**דוח רידוד מימדים, שיטות אישכול שונות ובדיקות סטטיסטיות**

מגישים: דוד ליפשיץ, עינב שנולד.

הדאטה: ספרות כתב יד, לקחנו רק את התמונות עם התיוגים 5,6,9. סה"כ עבדנו על 543 תמונות, 64 פיצ'רים עבור כל תמונה

תיאור תהליך השוואה בין מודלים:

הרצנו כל מדד כמותי על 10 subsets של הדאטה (דגמנו כל פעם 50% מהדאטה באופן ראנדומאלי)

השוואה בין הרצות עם ערכי hyper-parameters שונים (עבור כל בחירה של אלג' אישכול ואלג' רידוד מימדים):

* השוואה בעזרת מדד צללית\*
* במקרה והיו יותר מ2 בחירות אפשריות לערכים של הhyper-parameters לקחנו את 2 הבחירות שממוצע מדד הצללית עליהם היה מקסימלי
* הפעלת paired t-test –
  + אם היפותזת ה0 נדחתה לקחנו את הקונפיגורציה עם ממוצע ציוני הצללית הגבוה ביותר
  + אם ההיפותזה התקבלה – לקחנו את הקונפיגורציה הראשונה שבחנו עליה (לא הצלחנו להוכיח הבדל משמעותי בין 2 הבחירות, ולכן לא משנה איזו מהן ניקח)

מתוך השוואה זו, קיבלנו את הבחירה האופטימלית עבור כלל הhyper-parameters עבור כל בחירה של אלגוריתם אישכול ואלגוריתם רידוד ממדים.

השוואה בין אלגוריתמי רידוד ממדים בהינתן אלגוריתם אישכול:

* השוואה בעזרת מדד מידע משותף (mutual information)
* לקחנו את שני אלגוריתמי הרידוד שקיבלו את הממוצע (על המדד המשותף) הגבוה ביותר.
* השתמשנו בבחירות הhyper-parameters האופטימליות עבור כל אחד מהמקרים, בהתאם למה שהסקנו בהשוואות הקודמות.
* הפעלת paired t-test –
  + אם היפותזת ה0 נדחתה לקחנו את אלגוריתם הרידוד עם ממוצע ציוני המידע המשותף הגבוה ביותר
  + אם ההיפותזה התקבלה – לקחנו את אלגוריתם הרידוד הראשון שבחנו עליו (לא הצלחנו להוכיח הבדל משמעותי בין 2 הבחירות, ולכן לא משנה איזו מהן ניקח)

מתוך השוואה זו קיבלנו את הבחירה האופטימלית של אלגוריתם רידוד עבור כל אלגוריתם אישכול.

השוואה בין 3 אלגוריתמי אישכול:

* השוואה מול מדד מידע משותף (mutual information)
* עבור כל אלגוריתם אשכול, בחרנו באלגוריתם רידוד הממדים וערכי הhyper-parameters כך שייתן את התוצאה הטובה ביותר (לפי הסעיפים הקודמים)
* הפעלת anova test + ttest על השניים הכי גדולים
  + אם היפותזת ה0 נדחתה לשניהם לקחנו את אלגוריתם האישכול עם ממוצע ציוני המידע המשותף הגבוה ביותר
  + אם ההיפותזה התקבלה עבור אחד מהם – לקחנו את אלגוריתם האישכול הראשון שבחנו עליו (לא הצלחנו להוכיח הבדל משמעותי בין 2 הבחירות, ולכן לא משנה איזו מהן ניקח)

\*עבור קונפיגורציות שאלג' האישכול החזיר מחלקה יחידה, שמנו במדד הצללית ערך שלילי גדול (1000-) כיוון שמדד הצללית לא עובד במקרים כאלה ואם קיבלנו רק מחלקה יחידה ביצועי האלגוריתם מבחינתנו גרועים ביותר

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Kmeans parameters:

N\_clusters = 3

השוואה בין הייפר פרמטרים:

בדיקת השפעה של הרצת KMEANS בקונפיגורציות שונות על מדד צללית (עבור כל אלגוריתם רידוד מימדים בדקנו האם יש השפעה לכך שמריצים את אלג' האישכול לפני ביצוע רידוד מימדים או אחרי)

|  |  |  |  |
| --- | --- | --- | --- |
| Dim reduction algs | Dim reduction before clustering | Dim reduction after clustering | PVALUE |
| PCA | 0.268 | 0.27 | 0.000237 |
| CMDS | 0.264 | 0.27 | 0.0024 |
| LLE | 0.185 | 0.27 | 0.0000268 |
| ISO | 0.268 | 0.27 | 0.02 |
| EIGEN-MAPS | 0.267 | 0.27 | 0.005 |

(התוצאות מתבססות על האקסלים הבאים שמכילים את המדדים הכמותיים עבור כל הרצה: )

