****

**האוניברסיטה הפתוחה**

**המחלקה למתמטיקה ולמדעי המחשב**

סמינר בנושא מיוחד במדעי המחשב

20375

שיפור איחזור תמונה מבוסס תוכן  
בעזרת למידה חישובית

**העבודה הוכנה על-ידי:** שרי קריקשר 300481793

**בהדרכתה של** **ד"ר** מיה הרמן

**תאריך ההגשה:** 15.06.2015

תוכן עניינים

1. מבוא
2. רקע
   1. מהו איחזור תמונה מבוסס תוכן CBIR))
   2. עקרונות בסיסים באיחזור תמונה מבוסס תוכן CBIR))
      1. מבנה בסיסי למערכת CBIR
      2. רמות איחזור
      3. אומדן טיב האיחזור
   3. למידה חישובית
      1. למידה מפוקחת
      2. למידה בלתי מפוקחת
      3. למידה על ידי חיזוק
   4. מציאה וייצוג מאפיינים עבור תמונה
3. שיפור וייעול בעזרת עיבוד מקדים
   1. הקטנת מרחב המאפיינים
      1. חילוץ מאפיינים
      2. בחירת מאפיינים
4. שיטות שיפור CBIR בעזרת למידה חישובית
   1. שיטות למידה מפוקחת
   2. שיטות למידה בלתי מפוקחת
5. בחינה לעומק : Supervised SVM
   1. משוב רלוונטיות (RF) לאיחזור תמונה (CBIR) בעזרת מכונת ווקטורים תומכים (SVM) ובחירת מאפיינים
      1. תיאור
      2. ניתוח תוצאות
      3. מסקנות
   2. שיפור איחזור תמונה (CBIR) כאשר דוגמאות התיוג מרעישות
      1. תיאור
      2. ניתוח תוצאות
      3. מסקנות
6. הצגת מימוש : Supervised SVM
7. סיכום
8. מקורות

מבוא

תוכן מולטימדיה כיום, נמצא במגמת עלייה מתמדת. כמות התוכן הוויזאלי ותמונות בפרט אינם פרקטיים לשימוש באיחזור מבוסס טקסט שכן הוא תלוי בעיקר בתיוג ידני או נתוני מתאר וקצרה היריעה לעמוד בקצב גידול התמונות המסחרר.  
 בניגוד לאיחזור טקסטואלי, איחזור תמונה מבוסס תוכן הוא תהליך חיפוש ואיחזור תמונות מתוך מאגר תמונות בהתבסס על המידע הוויזואלי בתמונה. לאיחזור מבוסס תוכן שימושים קיימים רבים בינהם במערכות רפואיות, מערכות צבאיות ומשטרתיות מערכות ניהול זכויות יוצרים ועוד ועוד. אולם איחזור מבוסס תוכן מציג אתגרים לא פשוטים בניתוח, בשמירה ובאיחזור. למידה חישובית היא אחת הדרכים להתמודד עם בעיות אלו שכן למידה חישובית תסייע עם סביבות שהן דינמיות ולא ידועות במלואן ולהשיג בשאיפה שיפור ביצועים בהתבסס על נסיון העבר.   
 בעבודה זו נבחן דרכים לשפר איחזור מבוסס תוכן מעולם הלמידה החישובית, נציג מספר דרכים בקצרה ונבחן לעומק שיטה בשם מכונת ווקטורים תומכים.   
תחילה נסקור את הרקע הנדרש, בפרק 2 נציג הגדרות בסיסייות באיחזור מבוסס תוכן, מבנה בסיס של מערכת, קטגוריות למידה חישובית וסוגי מאפיינים לחילוץ מתמונה.  
בפרק 3 נציג דרכים להקטנת מרחב המאפיינים בעזרתן ניתן להקל על הניתוח השמירה והאיחזור של התמונות. פרק 4 מציג לעומק שיטות למידה חישובית בשתי קטגוריות הלמידה העיקריות (למידה מפוקחת ולמידה בלתי מפוקחת). פרק 5 מציג את שיטת מכונת ווקטורים תומכים לעומק ובפרק 6 מוצגת מערכת העושה שימוש בשיטה זו.

1. רקע
   1. **מהו איחזור תמונה מבוסס תוכן CBIR))**  
      איחזור תמונה מבוסס תוכן (Content based image retrieval) היא אחת הטכניקות למימוש איחזור תמונה, שהוא תהליך חיפוש ואיחזור תמונות מתוך מאגר תמונות, לרוב נרחב.  
      בשונה מטכניקות אחרות לאיחזור תמונה שברובן מבוססות נתוני מתאר (metadata) כגון תיוג או תיאור טקסטואליים, איחזור תמונה מבוסס תוכן מתייחס למידע הוויזואלי הגלום בתמונה. מידע זה למשל מאפייני צבע, צורה וכד' משמשים לבניית ייצוג התמונות ולייצירת אינדקס חיפוש.   
      דוגמאות לשימושים – מערכות רפואיות, מערכות צבאיות ומשטרתיות מערכות ניהול זכויות יוצרים ועוד.

מציאת מאפיינים (צבע, צורה, טקסטורה וכד')

הקטנת מרחב מאפיינים

שמירה במאגר

מאגר  
מאפיינים

השוואת מאפיינים

החזרת תוצאות

משוב רלוונטיות

עיבוד מקדים

מאגר תמונות

תמונה לחיפוש



* 1. **עקרונות בסיסים באיחזור תמונה מבוסס תוכן CBIR))** 
     1. מבנה בסיסי למערכת CBIR :  
        טיפול במאגר (שלב האינדוקס)

1. שלבי עיבוד מקדים לתמונות המאגר.
2. מציאת מאפיני התמונות.
3. הקטנת מרחב מאפייני התמונות.
4. שמירה ואינדוקס במאגר נתונים   
   (למשל בסיס נתונים).

טיפול בשאילתה (שלב השליפה)

1. עיבוד מקדים.
2. מציאת מאפיין התמונה.
3. הקטנת מרחב המאפיינים.
4. בחינת דמיון וקירבה מול אוסף   
   המאפיינים במאגר הנתונים.
5. החזרת תוצאות.

איור 1. מערכת CBIR, מבנה בסיסי

1. קבלת משוב מהמשתמש למידת ההתאמה.
   * 1. רמות איחזור  
        את תהליך האיחזור ניתן לסווג לשלוש רמות מורכבת :
2. איחזור לפי מאפיינים *פרימיטיביים* למשל צבע, צורה, טקסטורה ומיקום מרחבי, לדוגמא "מצא את כל התמונות המכילות כוכבים צהובים או כחולים המסודרים במעגל".
3. איחזור לפי תכונה הניתנת לגזירה מהמאפיינים הפרימיטיביים או מאפיינים *לוגיים*, לצורך כך נדרשת יכולת הסקה מסויימת בהקשר לאובייקטים בתמונה, לדוגמא "מצא תמונות רכבת נוסעים החוצה גשר".
4. איחזור לפי תכונות *מופשטות*, באמצעות הסקת מסקנות מתקדמת על בסיס האובייקטים או ההתרחשות בתמונה לדוגמא "מצא תמונות המציגות טקסים ותהלוכות".

צורך המשתמש לרוב הוא להגדיר את מושא החיפוש בצורה של הרמה השלישית, אולם כבר מהצגת הרמות השונות ניתן לראות כי קיים פער בין ייצוגה הממוחשב של תמונה, המושתת על המידע הגולמי – הפיקסלים המרכיבים את התמונה, לבין מה שאדם מבין מתוך התמונה, הפער הזה נקרא הפער הסמנטי. למידה חישובית יכולה לסייע בהקטנת פער זה ולתרום לשיפור מערכות איחזור נתונים.

* + 1. אומדן טיב האיחזור  
       במקרה של איחזור, כל תוצאה יכולה להיות "רלוונטית" או "לא רלוונטית". כלומר רלוונטיות התוצאה נמדדת בצורה בינארית. עבור כל תמונה נבחן האם חזרה או לא בהתאם לחיפוש והאם היא רלוונטית או לא בהתאם לחיפוש. לצורך כך מגדירים את שני המדדים הבאים :

דיוק (Precision) – מספר התוצאות הרלוונטיות שחזרו ביחס לכלל התוצאות שחזרו. ובכך דיוק הוא מדד לאפקטיביות התוצאות וסינון הלא רלוונטי.

החזר (Recall) – מספר התוצאות הרלוונטיות שחזרו ביחס למספר התוצאות הרלוונטיות הכולל. ובכך דיוק הוא מדד לאפקטיביות האיחזור.  
  
מכך נגזרות ארבעה מחלקות :  
רלוונטי אמיתי – התוצאה אכן רלוונטית   
לא רלוונטי אמיתי – התוצאה אכן לא רלוונטית  
רלוונטי שקרי – התוצאה לא רלוונטית, שגיאת איחזור  
לא רלוונטי שקרי – התוצאת רלוונטית, פיספוס

False negatives

True negatives

True positive

Relevant images

Retrieved  
images

* 1. **למידה חישובית**

למידה חישובית היא אחת הדרכים להתמודד עם סביבות שהן דינמיות ולא ידועות במלואן, בהן למתכנן המערכת לא הייתה היכולת להתחשב בכל התנאים והדרישות בשלב הכתיבה או שלא ניתן לבנות מערכת חוקים של סיבה ותוצאה. כמו כן באמצעות למידה ניתן להשיג בשאיפה שיפור ביצועים בהתבסס על נסיון העבר.

איור 2. מחלקות תוצאות איחזור

את סוגי הלמידה ניתן לחלק לשלושה סוגים עיקריים :

* + 1. למידה מפוקחת (supervised learning)  
       מתבססת על סט דוגמאות של קלט-פלט (כלומר נתונה תוצאה \ תיוג עבור דוגמאות הקלט) הנקראת קבוצת אימון, מתוכו נבנית פונקציית מיפוי \ סיווג עבור קלטים חדשים.   
       אלגוריתמים נפוצים – עצי החלטה, מכונת וקטורים תומכים (SVM) ועוד.

דוגמאות לשימוש – סינון ספאם, דירוג תוכן וכד'.

* + 1. למידה בלתי מפוקחת (unsupervised learning)  
       מתבססת על סט דוגמאות קלט, ללא פלט או משוב, בעזרת הדוגמאות יש להסיק תבניות של קלט\פלט ובכך תתאפשר חלוקה של קלטים חדשים.  
       אלגוריתמים נפוצים – K-means, רשת קוהנון (SOM) ועוד.

דוגמאות לשימוש – קיבוץ (Clustering) וכד'.

* + 1. למידה על ידי חיזוק (reinforcement learning)

מתבססת ע"י סדרת חיזוקים חיובים או שליליים (תגמול\עונש)

אלגוריתמים נפוצים – MDP  
דוגמאות לשימוש – מערכת רובוט לניווט במרחב ועוד.

