הסיכוי לרכישה פי 0.5 (כלומר בפועל מקטינה אותו פי 2). אם למשל OR של מין הלקוח הינו 3, פירוש הדבר שלגברים (שערכם ברגרסיה 1) יש סיכוי גדול פי 3 לרכוש iPad בהשוואה לנשים. OR השווה ל-1, פירושו כי המשתנה המסביר אינו תורם בפועל לניבוי המשתנה המוסבר (כיוון שמכפילים את ערכו ב-1). OR נמוך מאחד פוגע בניבוי רכישה, ולעומת זאת OR גבוה מאחד תורם לניבוי רכישה (בהנחה ומקדמים אלו מובהקים מבחינה סטטיסטית).

מצגת 7- Classification (decision) tree

העיקרון:

\* זיהוי גורמים

**\* ז"א:** חיפוש תכונה/פרמטר המנבאת התנהגות

\* מיון ע"פ חשיבות

\* הצגה - **עץ**

עץ החלטה - הוא מודל חיזוי בתחומי ה[סטטיסטיקה](http://he.wikipedia.org/wiki/%D7%A1%D7%98%D7%98%D7%99%D7%A1%D7%98%D7%99%D7%A7%D7%94),[כריית נתונים](http://he.wikipedia.org/wiki/%D7%9B%D7%A8%D7%99%D7%99%D7%AA_%D7%A0%D7%AA%D7%95%D7%A0%D7%99%D7%9D) ו[הלמידה החישובית](http://he.wikipedia.org/wiki/%D7%9C%D7%9E%D7%99%D7%93%D7%94_%D7%97%D7%99%D7%A9%D7%95%D7%91%D7%99%D7%AA) המספק מיפוי בין תצפיות לערכים המתאימים עבורן.

עץ החלטה הוא אחד המודלים הפופולאריים והשימושיים ביותר בכריית נתונים מהווה גרסת אלגוריתם סיווגי.

עץ החלטה הוא [עץ בינארי](http://he.wikipedia.org/wiki/%D7%A2%D7%A5_%D7%91%D7%99%D7%A0%D7%90%D7%A8%D7%99) מלא המורכב מצמתי החלטה שבכל אחד מהם נבדק תנאי מסוים על מאפיין מסוים של התצפיות ועלים המכילים את הערך החזוי עבור התצפית המתאימה למסלול שמוביל אליהם בעץ.

סוגים של עצי החלטה הם עצי ריגרסיה שבהם מותאם ערך רציף לכל תצפית ועצי סיווג שבהם מותאם ערך בדיד או מחלקת סוג לכל תצפית. כמו כן קיימים עצי החלטה מסוג (CART (Classification And Regression Tree המשלבים את שני סוגי החיזוי.

יתרונות:

- אין הנחות חלוקה

- יכול להתמודד עם קלטים ריאליים ואיכותיים

- מהירות ויכולת רחבה

- חוסן לחריגים וערכים חסרים

- interpretability

- קומפקטיות של כללי סיווג

-היתרון הוא שזה פשוט ללימוד וקל להסביר איך זה עובד.

* ייצוג ויזואלי נוח
* קלים להבנה – התוצאה מוצגת בצורה בהירה ואינה מהווה קופסא שחורה
* אפשרות נוחה להפיכתם לנוהלי עבודה
* רמת דיוק גבוהה למדי
* תהליך הבניה קצר יותר מאשר בטכניקות אחרות
* אפשרות לטיפול בנתונים לא נומריים
* אינדיקציה ברורה מהם השדות הדומיננטיים לסיווג

חסרונות

עץ **מסועף** הדורש **פישוט**

- Interpretability?

- מספר פרמטרים כוונון להגדיר עם הדרכה קטנה

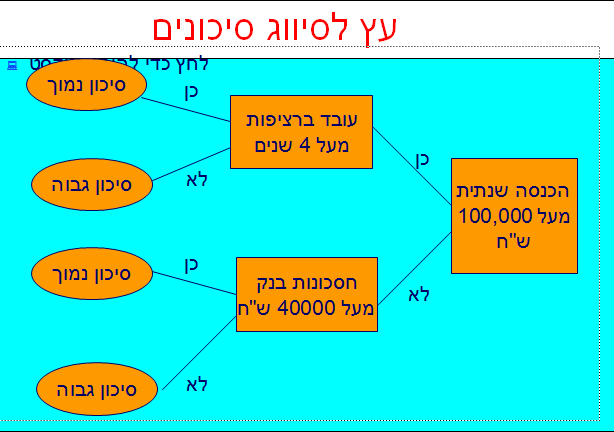
- גבול החלטה הוא לא רציף

-תלוי איך התחלנו והעצים יוצאים שונים. פתרון: random forests.

