یک روش مبتنی‌بر خوشه‌بندی سلسله‌مراتبی تقسیم‌کننده جهت شاخص‌گذاری اطلاعات تصویری

محمّدرضا کیوان‌پور[1]، سعیده رنجبران[2] و نجوا ایزدپناه[3]
[1]دانشکدهٔ فنی و مهندسی، دانشگاه الزهرا(س)،تهران، ایران
[2و3] دانشکدهٔ برق، رایانه و فناوری اطّلاعات،دانشگاه آزاد اسلامی قزوین، قزوین،ایران

چکیده

در سامانه‌های رایج بازیابی تصویر مبتنی‌بر محتوا از ساختارهای شاخص‌گذاری چندبعدی برای سرعت‌بخشیدن به عملیات جستجو استفاده می‌شود. در اکثرحوزه‌های کاربردی، ابعاد بالایی از بردارهای ویژگی چندبعدی برای توصیف تصاویر موردنیاز است؛ اما ساختارهای شاخص‌گذاری چندبعدی رایج کارایی خود را با افزایش ابعاد فضای ویژگی از دست می‌دهند. افزایش ابعاد فضای داده موجب افزایش نمایی اندازهٔ فضای جستجو و تعداد گره‌ها در ساختارهای شاخص‌گذاری چندبعدی و همچنین موجب افزایش هم‌پوشانی بین گره‌های ساختارهای شاخص‌گذاری چندبعدی می‌شود. این مسائل منجر به افزایش هزینهٔ جستجو از طریق ساختارهای شاخص‌گذاری چندبعدی رایج می‌شود. هدف این پژوهش ارائهٔ یک ساختار شاخص‌گذاری تصویر مبتنی‌بر خوشه‌بندی سلسله‌مراتبی تقسیم‌کننده، جهت مدیریت بردارهای ویژگی متناظر با تصاویر، در فضای ویژگی با ابعاد بالاست که درضمن از هم‌پوشانی نواحی گره‌ها نیز جلوگیری به‌عمل آورد. آزمون‌های مختلف نشان داده که ساختار شاخص‌گذاری پیشنهادی در فضاهای با ابعاد بالا کارایی مناسبی داشته و نسبت به رویکردهای پیشین دارای برتری است.

واژگان کلیدی: بازیابی تصویر مبتنی‌بر محتوا،ساختارهای شاخص‌گذاری چندبعدی، خوشه‌بندی سلسله‌مراتبی تقسیم‌کننده، روش‌های پیگرد افکنش.

۱- مقدّمه

درسال‌های اخیر با رشد فناوری و افزایش توانایی رایانه‌ها در مدیریت داده‌های تصویری و حجم وسیع تصاویر تولیدشده در حوزه‌های مختلف، طراحی و پیاده‌سازی پایگاه‌داده‌های تصویری و بازیابی از آنها به‌عنوان یک الزام مطرح است (Datta et al., 2008, Srinivasa Rao et al., 2010, Bohm et al., 2001). به‌طور متداول، اغلب سامانه‌های بازیابی تصویر به‌عنوان یک راهکار فعل و انفعالی برای بازیابی محتوا و تفسیر نیمه‌خودکار تصویر به‌کارمی‌روند (Asbaghi et al., 2008). انواع مختلفی از کاربردهای بازیابی تصویر مطرح است؛ برای مثال، بخشی از تلاش‌های صورت‌گرفته توسط محققان به بازیابی تصاویر متنی[1] اختصاص یافته است (Keyvanpour and Tavoli, 2013, Keyvanpour and Tavoli, 2012). روش‌های بازیابی تصویر مبتنی‌بر محتوا و مبتنی‌بر متن از روش‌های بنیادین برای بازیابی اطلاعات تصویری محسوب می‌شوند (Keyvanpour and Asbaghi, 2008). در بازیابی تصویر مبتنی‌بر محتوا، ویژگی‌های بصری[2] شامل ویژگی رنگ[3]،ویژگی بافت[4]،ویژگی شکل[5] و ویژگی‌های محلی[6] از هر تصویر استخراج می‌شوند و در قالب بردارهای ویژگی چندبعدی[7] سازماندهی می‌شوند. در مرحلهٔ جستجو، وقتی کاربر پرس‌وجوی خود را با انتخاب تصویر موردنظرش مطرح می‌کند، یک بردار ویژگی برای تصویر موردجستجو[8] محاسبه می‌شود؛ سپس، بااستفاده از

1Document Image Retrieval

[2]Visual features
[3]Color feature
[4]Texture feature
[5]Shape feature
[6]Local features
[7]Mult-dimensional Feature vectors
[8]Image query

معیارهای شباهت[1]، بردار ویژگی تصویر مورد جستجو با بردارهای ویژگی متناظر با تصاویر پایگاه‌داده مقایسه شده و تصاویری که بیشترین شباهت را با بردار تصویر موردجستجو دارند به‌عنوان نتایج برگردانده می‌شوند (Datta et al., 2008, Lew et al., 2006).