* 1. **מציאה וייצוג מאפיינים עבור תמונה**מאפיין בתמונה הוא פיסת מידע הנלקח מתוך התמונה המקורית ומשמש בסיס לתהליכי עיבודה.   
     מאפייני צבע -   
     מאפייני צבע הם בין הנפוצים ביותר בשימוש באיחזור תמונה, בעיקר בשל ההתאמה הגבוהה עם האובייקטים בתמונה. יתר על כך מאפייני צבע נשמרים גם בעקבות שינויי רקע, גודל, כיוון, וכו'...כמו כן ישנם מספר מודלי צבע רבים, אולם  
     HSV (Hue, Saturation, Value), HLS(Hue, Lightness, Saturation), ו- CIE נמצאו יעילים יותר לעומת RGB ומודלים אחרים בהשוואת דמיון בין תמונות.  
     היסטוגרמת צבע היא אחד ממאפיין הצבע השימושיים ביותר באיחור תמונה, כיוון שנמצא מועיל ביותר בתיאור חלוקת הצבעים בתמונה. בעזרת היסטוגרמת הצבע ניתן למצוא תמונות דומות מבחינת התפלגות הצבע.  
     מאפייני מרקם (טקסטורה) -   
     בעוד אין הגדרה מדוייקת למה היא טקסטורה, ניתן להתייחס אליה בתור חזרה של תבנית על גבי איזור בתמונה. ישנם שיטות רבות למדידת דמיון טקסטורה,   
     אחת השיטות הבולטות הינה מטריצת מאורעות משותפת (co-occurrence matrix), היא למעשה ייצוג מתמטי לתלויות המרחביות בין גווני אפור בטקסטורת התמונה. מטריצה זו ניבנת על המגמה והמרחק בין פיקסלים בתמונה ומתוך מטריצה זו ניתן לחלץ נתונים סטטיסטיים שייצגו את הטקסטורה למשל אנטרופיה, אנרגיה, ניגודיות והומוגניות.  
     כמו כן שיטות עיבוד אותות כגון התמרת Wavelet (Wavelet transforms) ובפרטGabor filters נמצאו שימושיים ביותר בניתוח טקסטורה.  
     מאפייני צורה -   
     באופן גס, ניתן להגדיר צורה בעזרת תיאור האובייקט ללא גודלו כיוונו ומיקומו. על נדרש ממאפיין צורה להיות עמיד לשינויי הזזה, סיבוב ושינויי גודל שכן סידור האובייקטים בתמונה אינו ידוע מראש. על מנת לבנות מאפיין צורה יש צורך בחלוקת התמונה למקטעים על מנת להבחין באובייקטים או בגבולות איזורים בתמונה. ניתן לחלק את השיטות לאיבחון צורה לשתי קטגוריות :  
     שיטות מבוססות גבול המשתמשות במתאר החיצוני של צורת האבייקט ושיטות מבוססות איזורים המשתמשות בכל איזור הצורה.  
     מאפייני מרחב (טופולוגיה) -תמונה דיגיטלית יכולה להיות מיוצגת ע"י אחת או יותר תוכונות טופולוגיות, תכונות אלו הן עמידות לשינויי הזזה, סיבוב ושינויי גודל. דוגמא לתכונה אחת כזו היא מספר אויילר (Euler number) המוגדר להיות ההפרש בין מספר הרכיבים הקשורים, לבין מספר החורים בתמונה כלומר .

H=1, C=1, E=0

(א)

H=2, C=1, E=-1

(ב)

איור 3. דוגמאות לחישוב מספר אויילר בתמונות בינאריות (שחור \ לבן)

ניתן להרחיב שימוש במספרי אויילר לווקטורי אויילר עבור תמונות בגווני אפור.

1. שיפור וייעול בעזרת עיבוד מקדים

|  |  |  |
| --- | --- | --- |
| Unsupervised | Supervised |  |
| PCA | LDA | Feature  Extraction |
| Category Utility, Laplacian Score Q-α | Feature  Subset Selection (Filters, Wrappers) | Feature  Selection  טבלה 1. החלוקה העיקרית עבור תהליך הקטנת מרחב המאפיינים [1] |

* 1. **הקטנת מרחב המאפיינים (dimension reduction)**תוכנים לא טקסטואלים וחזותיים בפרט הינם בדרך כלל מסדר מורכבות גבוה, כך שלעיתים הקטנת סדר המורכבות הוא שלב הכרחי בתהליכי ניתוח ומיצוי וכן בשימוש באלגוריתמי למידה. תהליך הקטנת סדר המורכבות של המידע שיש לעבד יכול להועיל מבחינה חישובית (הקטנת סיבוכיות החישוב), מבחינת דיוק ניתוח ומיצוי המידע ובפישוט ייצוג מודל הנתונים (ובכך להביא לצימצום מקום איחסון הנתונים).  
      תהליך הקטנת מרחב המאפיינים דורש לרוב זיהוי ייצוג מתאים של המידע שהינו מסדר נמוך יותר מהייצוג המקורי המורכב. את תהליך הקטנת מרחב המאפיינים ניתן לחלק באופן הבא :
     1. **חילוץ מאפיינים**   
        חילוץ או התמרת מאפיינים מבוססת על המרת המאפיינים מהייצוג המקורי שהינו בעל סדר גבוה לייצוג חלפי בו סדר המימד נמוך יותר בשאיפה לשמר כמה שיותר מידע מהייצוג המקורי. בהינתן סט דוגמאות הבנוי מ- דוגמאות אימון, כל דוגמא מתוארת ע"י קבוצת מאפיינים ובסך הכל ישנם אובייקטים המתוארים ע"י מאפיינים. ניתן לייצג זאת בעזרת מטריצת אובייקט למאפיין כך שכל עמודה מייצגת אובייקט.  
        מטרת התהליך היא להמיר את סט המאפיינים לסט   
        מאפיינים אחר כך ש וכתוצאה מכך מומר ל .   
        הדרך להמרה הזו היא בדרך כלל העתקה לינארית המקיימת כאשר הוא ממימד ו- ממימד .  
          
        את שיטות חילוץ המאפיינים ניתן לחלק לשתי תתי קטגוריות :
* השיטות העיקריות עבור למידה בלתי מפוקחת –  
  ניתוח גורמים ראשיים (Principal Component Analysis – PCA)  
  שיטה זו ממירה את המידע למרחב ממימד בעל סדר נמוך יותר המשמר כמה ניתן את *שונות המידע*. השיטה מחפשת העתקה לינארית הממירה את סט הדוגמאות הבנוי ממאפיינים שייתכן ותלוים לנארית זה בזה (אלה הם צירי המערכת המקורית) לסט הבנוי ממאפיינים שאינם תלוים לינארית (צירי המערכת החדשה), אלו נקראים גורמים ראשיים ומספרם קטן שווה למספר המאפיינים המקורי. השיטה בוחרת את הגורמי הראשיים הראשונים (בהם השונות הינה מקסימלית) ובכך מימד המידע המוטל קטן ממימד המקור.   
  ניתן לחשוב על השיטה בתור התאמת אליפסואיד על המידע כאשר צירי האליפסואיד מתאימים לגורמים הראשיים, אם אחד הצירים קצר יותר אז שונות המידע בציר זה קטנה גם היא ולכן השמטת גורם זה תגרור איבוד מידע מועט יחסית.   
  הבסיס לשיטה זו היא מטריצת השונות המשותפת[[1]](#footnote-1) : ובה הערכים באלכסון מייצגים את השונות[[2]](#footnote-2) של כל מאפיין ושאר הערכים מייצגים את השונות המשותפת של זוגות המאפיינים המתאימים.  
  מטרת ההעתקה ב- PCA היא לאפס את ערכי השונות המשותפת, כלומר ללכסן את המטריצה ובכך נקבל את המטריצה . את המידע ממירים בעזרת חישוב כאשר השורות של הם הווקטורים העצמיים של :

*f1*

*f2*

*f'1*

*f'2*

איור 4. PCA עבור מרחב מאפיינים דו מימדי, המומר ל ו- כך שהשונות בכיוון היא המקסמלית

הרכיב ה – המופיע באלכסון במטריצה מייצג את שונות המידע בכיוון הגורם הראשי המתאים. ההקטנה של מרחב המאפיינים מושגת ע"י השמטה של הגורמים הראשיים השוליים יותר, כלומר ל מימד כאשר הוא מספר הגורמים הראשיים שברצוננו לשמור.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| pth feature | … | 2nd feature | 1st feature |  |
|  |  |  |  | Image 1 |
|  |  |  |  | Image 2 |
|  |  |  |  | … |
|  |  |  |  | Image n |
| n \* p | | | |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| kth feature | … | 2nd feature | 1st feature |  |
|  |  |  |  | Image 1 |
|  |  |  |  | Image 2 |
|  |  |  |  | … |
|  |  |  |  | Image n |
| n \* k | | | |  |

feature 2

feature 1

feature 3

feature 4

feature 5

feature 6

C1

C2

Ck

C3

…

feature p

k < p

Ck

C1

C2

C3

…

איור 5. מתאר PCA במרחב מאפיינים מורכב יותר, הייצוג החדש של המידע הינו מתוצמת יותר

* השיטות העיקריות עבור למידה מפוקחת –  
   ניתוח לנארי מבחין (Linear Discriminant Analysis –LDA )  
  שיטתLDA לוקחת בחשבון את התיוג של קבוצת דוגמאות האימון ומחפשת דרך להדגיש את *החלוקה* במידע, הנובעת מן התיוג, בשונה מ-PCA המחפשת דרך לייצג את המידע בצורה הממקסמת את השונות.  
  אם כן יש לחפש הטלה הממקסמת את החלוקה בין המחלקות וממזערת את החלוקה הפנימית בתוך המחלקות.   
  לשם כך מגדירים את מטריצות הפיזור (בגודל , כל איבר במטריצה הוא ווקטור בגודל ) הבאות : עבור החלוקה בין מחלקות, ו- עבור החלוקה בתוך מחלקות  
  כאשר היא קבוצת המחלקות הנובעות מהתיוג,   
   הוא מספר הדוגמאות השייכות למחלקה   
   הוא הערך הממוצע של כל הדוגמאות ( וגודלו )  
   הוא הממוצע של כל הדוגמאות של המחלקה ( וגודלו )  
  הדרך למקסם את החלוקה בין המחלקות למזער את החלוקה הפנימית בתוך המחלקות ניתנת לביצוע ע"י קרטריון פישר :  
  כלומר יש למצוא כך שהמנה תקבל ערך מקסימלי[[3]](#footnote-3), את ערכו של יש לקבוע בהתאם לצורך, ובכך מתבצעת הקטנת מרחב המאפיינים אולם מגדירים בדרך כלל .  
   את ניתן להרכיב מהוקטורים העצמיים של ולשם כך יש לבצע ליכסון סימולטני, כלומר לפתור את המערכת הבאה :  
  כאשר המטריצה היא היא מטריצה אלכסונית, שערכי הם הערכים עצמיים הפותרים את :   
  בשיטה זו יש להזהר ממצב של Overfitting ולגרום לבחירה של ערכים מקריים או "רעש" לעומת ערכים מבחינים יותר.

**PCA**

**LDA**

איור 6. ניתן לראות כי PCA לא בהכרח יפיק חלוקה טובה של המידע כאשר הוא מחולק לקטגוריות, לעומת LDA המחפש הטלה שתדגיש את החלוקה [2]

* + 1. **בחירת מאפיינים**  
       בשונה מחילוץ מאפיינים, בחירת מאפיינים מחפשת את תת קבוצת המאפיינים המינימלית והטובה ביותר של המאפיינים המקוריים. למעשה המאפיינים, חלקיים אומנם, נשארים בצורת המקורית ולכן גם קלים יותר לפירוש. שיטה זו מוגדרת היטב עבור למידה מפוקחת ועבור למידה בלתי מפוקחת שיטה זו אינה כה ברורה.  
       שיטות בחירת מאפיינים מבוססות אחת משלושה גישות :

1. שיטות פילטור – בגישה זו מנסים להשמיט מאפיינים לא רלוונטיים *בשלב שלפני אימון אלגוריתם הלמידה*, תחילה המידע מנותח על מנת להבחין אילו מאפיינים הכרחיים על מנת לתאר את המידע.
2. שיטות מעטפת – בגישה זו *משתמשים באלגוריתם הלמידה* על מנת לבחור את סט המאפיינים הדרוש. מתבצעת בחינה של תתי קבוצות מאפיינים ע"י הערכת הדיוק של תת הקבוצה שנבחנת, מתוכן נבחרת שעבורה הדיוק הוא מקסימלי.
3. שיטות משובצות – בגישה זו בחירת המאפיינים מתבצעת כ*חלק מתהליך הלמידה*, כך יש יתרון בשל אינטרקציה עם תהליך הלמידה כמו בשיטת המעטפת אך בסיבוכיות טובה יותר בשל השילוב תוך כדי בניית רכיב הסיווג.