* מתאימים פחות למשימות הערכה עם משתנה רציף
* אינם נוחים לפתרון בעיות עתיות
* תמיכה במספר יחסית נמוך של משתנים
* שינוי קטן בפיצול הלוגי יכול לשנות את התוצאה בצורה דרסטית
* לעתים התוצאות שמתקבלות מאד מורכבות ונדרשים תהליכי פישוט
* יש מידע היסטורי נתון על הקבצות ידועות והמערכת מסייעת לבנות מודל לחיזוי ההקבצה של חברים חדשים
* תהליך בנית המודל כרוך בהפעלת טכניקות לימוד והסקת מסקנות מתוך ההקבצות הקיימות
* מספר ההקבצות קטן ובחלק ניכר מהבעיות כולל רק 2 מחלקות

תהליך הסיווג משמעותו בנית מודל לפיו יסווגו רשומות חדשות שאין להם עדיין סיווג ועדכון ה DB בהתאם

* המימוש מבוצע בדרך כלל בטכניקה של עצי החלטות או רשתות עצביות
* דוגמאות:
* סיווג רמת סיכון להענקת אשראי: סיכון גבוה /ממוצע/ נמוך
* סיווג תביעות לתביעה רגילה/חשודה
* ההסתברות ללקות במחלה מסוימת



בנייה של עץ החלטה, בנייה פשוטה מבחינת היגיון וקלה לבנייה אבל לא תמיד עובדת. כי אם קו הפרדה קצת מסובך (כמות הפרמטרים תהיה גדול כך שתהיה קשה ללימוד) אז צריך לעשות הפרדה שמאלה וימינה. תמיד אפשר לבנות עץ, אם הכמות של ההפרדות גדולה אז לפי תיאוריית הלימוד אז העץ לא ייצוגי.

הבנייה של העץ בנויה באופן רקורסיבי, כל פעם מחלקים לקטגוריות.

מבנה העץ אפילו תלוי מאיזה משתנה התחלנו, במידה והיינו מתחילים ממשתנה אחר היינו מקבלים עץ אחר.

מצגת 8- Neural Networks

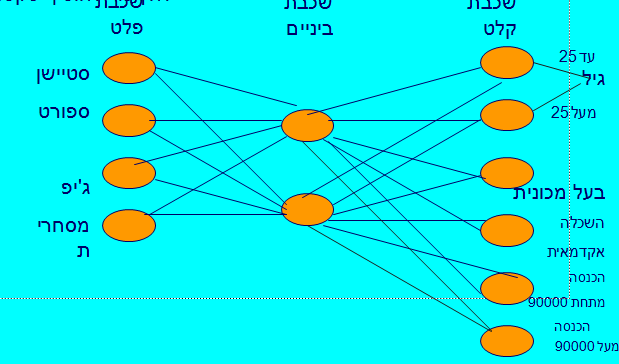
תוצאה של המודל הלינארי הקודם–כל נתון במודל זה הוא לינארריגרשיין בפני עצמו

The simplest model- the Perceptron



מודל מאוד פשוט ופופלארי החיסרון שלה שצריך המון נקודות

רשת נוירונים-

* מאפשרים ניתוח של בעיות מורכבות עם מאות פרמטרים
* זהו ניסיון לחיקוי מנגנון הלמידה של המוח
* המודל מבוסס על אוסף צמתים המחולקים לשכבת צמתי קלט, שכבת צמתי ביניים סמויה ושכבת צמתי פלט
* צמתי הקלט מקבלים את המשתנים, צמתי הפלט מציגים את התוצאה, צמתי הביניים משתתפים בתהליך הפתרון
* השימוש ברשת מורכב משני שלבים בשלב א , שלב הלימוד מלמדים את הרשת לפתור בעיה באמצעות אוסף דוגמאות בשלב ב משתמשים בפתרון לצורך סיווג
* האלגוריתם נקרא Back Propagation
* 

שלבים בשימוש ברשת:

השימוש ברשת מורכב משני שלבים:

\* שלב האימון בו מציגים לרשת אוסף דוגמאות מהעבר ומאמנים הרשת לפתור בעיות סיווג

\* שלב הפענוח בו מציגים מקרים חדשים ומסווגים אותם (מבצעים בהם פעולות חיזוי)

מעורבות המשתמש בתהליך:

* קביעת הקלטים
* קביעת הפלטים
* קביעת מספר הצמתים בשכבת הביניים
* הסבת נתוני הקלט לטווח 0-1
* קביעת גודל קבוצת הלימוד
* ניתוח וכוונון המודל

רשתות עצביות : תנאים הכרחיים להצלחה-

* הקלטים ברורים לחלוטין
* הפלטים ברורים לחלוטין
* קיים DB גדול עם ניסיון מצטבר איכותי של הפקת הפלטים כפונקציה של הקלטים
* יש לעדכן את המודל באופן תדיר
* קבוצת התרגול חייבת להיות איכותית

יתרונות-

* משמשים לפתרון מגוון רחב של בעיות
* נותנים פתרון טוב גם במקרים מסובכים
* מאפשרים טיפול הן בערכים רציפים והן בערכים בדידים
* רב חבילות התוכנה במגזר ה DM תומכות ברשתות עצביות

חסרונות-

* ערכי הקלט חייבים להיות בתחום 0-1
* לא ניתן להסביר במושגים עסקיים את התוצאות
* אם קבוצת המדגם נחותה הפתרון עלול להיות שגוי

# Overfitting. Regularization. Validation.מצגת 9-

התאמה לדאטא סט ספציפי שאינו מקטינה EOUT

Overfitting–במידה ויש לי יותר מידי פרמטרים ביחס למספר התצפיות ואני רוצה לעשות התאמה לנקודות בודדות אז עושים ש- ein=0. התאמה ל data set הספציפי שאינו עושה התאמה לeout. .