یکی از مسائل مطرح در حوزهٔ بازیابی تصویر مبتنی‌بر محتوا، به‌کارگیری روش‌هایی برای سرعت‌بخشیدن به عملیات جستجو در پایگاه‌داده‌های تصویری است. بدین منظور، از ساختارهای شاخص‌گذاری چندبعدی[2] برای سازماندهی بردارهای ویژگی چندبعدی تصویر استفاده می‌شود و سپس جستجوی پایگاه‌داده تصویری ازطریق ساختار شاخص‌گذاری برای رسیدن به سرعت مناسب در فرآیند بازیابی انجام می‌شود (Gaede and Gunther, 1998, Markov et al., 2008, Chakrabarti and Mehrotra, 1999). تاکنون تلاش‌های گوناگونی به‌منظور توسعه ساختارهای شاخص‌گذاری چندبعدی انجام گرفته است (Bohm et al., 2001, Gaede and Gunther, 1998, Keyvanpour and Izadpanah, 2011(a), Keyvanpour and Izadpanah, 2011(b), Li et al., 2002)؛ اما باتوجه به حجم وسیع تصاویر در پایگاه‌داده‌های تصویری و ابعاد بالای بردارهای ویژگی تصاویر، ارائهٔ یک ساختار شاخص‌گذاری با کارایی مطلوب بسیار دشوار است.

افزایش ابعاد در حوزهٔ ساختارهای شاخص‌گذاری چندبعدی بر پیچیدگی محاسباتی افزوده و هم‌گرایی روش جستجو را دچار اختلال می‌کند. این مسائل منجر به افزایش هزینهٔ جستجو و درنتیجه کاهش کارایی این ساختارها در پایگاه داده‌های چندبعدی با ابعاد بالامی‌شود. بنابراین ارائه ساختارهایی برای اداره داده‌های با ابعاد بالا ضروری است. (Keyvanpour and Izadpanah, 2011(b)) دسته‌بندی کلی از ساختارهای شاخص‌گذاری چندبعدی در فضاهای با ابعاد بالا را ارائه داده است.

تاکنون الگوریتم‌های خوشه‌بندی فراوانی جهت خوشه‌بندی داده‌های چندبعدی ارائه شده‌اند؛ اما تلاش‌های اندکی به‌منظور ارائه روش‌های خوشه‌بندی برای به‌کارگیری آنها، جهت شاخص‌گذاری اطلاعات تصویری صورت‌گرفته است (Li et al., 2002). محققان الگوریتم‌های خوشه‌بندی را به‌طور کلی به چهار دسته الگوریتم‌های افراز[3]، الگوریتم‌های سلسله‌مراتبی[4]، الگوریتم‌های مبتنی‌بر توری[5] و الگوریتم‌های مبتنی‌بر تراکم[6]، طبقه‌بندی کرده‌اند (Xu et al., 2007). الگوریتم‌های سلسله‌مراتبی به دو نوع سلسله‌مراتبی تقسیم‌کننده[7] شامل الگوریتم PDDP(Boley, 1998)، و سلسله‌مراتبی تجمیعی[8] شامل الگوریتم‌های BIRCH(Zhang et al., 1996)و CURE(Guha et al., 1998)، تقسیم شده‌اند که با هدف تجزیه فضای داده در یک ساختار سلسله‌مراتبی توسعه یافته‌اند.