Filter

Wrapper

Embedded method

Classifier

Classifier

Classifier

איור 7. שיטות בחירת המאפיינים

גם את שיטות בחירת המאפיינים ניתן לחלק לשתי תתי קטגוריות בהתאם ללמידה :

* 1. השיטות העיקריות עבור למידה בלתי מפוקחת –  
     שיטות פילטור  
     בהיעדר תיוג לקבוצת האימון (אשר ינתב את בחירת המאפיינים), יישום שיטה זו, קרי בחירת המאפיינים ללא שימוש באלגוריתם הלמידה, הופכת למשימה מורכבת למדי. במצב זה יש למצוא דרך לדרג את המאפיינים בצורה כזו שתשמר *לוקליות,* כלומר דוגמאות (תמונות) שהיו דומות במרחב הקלט יהיו דומות במרחב המצומצם, אחת הדרכים לביצוע דירוג זה היא בעזרת ציון לפלסיאן (Laplacian Score). ציון לפלסיאן מדרג את המאפיינים כך שאובייקטים שהיו קרובים (כלומר דומים) במרחב הקלט יהיו קרובים במרחב הפלט. מלבד תכונת הלוקליות, נובע כי אין מאפיינים לא-רלוונטים אולם ייתכן כי חלקם יתירים. חישוב ציון זה מבוסס על בניית גרף , המורכב מיחסי השכנויות בין דוגמאות הקלט, המיוצג ע"י מטריצה ריבועית :  
     עבור כל כאשר ו- הם שכנים

אחרת . ניתן לקבוע פרמטר נוסף בבנית שהוא מספר השכנים שיש לקחת בחשבון. הוא תוצאת התמרת לפלס[[4]](#footnote-4) על גרף כאשר היא המטריצה הדרגות האלכסונית .  
אם הוא ווקטור הערכים בקבוצת הדוגמאות המתאים למאפיין ה- אזי ציון לפלסיאן מוגדר ע"י :

כאשר , 1 הוא וקטור שכל ערכיו 1 באורך

ציון זה ישמש לדרג את מידת כל המאפיינים בשימור לוקליות.

שיטות מעטפת  
כאמור, בשיטות מעטפת משתמשים בתהליך הקלסיפיקציה או קיבוץ על מנת לאמוד את תת קבוצת המאפיינים נציג לכך דוגמא :  
תועלת הקטגוריה (Category Utility) –

כאשר הוא סט המקבצים (Clusters)  
ו- הוא סט המאפיינים. שיטת תועלת הקטגוריה מודדת את השונות בין ההסתברות המותנת של המאפיין כאשר ערכו במקבץ אל מול הההסתברות המקורית של המאפיין.   
 המטרה היא מדידת העליה במספר ערכי המאפיינים הניתן לחזות נכונה בהינתן קבוצת דוגמאות לעומת אלו הניתנים לחיזוי ללא קבוצת דוגמאות.  
תחילה בוחרים תת קבוצת מאפיינים, ניתן להתחיל עם קבוצה ריקה ולהוסיף מאפיינים או עם הקבוצה המלאה ולהסיר מאפיינים. בהמשך מריצים את תהליך הלמידה הבוחן כמה מערכי מאפיינים ניתנים לחיזוי נכון ולאחר מכן מודדים את המקבצים שנוצרו בעזרת ציון ה- Category Utility, בחירת הקבוצת תמשיך עד אשר ציון החיזוי חדל להשתפר, אז תת הקבוצה שקיבלה את הציון הכי גבוה תהיה הבסיס לשלב הבא (הוספת מאפיין \ הסרת מאפיין). כך עוברים משלב לשלב עד אשר ציון ה- Category Utility חדל להשתפר.

שיטות משובצות  
בשיטת בחירת המאפיינים המשובצת תהליך בחירת המאפיינים הוא חלק מאלגוריתם הסיווג. ישנם מספר טכניקות קיבוץ אשר תוך כדי תהליך הקיבוץ מביאות להקטנת מרחב המאפיינים, שיטות אלה מגלות מקבצים במידע המוגדרים ע"י תת קבוצה של המאפיינים המקוריים כאשר כל מקבץ יכול להיות מתואר ע"י תת קבוצה אחרת כלומר אלה הם מאפיינים *מקומיים*. החלופה לכך היא מציאת תת קבוצה *כוללת* שבאמצעותה ניתן לתאר את כל המקבצים. נציג דוגמא לשיטה הכוללת, אלגוריתם :  
דרך לאמוד כהלכה את טיב המקבץ היא לפי לכידותו, במונחים של תורת הגרפים הדבר מתבטא בכך שאובייקטים בתוך המקבץ הם קשירים היטב ואילו בין המקבצים קשירות רופפת. יש למזער את חתך הגרף בין המקבצים, על עיקרון זה מושתת אלגוריתם אשר מחפש חלוקה טובה לדוגמאות הקלט ומבצע בחירת מאפיינים בו זמנית :   
כאשר היא מטריצת קרבה וכל כניסה בה מייצגת את הקרבה (או הדמיון בין דוגמאות, במקרה זה מכפלה סקלרית) בין צומת לצומת .  
 מייצת את השורה ה- במספר במטריצה המנורמלת כך שמרכזה ב-0   
 הוא וקטור המשקלים עבור המאפיינים והמטרה שרוב המשקלים הללו יתאפסו. עבור מטריצה הבנוייה מ הווקטורים העצמיים של המתאימים ל הערכים העצמיים הגדולים ביותר.   
הציון למדידת קבוצת מאפיינים בהינתן ווקטור משקל מוגדר ע"י :   
אם כן המטרה של אלגוריתם היא "חידוד" המקבצים, בכך שנחפש את תת קבוצת המאפיינים שבאמצעותה נבנה מקבצים מופרדים היטב בהתאם לעיקרון החתך.

* 1. השיטות העיקריות עבור למידה מפוקחת –  
     שיטות פילטור   
     כאמור, תחילה מבצעים ניתוח בו מדרגים את מידת החיזוי של המאפיינים השונים, השיטות המקובלות לדירוג זה הן מדד כי בריבוע, מדד תוספת המידע (Information gain) ומדד סבירות היחס (Odds ratio). המטרה היא לבחון עד כמה ההתפלגות של ערכי המאפיין מאפשרת חיזוי שייכות למחלקה. באופן אינטואיטיבי ככל שהערכים קרובים יותר להתפלגות אחידה מידת החיזוי נמוכה יותר עבור המאפיין הנבחן למחלקה מסויימת. חשוב לציין כי עבור שיטות אלו נדרש כי המאפיינים יהיו בדידים (ניתן לבצע התאמה לכך). ניתן לראות את ההתפלגויות השונות בטבלה הבאה :

|  |  |
| --- | --- |
| מדד |  |
| כי בריבוע | מטרתה לבדוק את מידת הקירבה של מאפיין למחלקה בהתבסס על מידע  הוא מספר הדוגמאות החיוביות עם ערך מאפיין ו- הוא הערך הצפוי אם לא הייתה התאמה בין ל- |
| תוספת מידע | מטרתה למדוד את כמות המידע שהמאפיין תורם לקבוצת האימון, מוגדרת להיות ההפחתה הצפויה באנטרופיה[[5]](#footnote-5) בעקבות החלוקה של קבוצת האימון בעזרת המאפיין   *הוא תת הקבוצה של קבוצת האימון המתאימה לערך במאפיין* |
| סבירות היחס | בשימוש בעיקר כאשר המאפיינים הם בינאריים  *מתאים לערך החיובי, מתאים לערך השלילי* |

טבלה 2. מדדים לדירוג חיזוי המאפיינים

מלבד הניתוח יש לקבוע מדניות בעזרתה הפילטור יתבצע, לדוגמא :

1. בחירת המאפיינים הראשונים, מתוך , בהתבסס על ציונם (מהשלב הקודם), כלומר – בחר את 50% הראשונים.
2. בחר את המאפיינים העוברים סף ציון, כלומר בחר את אלו שציונם הוא בקרבת 50% מהציון המקסימלי.
3. התחל עם המאפיין שציונו הכי גבוה, אמוד את רכיב הסיווג שנבנה. לאחר מכן הוסף את המאפיין הבא שציונו הכי גבוה ואמוד בשנית, יש להמשיך בתהליך עד אשר אין שיפור ברכיב הסיווג.

שיטות מעטפת  
הביקורת המתבקשת לשיטת הפילטור נובעת מההפרדה בין בחירת המאפיינים לבין רכיב הסיווג, על כך מתגברים בשיטת המעטפת ע"י שימוש בהערכת הדיוק של רכיב הסיווג על מנת להנחות את חיפוש המאפיינים – כלומר רכיב הסיווג "עטוף" בתהליך בחירת המאפיינים. אולם הבדיקה עבור כל רכיב סיווג פוטנציאלי שיבנה הופכת שיטה זו ליקרה מאוד מבחינה חישובית – אם ישנם מאפיינים בבחינה, מרחב החיפוש יהיה , כלומר מרחב אקספוננציאלי. לכן כאשר נידונים ערכי גדולים חיפוש מקיף אינו פרקטי ובמצב כזה מפעילים אסטרטגיות חיפוש יעילות יותר כגון :

1. בחירה קדימה (Forward selection) – מתחילים עם סט מאפיינים ריק, בוחנים את כל האפשרויות של מאפיין אחד, בוחרים את הטוב ביותר כעת בוחנים את כל האפשרויות של המאפיין הנבחר ועוד מאפיין נוסף וחוזר חלילה עד אשר אין שיפור הדיוק רכיב הסיווג.
2. הסרה לאחור ( (Backward elimination– מתחילים עם סט המאפיינים המלא, ומכל האפשרויות להשמטת מאפיין אחד, בוחרים את הטובה ביותר וממשיכים בהשמטת מאפיינים זו עד אשר אין שיפור דיוק ברכיב הסיווג.
3. חיפוש גנטי (Genetic search) – כל מועמד מוגדר ע"י מסכת מאפיין, מתבצעת הצלבה של מועמדים ותהליך מוטציה ליצירת מועמדים חדשים.  
   אולם פונקציית הבחירה (כשירות) של המועמד ובמקרה זה מידת הדיוק של רכיב הסיווג היא יקרה.
4. הדמיית חישול (Simulated annealing) – משלב את היתרונות של השיטה הגנטית, בכך שהיא מאפשרת אקראיות אך יעילותה טובה יותר.  
   מבצעים חיפוש מקסימום מקומי ומאפשרים בעזרת חיפוש אקראי לשפר את ערך זה גם אם משמעות הדבר הוא קבלת ערך נמוך יותר, בהדרגה מקטינים את תדירותם וגודלם של מהלכי השיפור וכך שלבסוף השיטה מתייצבת על ערך.
5. שיטות שיפור CBIR בעזרת למידה חישובית
   1. **שיטות למידה מפוקחת**

למידה מפוקחת (Supervised learning) היא אחת השיטות הבולטות בתחום הלמידה החישובית בכלל, ובתחום ה CBIR בפרט. למידה מפוקחת מבוססת על מיפוי, או למעשה פונקציה, בין סט דוגמאות לערך תוצאה והחלת מיפוי זה על קלט חדש.

* דוגמאות לאלגוריתמים נפוצים בשימוש
  + - * + מכונת וקטורים תומכים (SVM)  
          בהינתן סט דוגמאות מסווגות, נבנה מודל סטטיסטי שבעזרתו ניתן לסווג קלטים חדשים. מודל SVM למעשה הוא ייצוג דוגמאות האימון בתור נקודות במרחב, באופן כזה שקטגוריות הסיווג מופרדות וכמה שיותר רחוקות זו מזו.  
            