Overfitting- יעד פונקציה פשוט

Ein=0, Eout is huge

Overfitting הוא "התאמת הנתונים יותר מבטיח "בגלל התאמת הרעש.

ההתאמה היא מזיקה ולכן נשתמש ב: פתרונות: regularization ו –validation

Validation– עושה התאמה של למדא, בונה 3 סטים:

Regularization- הסדרה היא שיטה להימנע overfittingעל ידי הימנעות מזעור ein.

זה יכול להיעשות עם עצירה מוקדמת לחישוב W ע"י הקטנה הEIN.

# Unsupervised learning מצגת 10-

למצוא את הדמיון בין הנתונים

2 אלגוריתמים שימושים:

ניתוח אשכולות (Cluster Analysis) אינו מודל חיזוי, אך משמש ככלי יעיל לתיאור דפוסים בלתי ידועים במסד הנתונים. מטרת השיטה אהי לקבץ רשומות על-פי מידת יהד מיון ביניה . ן ישנן שיטות שונות ליישום הדבר ש , העיקריים הם K-Means ושיטות ההתגבשות (Agglomeration Methods .) המטרה היא למצוא את המוקדים שמצמצמים את המרחק בין חברי הקבוצה זה מזה, ומגדילים את המרחק של קבוצה אחת מהאחרת (Banning 2000, p. 42 ). בשיטת K-Means מוגדרים מראש מספר האשכולות, והמחשב מציב ממוצעים אקראיים למשתנים של כל אשכול. כל רשומה משוייכת לאשכול הקרוב ביותר אליה על-פי חישוב המרחק לאחר-מכן, כל 20 האוקלידי (Euclidean Distance) בין הממוצעים שלה לבין ממוצעי האשכול .ות מרכזי האשכולות מחושבים מחדש על- פי הממוצעים של הרשומות שהוצבו אליהם. כך התהליך ממשיך, עד שאין יותר שינויים בממוצעי האשכולות . בשיטות ההתגבשות, הרשומות מתגבשות אחת עם האחרת לפי המרחק ביניהם, כך שבכל שלב הגושים מתגבשים לגושים גדולים יותר , עד להיווצרותו של גוש אחד. תוצאת שיטה זו מוצגת בעץ מידרגי (hierarchical , ) שבו ניתן לבחור את השלב הרצוי שמייצג את מספר האשכולות רצוי. שיטה זו היא הנפוצה ביותר במחקרים הארכיאולוגיים והטיפולוגיי .ם לשיטת ניתוח אשכולות יתרונות רבים: יש לה יכולת גילוי מידע בצורה בלתי ישירה, היא עובדת היטב עם משתנים בדידים ורציפים והיא קלה ליישום. לעומת זאת, קשה לבחור בה סולם מדידה ומשקלים מתאימים לייצוג המשתנים. כל המשתנים שווים בעיניה, כך שהמשתנים החשובים באמת , לעיתים, אינם מקבלים ביטוי ייצוגי הולם. כמו- , כן יש לה רגישות גבוהה לפרמטרים של הרשומות הראשונות, ופעמים רבות קשה לפרש את משמעות האשכולות שמתגלים. בנוסף לכך, היא אינה עובדת היטב עם ערכים חסרים, וגם-אם מתבצעת ההשלמה של הערכים החסרים לפני כן, לא תמיד 21 הדבר עובד היטב  
בעיה נוספת היא ש , יכול להיווצר מצב ששני חברים באותו אשכול יהיו שונים מאוד זה מזה, ויוצבו לאותו אשכול רק מפני שיש להם תכונות משותפות עם חברים אחרים באות ו אשכול. קיימים 22 אלג ירו תמים שונים לפיתרון בעיה זו, אך הם לא פותרים אותה באופן מוחלט. בצומת ניתוח האשכולות של EM 5.1 לא ניתן לקבל תוצאה מידרגית, ולצערי לא יכולתי לבחון את יעילות שיטה זו מול שיטות הלא מידרגיות . היתרון הגדול של השיטה מבחינתנו היא הקיבוץ הבלתי סובייקטיבי שהיא עורכת לרשומות, ולכן השתמשתי בעיקר בה בנסיונותי ליצור טיפולוגיה. בצומת זה ישנה אפשרות לקבוע את מספר האשכולות הרצוי, אך גם ניתן לתת לתוכנה לבחור זאת באופן אוטומטי. כמו , כן- ישנן עוד אפשרויות רבות שלא אפרט עליהן במסגרת זו