هدف از این پژوهش، ارائهٔ یک ساختار شاخص‌گذاری چندبعدی مبتنی‌بر خوشه‌بندی سلسله‌مراتبی تقسیم‌کننده[9]، جهت مدیریت بردارهای ویژگی متناظر با تصاویر، در فضای ویژگی با ابعاد بالاست که درضمن از هم‌پوشانی نواحی گره‌ها نیز جلوگیری به‌عمل می‌آورد. بدین منظور از الگوریتم FastICA (Hyvarinen and Oja, 2000)، با هدف پیداکردن یک مؤلّفهٔ غیرگوسی معنادار از داده‌ها، که افکنش داده‌ها[10] روی آن تقریب آشفتگی منفی[11] را بیشینه کرده و درنتیجه افکنش داده‌ها روی آن برای تجزیه داده‌ها به دو زیرخوشه می‌تواند حاوی اطلاعات مفیدی باشد، استفاده شده است. در ادامه ابتدا در بخش۲، نحوهٔ عملکرد سامانهٔ بازیابی تصویر براساس ساختار شاخص‌گذاری پیشنهادی، و سپس در بخش۳، ساختار شاخص‌گذاری چندبعدی پیشنهادی (NO-NGP-tree[12]) مطرح شده است. دربخش۴، پیاده‌سازی و آزمون مطرح شده است. و درنهایت بخش۵ به نتایج و فعالیت‌های آینده پرداخته است.

۲- نحوهٔ عملکرد کلی سامانهٔ بازیابی تصویر براساس ساختار شاخص‌گذاری پیشنهادی

سامانهٔ بازیابی تصویر براساس ساختار شاخص‌گذاری پیشنهادی، مانند معماری کلی سامانه‌های بازیابی تصویر مبتنی‌بر محتوای ارائه‌شده در(Rui and Huang, 1999)، ازچهار قسمت اصلی استخراج ویژگی، مجموعه‌های داده، موتور بازیابی و طرح شاخص‌گذاری چندبعدی تشکیل شده

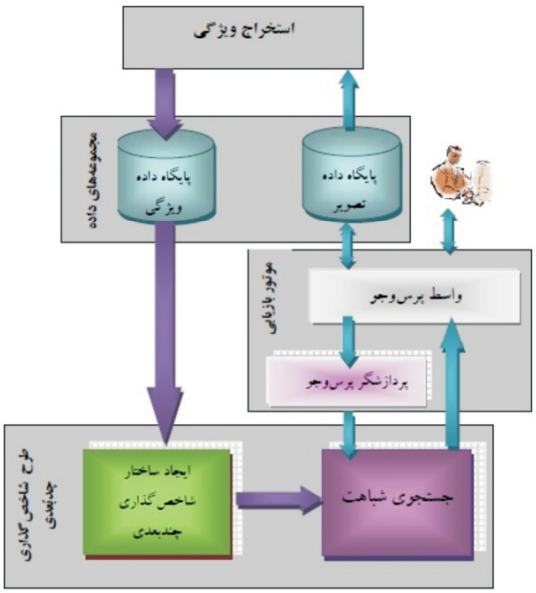
1 Similarity measures
2 Multi-dimensional indexing structures
3 Partitioning algorithms
4 Hierarchical algorithms
5 Grid-based algorithms
6 Density-based algorithms
7 Divisive
8 Agglomerative
9 Hierarchical divisive clustering
10 Data projection
11 Negentropy
12 No Overlapping-Non Gaussian Projection based-tree indexing

است. شکل(۲-۱) معماری پایه سامانهٔ بازیابی تصویر و جایگاه ساختار شاخص‌گذاری پیشنهادی را نشان می‌دهد.

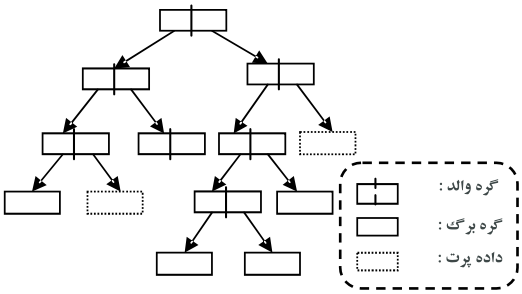
بردارهای ویژگی تصویر، در بخش استخراج ویژگی، از تصاویر درون پایگاه‌داده تصویری استخراج و در پایگاه‌داده ویژگی ذخیره می‌شوند. کاربر پرس‌وجوی خود را از طریق واسط پرس‌وجو مطرح می‌کند، تصویر موردجستجو توسط پردازش‌گر پرس‌وجو به بردار ویژگی چندبعدی تبدیل می‌شود. در قسمت طرح شاخص‌گذاری چندبعدی، در یک مرحلهٔ برون‌خط[1] کلیه بردارهای ویژگی تصویر از پایگاه‌داده ویژگی گرفته می‌شوند و در ساختار شاخص‌گذاری مبتنی‌بر خوشه‌بندی سلسله‌مراتبی تقسیم‌کننده NO-NGP-tree سازماندهی می‌شوند؛ سپس در مرحله جستجوی شباهت به صورت برخط[2]، بردار ویژگی تصویر موردجستجو گرفته می‌شود و ازطریق یک الگوریتم جستجوی شباهت بازگشتی، ساختار شاخص‌گذاری NO-NGP-tree موردجستجو قرارگرفته و k-نزدیک‌ترین بردارهای ویژگی به بردار پرس‌وجو برای نمایش تصاویر بازیابی‌شده به واسط پرس‌وجو فرستاده می‌شوند.