          נתבונן בדוגמה לינארית בה הסיווג הוא בינארי (אמת, 1 או שקר, 1-) :  
          יהי וקטור המאפיינים המייצג את מידע האימון ו המייצג את הסיווג.  
          מגדירים מישור-על (שהיינו תת מרחב שמימדו קטן ב-1 ממימדו של המרחב המקורי) ע"י כאשר ו-   
          מישור זה מחלק את המרחב לשני חצאי מרחבים כאשר כל אחד מהם מכיל קבוצת סיווג. כיוון שכל קבוצת סיווג היא נפרדת לינארית ניתן להגדיר את הפונקציה: כאשר   
          פונקציית ההחלטה יכולה להיות מתוארת ע"י   
          כאשר   
          כיוון שפונקציות רבות יכולות לענות הנדרש (הפרדה לינארית על דוגמאות האימון), מחילים אילוצים נוספים על הפונקציה על מנת להגיע למישור האופטימלי כזה שעבורו המרווח מקסימלי בין שתי הקבוצות. המרווח מוגדר להיות מרחק הנקודה הכי קרובה למישור מכל צד (מכל קבוצה).  
          מרחק נקודה מהמישור :   
          את האופטימיזציה ניתן לייצר ע"י :

וקטור תומך

X2

X1

איור 8. מציאת המפריד עבורו המרחק מקסימלי

הווקטורים התומכים הם נקודות האימון עבורן מתקיים שיווין עבור האופטימיזציה הנ"ל. בצורה כזו הם קרובים באותה מידה למישור האופטימלי.  
 במצבים בהם דוגמאות האימון אינן ניתנות להפרדה לינארית טובה, ניתן להרחיב לגירסה אחרת של SVM הנקראת Soft Margin

בגירסה זו ניתן לפשט את הבעיה ע"י הוספת גורם

במצב כזה ניתן לבטא את האופטימיזציה ע"י

ובעזרת הקבוע המשמש לכיוונון גודל המרווח לעומת טעויות סיווג.

* + - * + סיווג שכנים קרובים ביותר (NN)  
          בסיווג NN דוגמאות חדשות מסווגות לפי הסיווג של השכנים הקרובים ביותר לאותה דוגמאה (או K-NN כאשר הן מסווגות לפי שכנים שיקבעו את הסיווג).  
          בסיווג זה, תהליך ההחלטה מתבצע בזמן הריצה (למידה עצלה - Lazy Learning) ולכן דוגמאות האימון נדרשות להשמר בזיכרון.   
            
          לסיווג K-NN שני שלבים למעשה, בשלב הראשון מחפשים את השכנים הקרובים ביותר ובשלב השני קובעים את הסיווג עפ"י השכנים מהשלב הראשון.  
          בהינתן סט דוגמאות אימון המורכב מ דוגמאות המתוארות ע"י סט מאפיינים כלשהו , כאשר ערכים מספריים ינורמלו לערכים בין . כל דוגמאת אימון מסווגת והמטרה היא לסווג דוגמא חדשה .  
          לכל ניתן לחדש את המרחק ל ע"י :

איור 9. סיווג לפיk השכנים הקרובים ביותר K-NN   
אם k=3 הסיווג יהיה חיובי  
אם k=5 הסיווג יהיה שלילי

טווח הערכים האפשרי למרחק (עבור ערכים בדידים ורציפים)

*כלומר, במידה והפונקציה המייצגת את המאפיין היא בדידה, על ערכים זהים המרחק הוא 0, על ערכים שונים המרחק הוא 1 ואילו אם הפונקציה היא רציפה המרחק יוגדר להפרש בין ערכי הדוגמאות, בין זו המסווגת מסט הדוגמאות ובין זו החדשה הנדרשת לסיווג.   
 השכנים הקרובים ביותר נבחרים עפ"י אומדן המרחק הנ"ל ולאחר מכן ישנן יש לבחור את הסיווג המתאים של בהתבסס על שכנים אלו. לכך מספר שיטות, כאשר אחת השיטות המיידיות היא בחירת הסיווג הנפוץ ביותר מבין השכנים. שיטה נוספת, נותנת משקל לכל שכן לפי המרחק שלו מ ,* כך שכן קרוב יותר יקבל משקל גדול יותר. ניתן לחשב משקלים אלו ע"י :

השכנים "מצביעים" על סיווג כאשר לקולם משקל ההפוך ממרחקם מ- , כך להצבעה על סיווג ע"י השכן היא יחסית למרחק וככל שהשכן קרוב יותר כך גם משקל קולו גדל.  
פשטותו היחסית של רכיב הסיווג הופכת אותו לאפשרות מוצלחת עבור תוכן מורכב כמו זו העולה בבעיית ה-CBIR.

* + - * + שיטת הרכבה - Ensemble  
          הרעיון המרכזי מאחורי שיטה זו היא שאסופה של רכיבי סיווג יכולה להניב תוצאות טובות יותר מרכיב סיווג בודד, בהיבטים של דיוק ויציבות. יתרונה של גישה זו בולט במיוחד כאשר יש שימוש ברכיב סיווג פחות יציבים כגון רשתות נוירונים ועצי החלטה, בהם אף שינוי קל בסט האימון יכול לגרור שינויים נרחבים במודל שנבנה ומכך גם הסיווג שיתקבל. בנוסף לכך, שיטת ההרכבה יכולה לשפר את הדיוק עבור רכיבי סיווג יציבים יותר כמו K-NN או Naïve Bayes. השיפור בדיוק וביציבות מושפע ישירות מהגיוון של הרכיבים המרכיבים. בהנחה כי הרכיבים מסווגים נכונה בממוצע וטועים כל אחד בנפרד כלומר על הטעויות אינן כה משותפות, אזי הסיווג הקולקטיבי ככל הנראה צודק יותר מאשר הסיווג היחידני.  
          ניתן לחלק את סוגי הטעויות של רכיבי סיווג לשלוש קטגוריות עיקריות :

רעש – מייצג את הפגם במאפיינים בשימוש, למעשה הוא החסם התחתון על מידת השגיאה שניתן להשיג ע"י רכיב הסיווג, אין דרך לשפר זאת מבחינת רכיבי הסיווג.

הסט – מייצג עד כמה הסיווג הממוצע של אלגוריתם הלמידה (בהינתן כל קבוצות האימון) מתאים לקלט.

שונות – מייצג את השונות (התזוזה) בסיווג של אלגוריתם הלמידה על סטי אימון שונים.

על מנת לשפר את רכיב הסיווג ולהקטין את מידת השגיאה כדאי לדעת מה טיב השגיאה, לדוגמא ברשתות נוירונים מידת ההטיה היא נמוכה אולם שונות גבוהה, על כן כדאי להתאים עוד רכיב – מודל נסיגה שלו שנות נמוכה אך ההסט גבוה. בצורה זו ניתן להקטין את מידת השגיאה.

עוד דרך לייצר הרכבה נקראת עטיפה Bagging, הרעיון המרכזי בשיטה זו הוא בהינתן סט אימון וקלט לסיווג :

1. עבור הרכבה של רכיבי סיווג, יש לייצר תתי סטי אימון *מ ע"י*  bootstrap sampling *כלומר  
   לעיתים קרובות*
2. *לכל , נגדיר להיות תת סט הדוגמאות שלא נבחרו עבור (יכול לשמש קבוצת בקרה ולהגביל את ה* overfitting של *רכיב הסיווג שאומן עם )*
3. *אימון רכיבי סיווג , באמצעות*
4. *קבלת סיווגים עובר מכל אחד מרכיבי הסיווג*
5. *סיכום ערכים אלו ע"י פונקצית צבירה כלשהי (ממוצע וכד').*

*נוסחה כללית לחישוב הסיווג הכולל :*

*עבור ממוצע משוקלל :*

כאשר מתקיים , כלומר המשקלים משלימים ל-1. בהנחה שההרכבה הזו מגוונת, תכונה שאמורה להתקיים מהשימוש ב - bootstrap sampling, הסיווג יהיה מדוייק יותר מהסיווגים המרכיבים .

יש לציין כי בשיטות יציבות יותר כגון K-NN או Naïve Bayes תת קבוצות של סט האימון לא יניבו שונות (גיוון) בהרכבה. במצב כזה ניתן לקחת תת קבוצות של מאפיינים במקום, אופן זה משמעותי כאשר המידע מתואר ע"י כמות רבה של מאפיינים ושיש בהם יתירות.

* 1. **שיטות למידה בלתי מפוקחת**

למידה בתי מפוקחת (Un-supervised learning) הינה חשובה לא פחות מלמידה מפוקחת, שכן קיבוץ מידע או חלוקת המידע למחלקות בהיעדר דוגמאות תיוג הוא תרחיש נפוץ. יחד עם זאת היעדר דוגמאות תיוג הופכות את משימת אמידת טיב הקיבוץ \ החלוקה למשימה לא פשוטה כלל.

* דוגמאות לאלגוריתמים נפוצים בשימוש  
  *שיטות קלאסיות*  
  שיטות הקיבוץ הקלאסיות מתחילות עם פיתרון חלוקה ראשוני ומחפשות ליצור פונקציה המשפרת את החלוקה הראשונית הזו ולהגיע לחלוקה המיטבית.
  + קיבוץ מבוסס מרכז כובד (k-Means Clustering)  
    שיטה זו היא אחת השיטות הנפוצות ביותר עבור קיבוץ, העושה שימוש בסכמת מיקום מחדש של פריטי המידע על מנת ליצור מחלקות המקטינות את מידת העיוות (distortion)[[6]](#footnote-6) בין הפריטים המסומנים וסט של נציגי מחלקות. כל נציג שכזה, הנקרא בשם מרכז כובד (centeroid) מחושב ע"י הווקטור הממוצע של קבוצת פריטי המידע במחלקה מסויימת. מידת העיוות נמדדת כמרחק אוקלידי, כך שמטרת הקיבוץ הופכת להיות מיזעור שגיאת סכום הריבועים (sum-of-squared error או SSE) בין פריטי המידע ונקודת מרכז הכובד של הקבוצה :  
    התהליך הבסיסי לרוב מחולק לשני חלקים, בחלק הראשון כל פריט מידע ממוקם מחדש ביחס לנקודת מרכז הכובד הכי קרובה. לאחר שכל פריטי המידע עובדו מעדכנים את ווקטורי מרכז הכובד על מנת לשקף את החלוקה החדשה. תהליך זה חוזר על עצמו עד אשר מספקים קריטריון עצירה הנקבע מראש, בדרך כלל כאשר המחלקות מפסיקות להשתנות בין שלב לשלב (חלוקה יציבה) או עד הגעה לסף שינוי מסויים. שיטה זו עלולה לפספס מחלקה מועילה במקרה שבו פרטי המידע אינם מסודרים בצורת פיזור מעגלית או שהמעגלים שונים מאוד בגודלם.
  + קיבוץ מעורפל (Fuzzy Clustering)  
    שיטת FCM (Fuzzy c-means) מכלילה את השיטת קיבוץ מבוסס מרכז כובד הסטנדרטית בכך שהיא שמאפשרת לפריטי מידע להיות משויכים למספר קבוצות שונות עד סף מסויים הניתן ע"י משקולות הסתברותיים. ניתן לייצג משקולות אלו במטריצה בגודל שתאיה משקפים את מידת השיוך של פריט מידע לקבוצה כאשר .  
    מטרת הקיבוץ כאמור היא למזער את מידת העיוות בין פריטי המידע ומרכזי הכובד ובשיטה זו מידת העיוות נמדדת ע"י פונקציית הקרטריון המעורפל :  
    עבורה המאריך שולט במידת העירפול של שייכות פריטי המידע. בנוסף לכך, בשיטה זו מרכזי הכובד מחושבים ע"י :  
    שיטה נוספת לביצוע קיבוץ מעורפל היא מיקסום התוחלת (expectation maximization - EM) ובשונה משאר השיטות שתוארו שיטה זו מפעילה גישה מבוססת מודל למציאת החלוקה לקבוצות במידע. כלומר שיטה זו מניחה כי פריטי המידע נוצרו בעזרת מודל כלשהו המהווה תערובת של התפלגויות . במקרה זה פעולת הקיבוץ הופכת לזיהוי הגורמים שקרוב לוודאי שימשו את המודל, כאשר כל גורם מייצג קבוצה. השיטה מתחילה עם הערכה התחלתית לגורמי המודל ומפעילה אופטימיזציה חוזרת ונישנת של הערכת הגורמים הנוכחיים ובחירת גורמים מחודשת שתביא למיקסום ההסתברות. בדומה לשיטת קיבוץ מבוסס מרכז כובד, לבחירת הקבוצות הראשונית השפעה רבה על דיוק התוצאה הסופית.
  + קיבוץ היררכי (Hierarchical Clustering)  
    שיטות מסוג זה זו מייצרות קבוצות מקוננות, במבנה של עץ המייצגות קיבוץ היררכי הלא אחת מתאר את המידע בצורה יותר מציאותית מאשר קבוצות משוטחות. שיטות אלו מתחלקות לשתי קטגוריות עיקריות :  
    שיטות צבירה – בהן מתחילים כאשר כל פריט מידע שייך למחלקה נפרדת ולאחר מכן ממזגים את המחלקות הדומות ביותר בצורה מחזורית.   
    שיטות חלוקה – בהן מתחילים עם קבוצה אחת המכילה את כל פריטי המידע ולאחר מכן מחלקים קבוצה נבחרת כלשהי לתתי קבוצות.