השוואה בין אלגוריתמי רידוד (mutual information)

|  |  |  |
| --- | --- | --- |
| PCA | ISO | PVALUE |
| 0.7712 | 0.771 | 0.343 |

(התוצאות מתבססות על האקסל הבא שמכיל את המדדים הכמותיים עבור כל הרצה: )

מסקנה – אין הבדל משמעותי בין שני אלגוריתמי הרידוד שעבורם ממוצע הMI הוא מקסימלי לכן לקחנו את PCA

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Hierarchical clustering parameters:

N\_clusters = 3

השוואה בין הייפר פרמטרים:

בדיקת השפעה של הרצת אשכול היררכי בקונפיגורציות שונות על מדד צללית (עבור כל אלגוריתם רידוד מימדים בדקנו האם יש השפעה לכך שמריצים את אלג' האישכול לפני ביצוע רידוד מימדים או אחרי)

|  |  |  |  |
| --- | --- | --- | --- |
| Dim reduction algs | Dim reduction before clustering | Dim reduction after clustering | PVALUE |
| PCA | 0.262 | 0.257 | 0.09 |
| CMDS | 0.253 | 0.257 | 0.33 |
| LLE | 0.175 | 0.257 | 0.0015 |
| ISO | 0.252 | 0.257 | 0.17 |
| EIGEN-MAPS | 0.263 | 0.257 | 0.099 |

(התוצאות מתבססות על האקסלים הבאים שמכילים את המדדים הכמותיים עבור כל הרצה: )

השוואה בין אלגוריתמי רידוד (mutual information)

|  |  |  |
| --- | --- | --- |
| LLE | EigenMaps | PVALUE |
| 0.842 | 0.801 | 0.131 |

(התוצאות מתבססות על האקסל הבא שמכיל את המדדים הכמותיים עבור כל הרצה: )

מסקנה – אין הבדל משמעותי בין שני אלגוריתמי הרידוד שעבורם ממוצע הMI הוא מקסימלי לכן לקחנו את LLE

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

DBSCAN parameters:

Eps = 0.0007, min\_samples=5

DBSCAN parameters:

Eps = 10, min\_samples=22

DBSCAN parameters:

Eps = 0.01, min\_samples=10

DBSCAN parameters:

Eps = 4, min\_samples=8

DBSCAN parameters:

Eps = 1, min\_samples=19

השוואה בין הייפר פרמטרים:

עבור dbscan בדקנו, האם עדיף לבצע אישכול על הדאטה לפני או אחרי ביצוע רידוד המימדים, ערכי eps, וmin\_samples

"PCA":

Config1: "eps": 1.5, "min\_samples": 25, Config2: "eps": 1.5, "min\_samples": 20

"CMDS":

Config1: "eps": 5, "min\_samples": 7, Config2: "eps": 5, "min\_samples": 8

"LLE":

Config1: "eps": 0.01, "min\_samples": 5, Config2: "eps": 0.01, "min\_samples": 10

"ISO”:

Config1: "eps": 12, "min\_samples": 20, Config2: "eps": 10, "min\_samples": 10

"EigenMaps":

Config1: "eps": 0.0007, "min\_samples": 8, "eps": 0.0007, "min\_samples": 5

|  |  |  |  |
| --- | --- | --- | --- |
| Dim reduction algs | Config1 | Config2 | PVALUE |
| PCA | 0.219 | 0.193 | 0.03 |
| CMDS | 0.0054 | -0.021 | 0.016 |
| LLE | -0.098 | -19999 | 0.168 |
| ISO | 0.204 | 0.173 | 0.01 |
| EIGEN-MAPS | 0.173 | 0.14 | 0.011 |

(התוצאות מתבססות על האקסלים הבאים שמכילים את המדדים הכמותיים עבור כל הרצה: )

השוואה בין אלגוריתמי רידוד (mutual information)

|  |  |  |
| --- | --- | --- |
| EigenMaps | LLE | PVALUE |
| 0.787 | 0.695 | 0.0003 |

(התוצאות מתבססות על האקסל הבא שמכיל את המדדים הכמותיים עבור כל הרצה: )

מסקנה – יש הבדל משמעותי בין שני אלגוריתמי הרידוד עבורם הMI הוא מקסימלי. כיוון שממוצע הMI של EigenMaps גבוה יותר, לקחנו אותו.

השוואה בין אלגוריתמי אשכול:

|  |  |  |  |
| --- | --- | --- | --- |
| אשכול היררכי | DBSCAN | KMeans | PVALUE |
| 0.842 | 0.787 | 0.771 | 0.02 |

(התוצאות מתבססות על האקסל הבא שמכיל את המדדים הכמותיים עבור כל הרצה: )

מסקנה - יש הבדל משמעותי בין אלגוריתמי האשכול. כיוון שממוצע הMI על אשכול היררכי הוא הגבוה ביותר, בחרנו בו להיות אלגוריתם האשכול עם הביצועים הטובים ביותר.