در این پژوهش، حاصل بخش شاخص‌گذاری مبتنی‌بر خوشه‌بندی یک ساختار درختی باینری نامتوازن است (شکل(۲-۲))، که در بالاترین سطح - یا به‌عبارتی ریشه - شامل یک گره است که کل بردارهای ویژگی تصویر چندبعدی را شامل می‌شود.

در ساختار پیشنهادی، گره‌هایی که دارای دو فرزند هستند گره‌های والد، و گره‌هایی که فرزندی ندارند، درصورتی که تعداد داده‌های درون آن از یک حد آستانه، که با مشخصهٔ پیش‌نیاز Minpts تعیین می‌شود، کمتر باشند، "داده پرت[3]" و در غیراین‌صورت، "گره برگ" می‌گویند. به‌عبارت دیگر، Minpts حداقل تعداد داده‌ها در گره‌های برگ را مشخص می‌نماید. در بخش جستجوی شباهت، از الگوریتم‌های جستجوی نزدیک‌ترین همسایهٔ ارائه‌شده در (Taileb et al., 2007) برای جستجوی شباهت استفاده شده است.



(شکل(۲-۱)): معماری پایهٔ سامانهٔ بازیابی تصویر و جایگاه ساختار شاخص‌گذاری پیشنهادی



(شکل(۲-۲)): ساختار شاخص‌گذاری پیشنهادی NO-NGP-tree

۳- ساختار شاخص‌گذاری چندبعدی پیشنهادی NO-NGP-tree

در سامانهٔ بازیابی تصویر براساس ساختار شاخص‌گذاری پیشنهادی، هر تصویر با یک بردار ویژگی چندبعدی نمایش داده می‌شود و بردارهای ویژگی تصاویر در پایگاه‌داده ویژگی ذخیره می‌شوند. بنابراین پایگاه‌داده ویژگی را می‌توان به‌صورت ماتریس $M_0(n, m)$ درنظرگرفت. ستون‌های ماتریس M_0، بردارهای ویژگی می‌باشند. n برابرتعداد ابعاد بردارهای ویژگی و m برابر تعداد بردارهای ویژگی درون پایگاه داده ویژگی است. به‌عبارت دیگر، ماتریس M_0، ماتریس

[1] Out-line
[2] On-line
[3] Outlier data

حاوی بردارهای ویژگی چندبعدی استخراج‌شده از تصاویر (ماتریس مربوط به ریشه) است. برای ایجاد ساختار شاخص‌گذاری NO-NGP-tree، ماتریس M_0 با استفاده از یک الگوریتم تکرارکننده[1] به‌صورت بازگشتی افرازبندی می‌شود. در هر تکرار یک ساختار باینری ناتمام تشکیل می‌شود. روال تکرارکننده تا زمانی‌که تعداد گرهٔ برگ موردنظر (k) ایجاد شود، ادامه پیدا می‌کند.

شبه‌کد مربوط به فرآیند ایجاد ساختار شاخص‌گذاری چندبعدی پیشنهادی در ادامه، و فرآیند ایجاد ساختار شاخص‌گذاری چندبعدی پیشنهادی در شکل (۳-۱) نشان داده شده‌اند. همان‌طور که در شکل (۳-۱) مشخص است، فرآیند ایجاد ساختار شاخص‌گذاری چندبعدی پیشنهادی به‌صورت تکرارکننده از چهار مرحلهٔ اصلی شامل: پیش‌افرازبندی، افرازبندی، مرزبندی و ایجاد ساختار درختی تشکیل شده است. این مراحل در شبه‌کد زیر، به‌ترتیب با توابع Pre-Partitioning()، Partitioning()، Bounding()، Build tree structure() نشان داده شده‌اند.این مراحل تا زمانی‌که تعداد برگ تعیین‌شده توسط مشخصهٔ پیش‌نیاز k ایجاد شود، به‌صورت بازگشتی تکرار می‌شوند.

```
NO-NGP-tree (Mi, K, Minpts)
Begin
i =1
while LNoi<k
  {
    {Ii}= Pre-partitioning(Mi)
    {(MRi)s, (MLi )s}=Partitioning (Mi,Ii)
    {(BRi)s,(BLi)s,(MRi)s, (MLi )s}= Bounding ((MRi)s, (MLi)s,
    (ai)s)
    {TSi}= Build tree structure((BRi)s,(BLi)s ,(MRi)s, (MLi)s)
    i=i+1
  }
Return (NO-NGP-tree)
End
```