*שיטות עדכניות*

* + קיבוץ מבוסס גרעין (Kernel Clustering)  
    שיטות מסוג זה כרוכות בהתמרת המידע למרחב חדש (ייתכן מסדר גבוה יותר) בו הקשרים הלא לינאריים בין פריטי המידע יהיו יותר קלים לזיהוי. במקום לחשב את ייצוג החדש (לאחר ההתמרה) לכל פריט מידע מבצעים שימוש בתכסיס הגרעין (kernel trick) המאפשר בחינת מידת הקירבה בין זוג פריטי מידע ע"י ערך פונקציית הגרעין :  
    בפועל הפונקציה מיוצגת ע"י מטריצה סימטרית המכונה מטריצת גרם . העבודה עם אומדן קירבה מקל את תהליך הלמידה במרחב החדש. ניתן להטמיע את השימוש בפונקציית גרעין בשיטות שתוארו לעיל.
  + קיבוץ מבוסס ערכים עצמיים (Spectral Clustering)  
    שיטות אלו מבוססות על תורם הגרפים, וכוללות בניית מרחב מסדר נמוך יותר בעזרת הפירוק לפי הערכים העצמיים של המטריצה (Eigen Value Decomposition) המייצגת את גרף הקשרים במידע. ניתן לייצג את הקשרים במידע ע"י מטריצת קירבה שבה כל תא מייצג את הקשר בין ו- . במקרה זה בעית הקיבוץ הופכת לבעית חלוקת גרף כאשר היא מטריצת שכנויות עבור גרף ממושקל לא מכוון וכאשר קבוצת הקודקודים היא קבוצת פריטי המידע ו- קבוצת הקשתות היא דמיון בין שני פריטי מידע. בעזרת ניסוח זה הבעיה הופכת למציאת חלוקה לקבוצות של גרף המביאה קריטריון מחיר כלשהו לערך האופטימלי. לדוגמא קריטריון החתך המינימלי (minimum cut) המודד את משקל הקשתות החוצות את הקבוצה.
  + רשת קוהונן (Self-organizing Maps SOM -)  
    שיטה נפוצה זו, המבוססת רשת נוירונים, מספקת מיפוי משמר טופולוגיה ממרחב שמימדו מסדר גבוה למרחב בעל סדר נמוך. במימושים רבים המרחב הנוצר הינו מרחב פשוט, מרחב דו מימדי, למרות שגם מרחבים מסדר גבוה יותר אפשריים. במיפוי זה שיטת SOM גם מבצעת הקטנה של מרחב המאפיינים וגם חיפוש נציגי קבוצות אפשריים.  
    השיטה מתייחסת אל פריטי המידע בתור רשת תאים כאשר כל תא מייצג תכונה כלשהי. ל-SOM שלב אימון בו המפה נבנת ומותאמת כך שתתאר את תכונות המידע בצורה הטובה ביותר. לאימון מספר תתי שלבים :  
    איתחול – בניית המפה הראשונית, איתחול פשוט כולל השמת כל תא עם ווקטור מודל אקראי. צורה מורכבת יותר היא שימוש ב- PCA כאשר הערכים בתאים יהיו הווקטורים העצמיים של הערכים העצמיים של המידע בקבוצת האימון.  
    לימוד מחזורי - עבור דוגמאת למידה אקראית מוצאים את הווקטור (תא) הכי קרוב אליה במפה (קירבה זו נמדדת לרוב עפ"י מרחק אוקלידי) ולאחר מכן מעדכנים את המפה בחלקים הרלוונטים (התא והשכנים הקרובים אליו בהתאם לפונקציית בחירת שכנים הנקבעת מראש) ומקרבים אותה לדוגמאת הלמידה.  
    טיוב – בסיום שלב הלמידה עוברים שוב על דוגמאות הלמידה ומעדכנים על פיהם רק את התאים המתאימים ביותר לכל דוגמא (ללא השכנים).

1. בחינה לעומק :Supervised SVM

בחלק זה נציג שתי שיטות בהן ניתן לשפר את תהליך האיחזור בעזרת מכונת ווקטורים תומכים SVM. שיטות אלו מתבססות על תהליך אימון אינטרטקטיבי ע"י משוב רלוונטיות מהמשתמש לאחר ביצוע שליפה. בכל אחת מהשיטות המוצגות מאומן רכיב SVM בהתאם למשוב הרלוונטיות ולפיו מתבצע טיוב של תוצאות האיחזור.

* 1. **משוב רלוונטיות (RF) לאיחזור תמונה (CBIR) בעזרת מכונת ווקטורים תומכים (SVM) ובחירת מאפיינים** 
     1. תיאור כלליאחת הדרכים להתמודד עם הפער הסמנטי היא באמצעות משוב רלוונטיות RF (רלוונטי או לא רלוונטי) הניתן ע"י המשתמש, המשוב מאפשר למשתמש "ללמד" את המערכת ולטייב את תוצאותיה בהתאם לשליפת המשתמש, המשוב ישפיע על דירוג מאגר התמונות לאותה שליפה ויאפשר החזרת תמונות יותר רלוונטיות למשתמש. אולם משוב זה מוגבל כיוון שהמשתמש יכול לספק משוב על סט מאוד מצומצם של תמונות. יעילותו של רכיב סיווג, המאומן עם סט דוגמאות מזערי יחסית וכאשר סט המאפיינים (הנדרש לתיאור תמונה) הוא גדול, היא נמוכה. לשם כך תוצג שיטה, המנסה להתמודד עם בעיה זו, אשר תבחר סט מאפיינים קטן ומייטבי עבור הסיווג בהתבסס על משוב הרלוונטיות. אימון רכיב הסיווג יבוצע ע"י שימוש ב-SVM.
     2. פירוט האלגוריתם  
        *סכמה כללית*  
        1. קבלת שאילתה מהמשתמש וביצועה על מאגר התמונות.  
        2. קבלת משוב רלוונטיות מהמשתמש.   
        3. אימון רכיב סיווג SVM והקטנת מרחב המאפיינים ע"י בחירת מאפיינים.   
        4. דירוג מאגר התמונותבהתאם לרכיב הסיווג שאומן בשלב 3.  
        5. החזרת תוצאות מחדש לפי הדירוג בשלב 4.  
        6. חזרה לשלב 2 במידה ויש עוד סבבי משוב רלוונטיות.  
        *מאפיינים מחולצים*
  + מאפייני צבע –  
    היסטוגרמת HSV בת 256 מימדים : עבור בהתאמה.
  + מומנט צבע בן 9 מימדים במרחב צבע (הרחבה של CIE L – מידת בהירות הצבע) המכיל את המומנטים צבע חציוני, התפלגות סטנדרטית ומידת הטיה) לכל אחת משלוש הקורדינאטות .
  + מאפייני טקסטורה –   
    ווקטור בן 104 מימדים הנבנה ע"י שלושה רמות של התמרת wavelet
  + מאפיינים צורה –   
    ווקטור בן 80 מימדים של היסטוגרמת קצוות הנבנה ע"י חלוקה שווה של התמונה ל-16 תתי תמונות וחישוב עבור כל אחת מתתי התמונות את תדירות הופעתם של 5 סוגי קצוות (אופקי, אנכי, אלכסוני וכו').

כל המאפיינים לעיל מאוחדים לווקטור אחד בן 440 מימדים והוא מהווה את ייצוגה של תמונה במאגר.

*SVM*פונקציות הגרעין המקובלות ביותר (עבור שימוש רכיב ה-SVM), שייכות למשפחת הפונקציות הנקראת Redial Basis Functions (RBF) ובהן Gaussian RBF היא אחת הפונקציות הנפוצות ביותר:

לאחר אימון רכיב הסיווג, פונקצית ההחלטה הנוצרת היא :

כאשר הוא פרמטר ההיסט.   
 הוא פרופורציונאלי למרחק של הקלט מגבול ההחלטה, לכן ניתן להתייחס אל הערך כאל מידת הוודאות לגבי הסיווג של ; ערכים חיוביים גדולים עבור סיווג חיובי 1+ וערכים שליליים קטנים עבור סיווג שלילי 1-. לעומתם ערכים בסביבת אפס אינם מספקים וודאות לגבי סיווגו של .

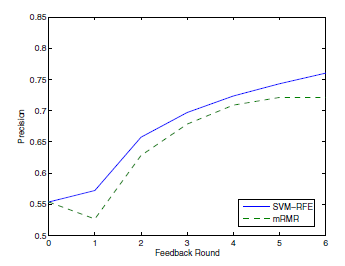
עבור כל איטרציית משוב פותרים את בעיית הסיווג הזו, כאשר אימון הסיווג מתבסס על דוגמאות המשוב, בתור ווקטורי מאפיינים המסווגים רלוונטי 1+ \ לא רלוונטי 1- . תמונת החיפוש, כלומר תמונת הקלט נחשבת לדוגמא רלוונטית ומסווגת 1+. רכיב הSVM יקבע לגבי כל תמונה במאגר החיפוש האם היא רלוונטית \ לא רלוונטית וערך הפונקציה יקבע את מידת הרלוונטיות לכל תמונה, ככל שהערך גדול יותר כך גם מידת הרלוונטיות בעיני המערכת.  
 *בחירת מאפיינים*בחירת המאפיינים תתבצע ע"י שיטה הנקראת סילוק מאפיינים מחזורי -SVM Recursive Feature Elimination (SVM-RFE). שיטה זו היא שיטה משובצת המסלקת מאפיינים באופן רקורסיבי בהתבסס על תוצאות הסיווג של דוגמאות אימון ע"י רכיב ה SVM. SVM-RFE משתמש בגרעין לינארי:   
נוסחת הדירוג תקבל את הצורה הבאה:

כאשר הווקטור הוא מאותו מימד של דוגמאת אימון . ניתן להסיק מנוסחת דירוג זו כי ככל שהערך של  *גדול יותר עבור האיבר ה- ב- כך גם השפעתו על ערך פונקציית ההחלטה לתמונה לא ידועה המיוצגת ע"י . מכך ניתן לגרוס קריטריון לדירוג המאפיינים בהתאם לתרומתם לתהליך הסיווג.*   
שיטת SVM-RFE היא רקורסיבית, ראשית מתחילים עם סט המאפיינים המלא ובכל חזרה מעדכנים את סט המאפיינים ע"י סילוק המאפיין שתורם הכי פחות לתהליך הסיווג. התרומה נמדדת ע"י בחינת רכיב ה SVM שנבנה מדוגמאות האימון ומוגבל למאפיינים אשר ב באותה איטרציה. לאחר האימון המאפיין בעל הערך הקטן ביותר נחשב לזה שתרומתו זניחה והוא מסולק מ . התהליך חוזר עד אשר מגיעה לגודל הרצוי שנקבע מראש. מוחזרת בתור קבוצת המאפיינים הדרושה.  
 שיטה נוספת לבחירת מאפיינים היא מינימום יתירות מקסימום רלוונטיות Minimum Redundancy Maximum Relevance (mRMR), שיטה זו היא שיטת פילטור המבוססת על אינפורמציה הדדית המהווה מדד למידת הקשר בין שני משתנים אקראיים ומוגדרת ע"י:

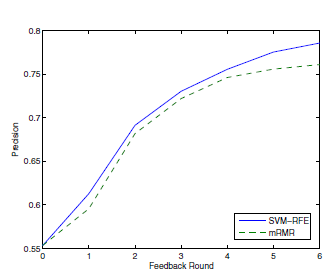
כאשר הן פונקציות הצפיפות של ו- היא פונקציית הצפיפות המשותפת. בשיטת mRMR מתייחסים אל המאפיינים ואל הסיווג בתור משתנים אקראיים, תחת הנחה זו ייבחרו המאפיינים שהרלוונטיות שלהם לסיווג גבוהה ובו בזמן אין בינהם יתירות או חזרה על מידע.

כאשר מסמנים משתנים אקראיים המתאימים למאפיינים ו- לסיווג. מסמן את סט המאפיינים שנבחר עד כה ו- מסמן את סט המאפיינים המלא. *תחילה ריקה, באיטרציה הראשונה המאפיין בעל הערך המקסימלי (כלומר הרלוונטי ביותר לסיווג) נבחר ומוכנס ל . באיטרציות הבאות, מאפיין מוכנס ל , העומד בתנאים של רלוונטיות גבוהה ביחס לסיווג ורלוונטיות נמוכה עם המאפיינים שנמצאים ב עד כה , על מנת למקסם את . נחזור על שלבים אלו עד אשר נגיע למספר המאפיינים הרצוי ב-.*

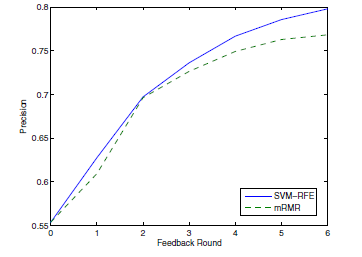
* + 1. ניתוח תוצאות  
       האלגוריתם המוצע לעיל נבחן על סט תמונות המכיל 3740 מתוך המאגר  
       Microsoft Research Cambridge : Object Recognition Image Database, סט זה סווג ידנית ל-17 קטגוריות ועל בסיס סיווג זה ייבחן המשוב.

המדד לבחינת יעילות האלגוריתם יהיה מדד הדיוק (Precision), המחושב ע"י היחס בין כמות התמונות הרלוונטיות מתוך התמונות הראשונות שהוחזרו. תמונה תחשב רלוונטית אם היא נמצאת באותה קטגוריה עם תמונת החיפוש ותחשב לא רלוונטית אם הן נמצאות בקטגוריות שונות.בניסוי זה 1000 תמונות מתוך המאגר משמשות פעם אחת בתור תמונות לחיפוש. לכל תמונה לחיפוש מבצעים סימולציה של שישה סביבי משוב רלוונטיות. בכל סבב משוב לכל היותר שלוש דוגמאות רלוונטיות ושלוש דוגמאות לא רלוונטיות נבחרות בצורה אקראית מתוך חמישים התמונות הראשונות בדירוג. תמונות אלו משמשות ביחד עם התמונות מהסבב הקודם לבחירת – מספר המאפיינים הנדרשים שישמשו את רכיב ה-SVM וכתוצאה מכך הקטנה של מרחב המאפיינים. באמצעות רכיב הסיווג החדש, מאגר התמונות מדורג מחדש. עבור הדירוג הראשוני כאשר אין עוד משוב משתמש, לכן לא ניתן לבחור מאפיינים או לבנות רכיב סיווג, ייתבצע שימוש במרחק אוקלידי במרחב המאפיינים המקורי.נציג כעת השוואה בין שיטות בחירת המאפיינים שתוארו בחלק הקודם :   
מספר המאפיינים שיבחרו מתקבל בתור פרמטר לשיטות SVM-RFE ו- mRMR. איורים 9, 10, 11 מציגים את הדיוק הממוצע עבור בחינת 20 התמונות ראשונות בסט התוצאות מול מספר מאפיינים שונה  
 . איור 12 מציג דיוק ממוצע לשיטת SVM-RFE לעומת ללא בחירת מאפיינים עבור סטי מאפיינים בגדלים שונים  
.  
כפי שניתן לראות (מאיורים 9, 10, 11) מדד הדיוק בשיטה המוצעת SVM-RFE טובים יותר וראוי לציין כי גם ביצועיה עדיפים (שיטה זו מהירה פי 6 משיטת mRMR – עבור הזמן הממוצע היה 2.5 שניות לחיפוש עבור 6 סבבי משוב רלוונטיות ואילו בשיטה השניה נדרש 15 שניות). מאיור 12 ניתן לראות כי בעזרת מספר מאפיינים קטן יחסית ניתן להשיג דיוק שווה או דיוק העולה על השיטה ללא בחירת מאפיינים גם עבור סט המאפיינים המלא.  
 ****

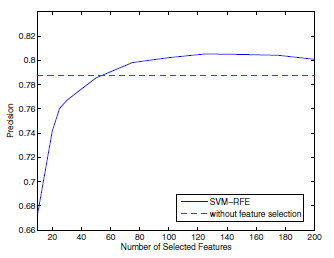
איור 10. דיוק ממוצע עבור (מספר התמונות לבחינה)   
ו (מספר המאפיינים)

****

איור 11. דיוק ממוצע עבור (מספר התמונות לבחינה)   
ו (מספר המאפיינים)



איור 12. דיוק ממוצע עבור (מספר התמונות לבחינה)   
ו (מספר המאפיינים)



איור 13. דיוק ממוצע עבור עבור האיטרציה מספר 6 של משוב הרלוונטיות,   
עבור כמות מאפיינים משתנה

* + 1. מסקנות

הוצגה שיטת שיפור איחזור תמונה בעזרת משוב רלוונטיות ע"י אימון רכיב סיווג SVM. במסגרת שיטה זו שולב מהלך מקדים של הקטנה של מרחב המאפיינים המתאימה ל SVM. ההקטנה בוצעה ע"י בחירת מאפיינים בשני אופנים RFE ו- mRMR, כמו כן בוצעה השוואה לקביעת בחירת המאפיינים המוצלחת יותר ומסתמן כי RFE מדוייקת יותר על מאגר תמונות הניסוי. בנוסף לכך בוצעה השוואה בין השיטה המוצעת לבין משוב רלוונטיות ללא בחירת מאפיינים ואימון, ניתן לראות מהתוצאות כי גם עבור הקטנה משמעותית של מרחב המאפיינים ניתן להגיע לדיוק טוב לפחות אם לא יותר טוב לדיוק המתקבל בשימוש בסט המאפיינים המלא.

* 1. **שיפור איחזור תמונה (CBIR) כאשר דוגמאות התיוג מרעישות**
     1. תיאור כלליבשיטות איחזור תמונה המשלבות משוב רלוונטיות עלולות להתעורר בעיות יישומיות שעשויות לפגוע בטיב האיחזור: ראשית, רובם של המשתמשים מעוניינים להשלים את תהליך החיפוש באופן מיידי ולא להאלץ לעבור מספר איטרציות, בייחוד בסביבת האינטרנט. שנית, בעוד תיוג דוגמאות חיוביות (רלוונטיות) הוא קל יחסית, תיוג מספר דוגמאות שליליות (לא רלוונטיות) רב ומגוון אורך זמן ואינו קל לביצוע. בנוסף לכך המשתמשים עלולים להכניס דוגמאות "מרעישות" שיפגעו בתהליך האיחזור. לשם התמודדות עם בעיה זו תוצג שיטה ל"ניקוי" המידע ורכיב סיווג העמיד יותר לרעש.
     2. פירוט האלגוריתם   
        *סכמה כללית*

1. קבלת שאילתה משמשתמש וביצועה על מאגר התמונות.
2. קבלת משוב רלוונטיות מהמשתמש.
3. זיהוי וסילוק רעש מתוך דוגמאות התיוג.
4. אימון רכיב סיווג SVM + דירוג רלוונטיות עמיד לרעש.
5. החזרת תוצאות מחדש לפי הדירוג בשלב 4.

*מאפיינים מחולצים*  
הדמיון בין תמונות נמדד ע"י שילוב המרחקים בין מאפיינים שונים כגון מאפיני צבע וטקסטורה, שילוב מאפיינים זה נקרא קיבוץ מאפיינים :  
  
*קיבוץ מאפיינים*לא מעט אופן שילוב המאפיינים המחולצים מתוך תמונה היא נקודה קריטית בייצירת מערכת איחזור אפקטיבית, דרך אחת לשלב את סט המאפייינים נקראת קיבוץ מאפיינים.  
קיבוץ מאפיינים היא שיטה לאמידת דמיון בין תמונות ע"י שילוב מרחקים בין מאפיינים. התמונות הרלוונטיות (ותמונת השליפה) מדורגותלפידמיוןקיבוצי של סט מאפיינים המחושב ע"י כאשר הוא המרחק במאפיין ה בין תמונת השליפה לתמונה במאגר ו- הוא המשקל לכל מאפיין.  
 *מרחב אי תאימות*תוצג שיטת קיבוץ מאפיינים חדשה עבור הסיווג :  
בהינתן מאגר תמונות המכיל תמונות , מכל תמונה מחלצים מאפיינים שישמשו עבור ייצוגה.   
המשתמש מספק מספר דוגמאות תיוג חיוביות המפרטים את השאילתה. תהי תמונת המקור, על בסיסה נבנה מרחב ע"י  
כאשר מייצגת את אי התאימות בין ו- ביחס למאפיין ה- ו- הוא ווקטור במרחב בן מימדים הנקרא גם מרחב אי תאימות.  
נסמן ב- את המרחק עבור המאפיין ה- ולכן   
לכן כל התמונות ב- הן ווקטורים ב .  
 תחילה בונים הרכבה של מספר רכיבי SVM על מנת לסנן את דוגמאות התיוג המרעישות במטרה להסיר את הדוגמאות המזיקות ולשמור על הדוגמאות המועילות. על מנת ליצור את מרחב אי התאימות שתואר יש לבחור תמונת מקור. במקום לבחור בצורה אקראית או לפי שיטות מיצוע כלשהן אשר רגישות לרעש תמונת המקור תיבחר ע"י :  
כאשר הוא המרחק בין הדוגמא החיובית ה- המסומנת לבין מועמד המקור עבור מאפיין . עבור דוגמאות שליליות יבחרו תמונות באופן אקראי מהמאגר, במאגר גדול דיו ההסתברות לכך שתיוג זה יהיה נכון תהיה גבוה יחסית.  
 *SVM*

איור 14. קיבוץ מאפיינים עבור CBIR

Feature 1

Feature n

Feature 1

Feature n

…

…

Query Image

Distance function

Distance function

…

Determine feature weights

Database Image

…

Determine feature weights

…

Final ranking

במרחב האי תאימות, בו דוגמאות שליליות וחיוביות, ניתן לאמן רכיב סיווג SVM. על מנת להתגבר על הדוגמאות החיוביות המרעישות, יתבצע שימוש במספר סטים שונים של דוגמאות שליליות על מנת האימון מכך יתקבלו מספר רכיבי סיווג. לאחר מכן מפעילים את רכיבי הסיווג על דוגמאות התיוג שניתנו ע"י המשתמש, אלו שסווגו שליליות (לפי עיקרון הרוב) ייחשבו לדוגמאות תיוג מרעישות ויסולקו.  
 עם זאת ייתכן ובצורה זו לא הוסרו את כל דוגמאות התיוג המרעישות, לצורך כך חישוב הרלוונטיות יבנה כך שיהיה עמיד יחסית לרעש ע"י הסתברות מידת הרלוונטיות :

* אומדן הסתברות הרלוונטיות –

על מנת לאמוד את ההסתברות נשתמש בפונקציית sigmoid[[7]](#footnote-7) יחד עם תוצאת רכיב סיווג SVM על דוגמאת תיוג חיובית   
כאשר הוא ערך פונקציית ההחלטה ו- הוא ערך הסיווג הצפוי, ושניהם מחושבים ע"י רכיב ה SVM. בסיווג בינארי מייצג סיווג חיובי ו- סיווג שלילי. מבצעים שימוש ברכיב הSVM שאומן לצורך סינון הדוגמאות המרעישות על מנת לסווג את הדוגמאות הנותרות. לאחר מכן התוצאות משמשות לחישוב ההסתברות המותנית לפי נוסחאת בייס :  
כאשר הוא האומדן להסתברות הרלוונטיות ו- הוא מספר רכיבי ה- SVM.