ماتریس M_i مجموعه ماتریس‌های مربوط به برگ‌ها در تکرار iام است. به‌عبارت دیگر، در هر تکرار، مجموعه ماتریس‌های مربوط به برگ‌ها در تکرار i (M_i) دریافت می‌شود. M_i طبق رابطهٔ زیرتعریف می‌شود.

(۱)

$$M_i = \{M_{i-1} - (M_{i-1})_s\} \cup (MR_{i-1})_s \cup (ML_{i-1})_s - outlier_{i-1}$$

در رابطهٔ (۱)، M_{i-1} حاوی مجموعه برگ‌ها در تکرار i-1، $(M_{i-1})_s$ ماتریس مربوط به برگ انتخاب‌شده در تکرار i-1، $(MR_{i-1})_s$ و $(ML_{i-1})_s$ به‌ترتیب زیرخوشه‌های راست و چپ خوشهٔ انتخاب‌شده در تکرار i-1 هستند و outlieri-1، زیرخوشه‌ای با تعداد داده کمتراز مقدار تعیین‌شده توسط پارامتر پیش‌نیاز Minpts در تکرار i-1 است. $(I_i)_j$ طبق رابطه زیر تعریف می‌شود.

(۲)

$$(I_i)_j=\{F_j,(a_i)_j,IDX1,IDX2,CP1,CP2\}$$

در رابطهٔ (۲)، $(I_i)_j$ اطلاعات استخراج‌شده از $(M_i)_j$ است، که ماتریس $(M_i)_j$، ماتریس مربوط به عضو jام مجموعهٔ M_i می‌باشد، و j شمارندهٔ تعداد خوشه‌های موجود در مجموعهٔ M_i است. F_j بردار افکنش ماتریس[2] مربوط به jامین برگ، $(a_i)_j$ مؤلّفه غیرگوسی معنادار به‌دست‌آمده از ماتریس مربوط به‌jامین برگ، در مجموعه برگ‌های تکرار i موجود در مجموعه M_i، IDX1,IDX2 برابر زیرخوشه‌های افکنش وCP1,CP2 برابر مراکز زیرخوشه‌های افکنش هستند.

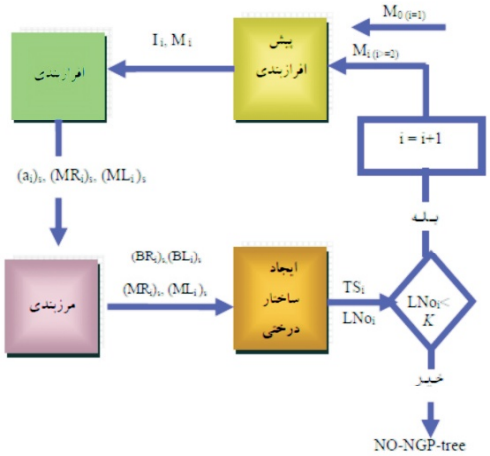
لازم به ذکر است که در اولین تکرار الگوریتم (i=1)، ماتریس M_0 حاوی کل بردارهای ویژگی تصویر، به‌عنوان ورودی درنظرگرفته می‌شود. مرحله پیش‌افرازبندی، عملیات لازم را برای به‌دست‌آوردن اطلاعات I_i موردنیاز برای انتخاب و تجزیهٔ خوشهٔ بعدی به دو زیرخوشه در مرحلهٔ افرازبندی فراهم می‌کند. I_iاطلاعات استخراج‌شده از مجموعه M_i را دربرمی‌گیرد. سپس در مرحلهٔ افرازبندی، خوشهٔ بعدی برای افرازبندی از مجموعه M_i انتخاب می‌شود و به دو زیرخوشهٔ راست $(MR_i)_s$ و زیرخوشهٔ چپ $(ML_i)_s$ تجزیه می‌شود.