* חישוב רלוונטיות -על מנת להסיק את מידת הרלוונטיות של תמונה במאגר לתמונת החיפוש משתמשים במספר הרכבות של רכיבי SVM באופן הבא:

1. עבור מ- עד (כאשר היא קבוצת דוגמאות התיוג לאחר סינון)
   1. ניצור את מרחב אי התאימות עבור
      1. עבור מ-1 עד (כאשר הוא מספר רכיבי הסיווג)
         1. בחר סט דוגמאות אקראי מ שיסומן , כאשר
         2. (כאשר הוא רכיב סיווג הSVM הכללי)
2. צור את , ע"י שילוב כל רכיבי
3. החזר את

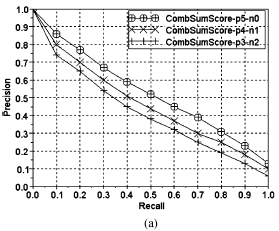
תחת סכמה זו ייבדקו שלושה מודלי קיבוץ מאפיינים :

* SVM-Weighted-MVR **-** הצבעת רוב משוקללת (Majority Vote Rule)

לכל תמונה מחשבים את הצבעת הרוב המשוקללת על מנת לקבוע האם התמונה רלוונטית או לא רלוונטית לחיפוש, חישוב זה מתבצע ע"י :  
כאשר הוא מאגר התמונות, הוא התיוג 0 או 1 המתקבל ע"י רכיב הסיווג ה עבור תמונת התיוג ה- (שנשמרה לאחר הסינון) ו- הוא המשקל שניתן לדוגמא זו, כלומר המייצג את מידת הרלוונטיות של דוגמאת התיוג החיובית ה . לאחר מכן, מודדים את מידת הרלוונטיות של תמונת המאגר מול תמונת החיפוש הניתנת כתוצאת הסיווג של רכיב SVMהנותן את אותו תיוג כמו זה של ה Weighted-MVR ועבורו הציון הכי גבוה (המשקל המוחלט של פונקציית ההחלטה של רכיב ה SVM)

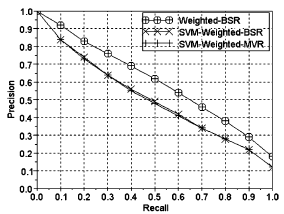
* SVM-Weighted-BSR - נוסחאת בייס משוקללת (Bayesian Sum Rule)

לכל תמונה משתמשים בערך ה BSR על מנת להסיק אם היא רלוונטית או לא רלוונטית : כאשר מייצג את ההסתברות המותנית לסיווג.לאחר מכן, מודדים את מידת הרלוונטיות של תמונת המאגר מול תמונת החיפוש הניתנת כתוצאת הסיווג של רכיב SVMהנותן את אותו תיוג כמו זה של ה Weighted-BSR ועבורו הציון הכי גבוה.

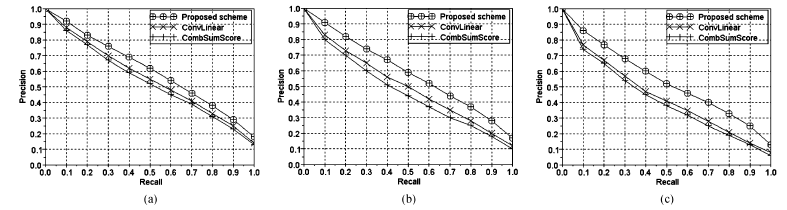
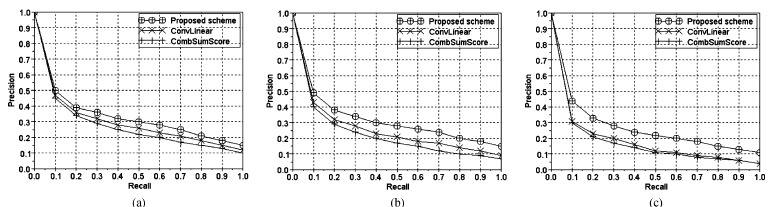
* Weighted-BSR  
  ניתן גם לבצע שימוש באופן ישיר בתוצאת ה BSR   
  בתור אומדן רלוונטיות של תמונת המאגר מול תמונת החיפוש.  
  + 1. ניתוח תוצאות  
       נציג השוואה בין שתי שיטות מימוש לסכמת האיחזור המוצעת, CombSumScore ו- ConvLinear. ב- CombSumScore המדד לדמיון בין תמונת המאגר לתמונת החיפוש הוא הממוצע של מספר מרחקי מאפיינים (לאחר נירמול) ומידת הרלוונטיות מוגדרת להיות הממוצע של מספר נקודות הדמיון לתמונת השליפה. ConvLinear מרחיב את CombSumScore ע"י שימוש בפונקציית משקל לינארית על מנת לשלב את מרחקי מאפיינים.  
       לצורך ההשואה נעשה שימוש בחמישה מאפיינים סטנדרטים של MPEG-7 יחד עם אומדני המרחק המומלצים. לכל תמונה סופקו חמישה דוגמאות תיוג חיוביות.יש לציין כי הטיפול בדוגמאות המרעישות אינו משפיע לרעה על תהליך האיחזור משום שכמות התמונות לטיפול קטנה, כלומר זמן זה יחסית זניח.  
       ההשואה נערכה על שני מאגרי תמונות מוכרים, ע"י כ-300 חיפושים ומדדי היעילות של השיטה הם דיוק והחזר.

1. בחינת השפעת הדוגמאות החיוביות המרעישות  
   על מנת לבחון ולהגדיש את השפעת דוגמאות חיוביות מרעישות הוכנסו באופן ייזום תיוגים שגויים לשליפה. עבור ניסוי זה נבחרה שיטת CombSumScoreבשל פשטותה. התוצאות מראות כי לתיוגים שגויים יש השפעה דרמטית על טיב האיחזור. (איור 14)

איור 15.בחינת השפעת רעש על טיב האיחזור  
<מספר דוגמאות שגויות>n  
<מספר דוגמאות נכונות>p

1. בחינת מודל קיבוץ המאפיינים  
   בחינת מודל הקיבוץ ע"י השוואת SVM-Weighted-MVR, SVM-Weighted-BSR ו- Weighted-BSR. ניתן לראות כי Weighted-BSR יעיל יותר משתי השיטות האחרות, הסיבה לכך יכולה להיות שעבור מקרים בהם יש מעט מידי דוגמאות SVM חלש מידי עבור מדידת רלוונטיות.

איור 16. בחינת מודל קיבוץ המאפיינים

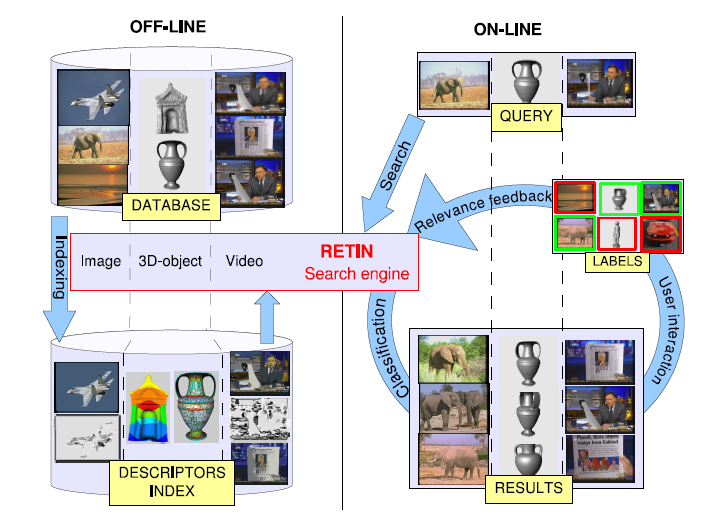
1. בחינת מודל האיחזור  
   בניסוי זה בוחנים את מודל האיחזור בנוכחות דוגמאות תיוג שגויות. מודל Weighted-BSR נבחר על בסיס הניסוי הקודם. ניתן לראות מאיור 16 ו-17 כי השיטה המוצעת יעילה יותר מ CombSumScore ו- ConvLinear בייחוד כאשר מספר דוגמאות התיוג השגויות עולה.  
   * 1. מסקנותהוצגה שיטת שיפור איחזור תמונה בעזרת משוב רלוונטיות המכיל דוגמאות תיוג מרעישות ודרך להתגבר על כך ע"י אימון רכיב סיווג SVM.  
         בוצעה הקטנת מרחב המאפיינים ע"י קיבוץ מאפיינים, הנוסח כבעיית סיווג, לצורך כך נבנה רכיב SVM. הטיפול בדוגמאות התיוג המרעישות מבוצע בשני שלבים, בשלב הראשון הרכבה של רכיבי SVM נבנת על מנת לזהות ולסלק דוגמאות חיוביות מרעישות (שגויות) ובשלב השני חישוב רלוונטיות עמיד לרעש מבוצע. הוצגה השוואה בין השיטה המתוארת לשיטות קיבוץ מאפיינים מקובלות, ומסתמן כי השיטה המוצגת יעילה יותר כאשר קיימות דוגמאות תיוג מרעישות.   
          
        \* יש לציין כי אין פירוט רב על המאפיינים שחולצו עבור תמונות המאגר ובהם בוצע שימוש.

איור 18. בחינת סכמת האיחזור בעזרת Weighted-BSR על מאגר תמונות נוסף

איור 17. השוואת טיב האיחזור בין השיטות : (a) ללא תיוג מרעיש,   
(b) דוגמאת תיוג מרעישה אחת, (c) שתי דוגמאות תיוג מרעישות

1. הצגת מימוש Supervised SVM :

בחלק זה נציג מערכת המבצעת שימוש בSVM , מערכת איחזור מולטימדיה העונה לשם "RETIN". מערכת זו מאפשרת לחפש במאגרי מידע נרחבים ובבסיסה תהליך סיווג הנבנה לפי תיוג משתמש. סיווג זה נעשה ע"י שיטות גרעין עבור דמיון וSVM עבור בעזרת תיוג המשתמש. בנוסף לכך מערכת זו מציעה תוכן כדאי לתיוג המשתמש ע"י שימוש בלמידה. המערכת מטפלת בתמונות (בהן נתמקד), אובייקטים תלת מימדיים, סרטים וגם סוגי תכנים אחרים ניתנים להוספה (מולטימידה).   
 מערכת RETIN מורכבת משני חלקים, חלק לא מקוון וחלק מקוון. בחלקה הראשון, הלא מקוון מתבצע אינדוקס המאגר ובו מחלצים מאפיינים שונים לתוכן ושומרים אותם במסד נתונים מאונדקס. חלקה השני, המקוון הוא מנגנון החיפוש, הכולל מנוע חיפוש ונגיש למשתמש בצורת ממשק גרפי ומורכב מחלק עבור השליפה, חלק עבור התוצאות ותיוג משוב רלוונטיות.