از آن‌جا که افرازبندی براساس افکنش داده‌ها روی مؤلّفهٔ غیرگوسی معنادار به‌دست‌آمده از ماتریس مربوط به خوشهٔ انتخاب‌شده $((a_i)_s)$ صورت می‌گیرد، که در مرحلهٔ پیش‌افرازبندی به‌دست‌آمده، مؤلّفهٔ $(a_i)_s$ به مرحلهٔ مرزبندی زیرخوشه‌ها فرستاده می‌شود. در مرحلهٔ مرزبندی، زیرخوشه‌های راست و چپ توسط مستطیل‌های مرزبندی کمینه‌ای که در جهت مؤلّفهٔ $(a_i)_s$ گسترش می‌یابند، مرزبندی می‌شوند. نتایج مرحلهٔ مرزبندی $(BR_i)_s$, $(BL_i)_s$ هستند، که به‌ترتیب شامل اطلاعات مربوط به مرزبندی زیرخوشهٔ راست و زیرخوشهٔ چپ خوشهٔ انتخاب‌شده در تکرار i می‌باشند؛ سپس در مرحله تشکیل ساختار درختی، دو زیرخوشه راست و چپ ایجادشده در تکرار i بررسی شده و درصورتی‌که تعداد اعضای آن خوشه‌ها از حد آستانهٔ تعیین‌شده کمتر باشد، به‌عنوان داده‌های پرت و در غیر این‌صورت به‌عنوان گرهٔ برگ علامت زده می‌شوند. در هر دو

[1]Iterative

[2]Matrix projection vector

حالت $(BR_i)_s$ و $(BL_i)_s$ ذخیره می‌شوند؛ سپس، Ts_i که نشان‌دهنده درخت باینری تشکیل‌شده در تکرار i است، به‌همراه LNO_i نتایج این مرحله را تشکیل می‌دهند. LNO_i تعداد برگ‌های ساختار درختی باینری تشکیل‌شده در تکرار i است.



(شکل(۳-۲)): فرآیند ایجاد ساختار شاخص‌گذاری چندبعدی پیشنهادی NO-NGP-tree

۳-۱-پیش‌افرازبندی

در این مرحله، اطلاعات I_i، شامل اطلاعات مورد نیاز برای انتخاب و تجزیهٔ خوشهٔ بعدی به دو زیرخوشه در مرحلهٔ افرازبندی فراهم می‌شود. مرحلهٔ پیش‌افرازبندی، مجموعه M_i شامل ماتریس‌های مربوط به برگ‌های موجود در تکرار i-1 را می‌گیرد و برای هریک از اعضای مجموعه M_i، طی سه مرحله، اطلاعات $(I_i)_j$ به‌دست می‌آید و اطلاعات نهایی برای j=1 تا $j=LNO_i$ در مجموعه I_i قرارمی‌گیرند و همراه مجموعه M_i به مرحله افرازبندی فرستاده می‌شوند (jشمارندهٔ تعداد برگ‌ها در M_i است، بنابراین $(M_i)_j$ برابر jامین برگ در مجموعه M_i می‌باشد).

فرآیند پیش‌افرازبندی از سه زیرمرحله که برای j=1 تا $j=LNO_i$ تکرارمی‌شوند، تشکیل شده است. در این شبه‌کد، زیرمراحل به‌ترتیب با توابع Find meaningfulnon-guassian component()، Projection() و 2-mean clustering() نشان داده شده‌اند.

یافتن مؤلفهٔ غیرگوسی معنادار: هدف این مرحله، پیداکردن مؤلفه $(a_i)_j$ از داده‌های ماتریس $(M_i)_j$ است، به‌طوری که این مؤلفه در جهتی قرار گرفته باشد که بتواند به‌شکل مطلوبی ساختار داده‌ها و یا به‌عبارتی خوشه‌های طبیعی موجود در توزیع داده‌ها را نشان دهد.

افکنش: هدف از این مرحله، افکنش داده‌های ماتریس $(M_i)_j$ روی مؤلفه به‌دست‌آمده در مرحلهٔ یافتن مؤلفه غیرگوسی معنادار از داده‌هاست.

خوشه‌بندی داده‌های افکنش‌شده: هدف این مرحله، به‌کارگیری الگوریتم خوشه‌بندی 2-means روی داده‌های افکنش‌شده در مرحلهٔ افکنش، برای به‌دست‌آوردن دو زیرخوشه به نام خوشه‌های افکنش و مراکز ثقل آنهاست.

```
Pre-partitioning (Mi)
Begin
For j=1 to LNOi
  {
{(ai)j} = Find meaningful non-gaussian component ((Mi)j)
{Fj} = Projection ((ai)j, (Mi)j):
  {CP1, Cp2, IDX1, IDX2} = 2-mean clustering (Fj)
  }
Return (Ii)
End
```

۳-۱-۱-یافتن مؤلفهٔ‌غیرگوسی معنادار

برای ایجاد ساختار پیشنهادی، در هر تکرار، ماتریس مربوط به برگ انتخاب‌شده (از برگ‌های درخت باینری تشکیل‌شده تا تکرار مربوطه) در مرحلهٔ افرازبندی، به دو زیرخوشه تجزیه می‌شود. بدین منظور، داده‌های ماتریس مربوط به برگ انتخاب‌شده روی یک مؤلفه از داده‌ها افکنش شده و سپس در مرحلهٔ افرازبندی از نقطه‌ای روی داده‌های افکنش‌شده به دو زیرخوشه تجزیه می‌شوند.

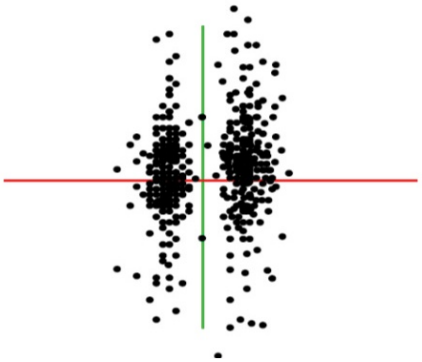
افکنش داده‌ها روی یک مؤلفه از داده‌ها ایده جدیدی نیست و در قبل در الگوریتم خوشه‌بندی سلسله‌مراتبی تقسیم‌کننده PDDP، برای خوشه‌بندی داده‌های با ابعاد بالا استفاده شده است. درPDDP، داده‌ها روی اولین مؤلفه اصلی[1] افکنش‌شده و بعد با ابرصفحهٔ گذرنده از مرکزثقل داده‌ها و عمودبر اولین مؤلفهٔ اصلی داده‌ها به دو زیرخوشه تجزیه می‌شوند. این عملیات به‌صورت بازگشتی تکرار می‌شود تا تعداد خوشه موردنظر حاصل شود (Boley, 1998). درحقیقت درPDDP داده‌ها با استفاده از تبدیل خطی PCA[2] (Jackson, 1991, Jolliffe, 2002) روی اولین مؤلفهٔ اصلی افکنش می‌شوند. اولین مؤلفهٔ اصلی، مؤلفه‌ای از داده‌هاست که امتداد آن منطبق با بیشترین پراکندگی قابل مشاهده در توزیع داده‌هاست. افکنش داده‌ها روی اولین

[1] First principal component
[2] Principal component analysis (PCA)

مؤلّفه اصلی منجربه تولید خوشه‌های متراکم می‌شود؛ اما درصورتی‌که توزیع داده‌ها به‌گونه‌ای باشد که خوشه‌های طبیعی آن در راستای اولین مؤلّفه اصلی ازهم جدا نباشند، اولین مؤلّفه اصلی نمی‌تواند حاوی اطلاعات مفیدی برای تجزیه داده‌ها باشد.

در این پژوهش از یک روش پیگرد افکنش[1] (الگوریتم FastICA) با هدف پیداکردن یک مؤلّفه غیرگوسی معنادار از داده‌ها استفاده شده است.افکنش داده‌ها روی این مؤلّفه غیرگوسی معنادار تقریب آشفتگی منفی را بیشینه می‌کند و درنتیجه افکنش داده‌ها روی آن برای تجزیه داده‌ها به دو زیرخوشه مفید خواهد بود. نمونه‌ای از مطلوب‌بودن مؤلّفه‌هایی با بیشترین خاصیت غیرگوسی (مؤلّفه حاصل از به‌کارگیری روش‌های پیگرد افکنش) نسبت به مؤلّفه‌های اصلی، ازنظر توانایی در نمایش توزیع معنادار داده‌ها در شکل (۳-۲) نشان داده شده است.