איור 19. סכמת מערכת RETIN

Initialization

Classification

Correction

Display

User  
Satisfaction?

End

Selection  
Pre-selection  
Criteria  
Diversification

Display

User labels

Yes

No

כאמור, המערכת משתמשת בלמידה מבוססת סיווג על מנת לשפר את תוצאות האיחזור עבור המשתמש. פירוט השלבים :

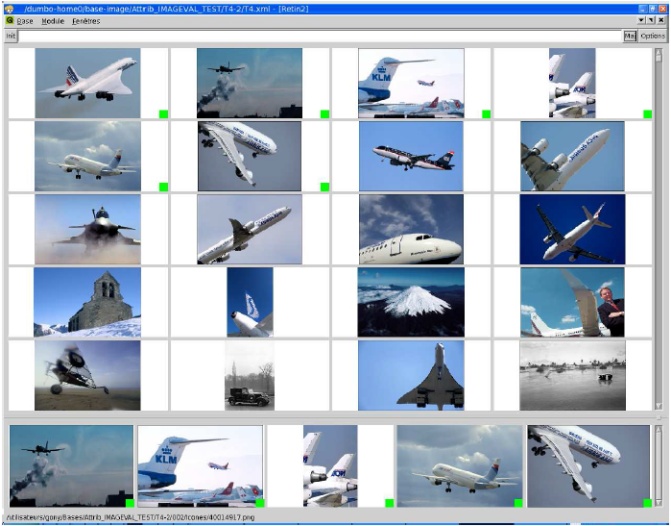
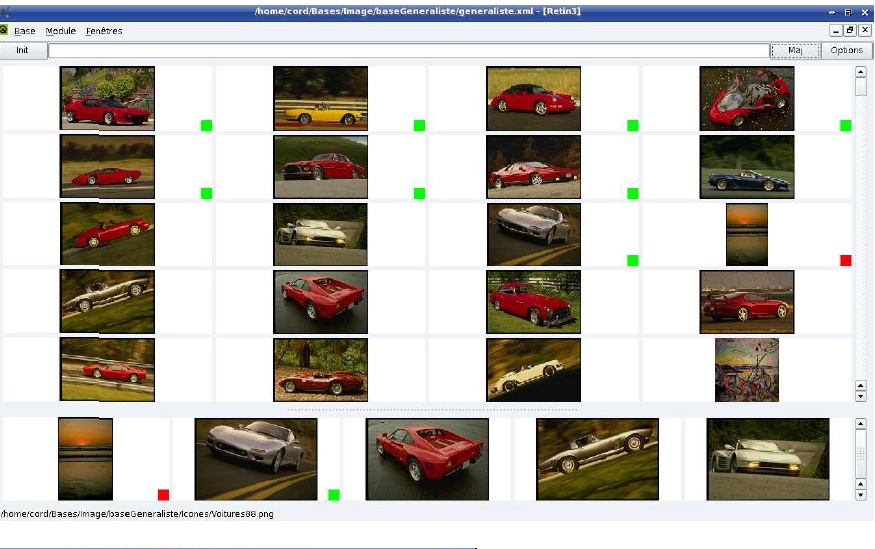
1. איתחול(Initialization) – תהליך האיחזור מתחיל עם שליפת משתמש המספק דוגמא לתוכן בו הוא מעוניין. מחשבים מאפיינים עבור התוכן ומכניסים זאת לבסיס הנתונים. תוכן זה מתוייג כתוכן רלוונטי והתוצאות הקרובות ביותר מוצגות למשתמש.
2. סיווג (Classification) – רכיב סיווג בינארי מאומן באמצעות התיוגים הניתנים ע"י המשתמש. רכיב הסיווג המאומן הוא רכיב SVM עם גרעין Gaussian , התוצאה היא פונקציית סיווג המחזירה את מידת הרלוונטיות לכל מסמך תוכן בהתאם לדוגמאות הניתנות .

איור 20. סכמת הלמידה במערכת RETIN

1. בחירה ייזומה (Active selection) – כיוון שדוגמאות התיוג מועטות והחלוקה בין התוכן הרלוונטי לתוכן הלא רלוונטי אינה מאוזנת, נוסף שלב של תיקון הגבול בין המחלקות רלוונטי \ לא רלוונטי. מטרת שלב זה היא להדגיש את החלוקה, שכן "הגבול" הוא תחום של אי וודאות. המערכת בוחרת את סט מסמכי התוכן מהמאגר אותם רצוי שהמשתמש יתייג. הבחירה של סט מסמכי התוכן צריכה להיות כזו שבהינתן תייוג המשתמש על סט זה, החלוקה בין המחלקות בסיווג תהיה הכי מודגשת. בחירה זו מתבצעת בשלושה חלקים : ראשית על מנת להקטין את זמני החישוב נבחרים תחילה כ-500 מסמכי תוכן העשויים להיות האופטימליים (כאלו שעשויים לשפר את החלוקה). לאחר מכן מחשבים את קריטריון החלוקה (הבאת הדיוק הממוצע למקסימום שניתן) לכל מסמך תוכן שנבחר מראש. לבסוף מבצעים את החלוקה ע"י הקריטריון.
2. משוב (Feedback) – המשתמש מתייג את סט המסמכים שנבחר עבורו ובעזרת תייוג זה ניתן לחשב סיווג ותיקון חדשים. חוזרים על התהליך כמספר הפעמים שנדרש.

תהליך מיפתוח (אינדוקס) עבור תמונות  
כל מסמך תוכן במאגר מיוצג ע"י מאפיינים השואפים לתת תיאור מלא וקומפקטי לתוכן הגלום במסמך באופן כזה שיהיה קל לעיבוד וביצוע חישובים. המאפיינים המחולצים הם מאפייני צבע וטקסטורה. מאפייני הצבע מיוצגים במודל HSV ומאפייני הטקסטורה מחולצים ע"י פילטרי Gabor[[8]](#footnote-8) בעזרת שניים עשר גדלים וכיוונים שונים. בעזרת מאפיינים אלו ניתן למקם את התמונות במרחב המאפיינים וע"י קיבוץ לחשב היסטוגרמות תמונה. מכך אפשר ליצור עבור התמונות חתימות שהן התפלגות סטטיסטית לצבעים ולטקסטורות במרחב המאפיינים. עבור מאגר בגודל סביר (כ 10,000 תמונות) דיי בכמות מחלקות קטנה על מנת ליצור חתימות אפקטיביות.

ממשק המשתמש  
חלון השליפה מחולק לשני חלקים, בחלקו העליון מוצגות התמונות בסדר רלוונטיות יורד בהתאם לרכיב הסיווג הנוכחי. בחלקו התחתון של חלון השליפה מוצגות התמונות המוצעות לתיוג המשתמש.



1. סיכום

איחזור תמונה מבוסס תוכן הינו תחום מורכב ומאתגר, ובו נצבים מכשולים החל בשמירת מאגר התמונות באופן יעיל לאיחזור, התגברות על הפער בין ייצוגה של תמונה במחשב לבין ראיתו של המשתמש וכלה בזמן הדרוש לתהליך. לצורך כך ניתן לרתום מנגנוני למידה חישובית על מנת לשפר את טיב האיחזור ובעיבוד מקדים מתאים הן לשיפור טיב האיחזור והן לשיפור זמן האיחזור.   
 בעבודה זו הוצגה סקירה קצרה על שיטות למידה חישובית עיקריות בתחום והרקע הדרוש להבנתן ונבחנה שיטת מכונת וקטורים תומכים לעומק בהקשר משוב רלוונטיות. הוצגו שתי שיטות לשימוש במשוב רלוונטיות ללמידת רצון המשתמש וטיוב האיחזור בעזרתו ונראה כי שיטות אלו מביאות לשיפור לא מבוטל בדיוק האיחזור. עם זאת במערכות הנעזרות במשוב רלוונטיות מהמשתמש מתעוררות בעיות נוספות בהן תיוג מוגבל ותיוג שגוי של המשתמש עלול לפגוע בתהליך הלמידה ובעקבות כך באיחזור. השיטה השנייה מציגה דרך להתמודד עם בעיה זו.  
לאחר מכן הוצגה מערכת RETIN העושה שימוש במכונת וקטורים תומכים כחלק מתהליך האיחזור.  
 יש לציין כי חילוץ המאפיינים ובחירת המאפיינים חשובה לא פחות מבחירת אלגוריתם הלמידה שכן המאפיינים המחולצים הם בסיסו של תהליך הלמידה, להם יש השפעה ישירה על כל חלקי מערכת האיחזור – שמירה במאגר, השוואת תמונות ובעקבות זאת על תהליך הלמידה עצמו.

1. מקורות

Ayodele, T. O. (2010). Types of Machine Learning Algorithms. In Y. Zhang, *Types of Machine Learning Algorithms, New Advances in Machine Learning* (pp. 19-48). InTech.

Buckland, M. K., & Gey, F. C. (1994). The relationship between recall and precision. *Journal of the Association for Information Science and Technology(JASIS) 45.1* , 12-19.

Eakins, J. P. (2002). Towards intelligent image retrieval. *Pattern Recognition 35.1* , 3-14.

Ivica Dimitrovski, B. J. (2010). Efficient Content-based Image Retrieval Using Weighted Feature Aggregation Scheme. In K. E. Tarek Sobh, *Innovations in Computing Sciences and Software Engineering* (p. 321). Springer Science & Business Media.

Ma, Z. (2009). *Artificial Intelligence for Maximizing Content Based Image Retrieval.* Hershey, PA: IGI Global.

Marakakis, A. e. (2009). Relevance feedback for content-based image retrieval using support vector machines and feature selection. *Artificial Neural Networks–ICANN* , 942-951.

Matthieu Cord, P. C. (2008). *Machine Learning Techniques for Multimedia, Case Studies on Organization and Retrieval.* Berlin: Springer.

Stuart Russell, P. N. (2010). *Artificial Intelligence: A Modern Approach.* Prentice Hall Pa.

Tinku Acharya, A. K. *Image processing Principles and Applications.*

Zhang, J. a. (2009). Content based image retrieval using unclean positive examples. *IEEE transactions on image processing* , 2370-2375.

1. שונות - מדד לפיזור הערכים עבור משתנה מסויים, שונות נמוכה מצביעה על כך שהערכים קרובים לחציון, שונות גבוהה מצביעה על פיזור גבוה של הערכים מהחציון וזה מזה. [↑](#footnote-ref-1)
2. שונות משותפת – מדד לקשר בין שני משתנים, השונות חיובית אם לשניהם אותה מגמה ושלילית אם לשניהם מגמות מנוגדות. [↑](#footnote-ref-2)
3. , מוגדרת להיות קבוצת הערכים עבורה מקבלת מקסימום [↑](#footnote-ref-3)
4. התמרת לפלס – [↑](#footnote-ref-4)
5. אנטרופיה (Entropy) – מדד לאי הוודאות (הממוצעת) של משתנה מקרי. [↑](#footnote-ref-5)
6. מידת העיוות (distortion) – המידה בה פונקציה (מהמישור האוקלידי למישור האוקלידי) משנה או מעוותת גודל וצורה למשל מעגלים לאליפסות. [↑](#footnote-ref-6)
7. פונקציית sigmoid ולה צורת S [↑](#footnote-ref-7)
8. פילטרי Gabor – פילטר לינארי המשמש לזיהויי קצוות (Edge detection). [↑](#footnote-ref-8)