همان‌طور که در شکل(۳-۲) مشاهده می‌شود، اولین مؤلّفه اصلی که با رنگ سبز (در جهت عمودی) نشان داده شده است، ساختارخوشه‌ای داده‌ها را نشان نمی‌دهد؛ اما مؤلّفه غیرگوسی حاصل از به‌کارگیری روش‌های پیگرد افکنش که با رنگ قرمز (در جهت افقی) نشان داده شده است، جداسازی بین خوشه‌ها را نشان می‌دهد و بنابراین ساختار داده‌ها را مشخص می‌کند. گوسین‌بودن مؤلّفه‌ها سادگی محاسباتی و کاهش میزان نیاز به اطلاعات اضافی را دربردارد.



(شکل(۳-۲)): مطلوب‌بودن مؤلّفه غیرگوسین نسبت به اولین مؤلّفه اصلی ازنظر توانایی در نمایش توزیع معنادار داده‌ها

ایدهٔ اصلی روش‌های پیگرد افکنش، استخراج مؤلّفه‌های معنادار توسط بهینه‌سازی محلی تابع شاخص افکنش[2] (تابع هدف[3]) موردنظر است. تحقیقات در این زمینه نشان داده است که توزیع گوسین[4] کمترین معنا را در توزیع داده‌ها نشان می‌دهد و جهت‌های مطلوب معنادار، جهت‌هایی هستند که از توزیع گوسین فاصله دارند.

بهینه‌سازی محلی تابع هدف برای به‌دست‌آورن مؤلّفه‌های غیرگوسی، به‌طور دقیق در تحلیل مؤلّفه‌های مستقل[5]، که یکی از روش‌های پیگرد افکنش محسوب می‌شود نیز انجام می‌شود (Hyvarinen and Oja, 2000). در (Hyvarinen and Oja, 1997) با استفاده ازتبدیل قوانین یادگیری شبکه عصبی[6] به تکرار ممیز ثابت[7]، الگوریتمی به نام FastICA ارائه شده که در اصل برای پیداکردن مؤلّفه‌های مستقل پیشنهاد شده، اما درحقیقت یک روش پیگرد افکنش است که می‌توان از آن برای پیداکردن مؤلّفه‌های غیرگوسی استفاده کرد. این الگوریتم هر مؤلّفه را تک‌به‌تک پیدا می‌کند، بنابراین می‌توان از آن برای پیداکردن تنها یک مؤلّفه استفاده کرد. تابع هدف در این الگوریتم می‌تواند هریک از توابع هدفی که در تحقیقات برای به‌دست‌آوردن مؤلّفه‌های پیگرد افکنش استفاده شده، باشد. در این پژوهش از آن‌جایی‌که هدف به‌دست‌آوردن مؤلّفه‌ای است که بتواند ساختار خوشه‌ای داده‌ها را نشان دهد، از نسخه‌ای از الگوریتم FastICA که در آن تقریب آشفتگی منفی به‌عنوان تابع هدف درنظرگرفته می‌شود، استفاده شده است (Hyvarinen and Oja, 2000).

از آن‌جایی‌که متغیرهای گوسین بیشترین آشفتگی[8] را درمیان متغیرهای با واریانس یکسان دارند، بنابراین آشفتگی می‌تواند یک معیار اندازه‌گیری عدم‌گوسین‌بودن قرارداده شود. درحقیقت، این امر نشان می‌دهد که توزیع گوسین کمترین ساخت‌یافتگی را درمیان انواع توزیع داده‌ها دارد. در تحقیقات برای به‌دست‌آوردن معیار مناسبی برای اندازه‌گیری میزان عدم‌گوسین از آشفتگی منفی استفاده می‌شود. آشفتگی منفی J در رابطهٔ زیرتعریف می‌شود (Hyvarinen and Oja, 2000).

$$J(y) = H(y_{gauss}) - H(y) \qquad (۳)$$

در رابطهٔ (۳)، y_{gauss} متغیرگوسین با واریانس یکسان y و H(y) برابر آشفتگی متغیرگسسته y است. از آن‌جاکه

[1] Projection pursuit
[2] Projection index
[3] Objective function
[4] Gaussiandistribution
[5] Independent component analysis(ICA)
[6] Neural network
[7] Fixed point iteration
[8] Entropy

برآورد آشفتگی منفی دشوار است، در مرجع (Hyvarinen and Oja, 2000) درعمل از تقریب آشفتگی منفی به‌عنوان تابع هدف استفاده شده است. تقریب آشفتگی منفی در رابطهٔ زیرتعریف شده است.

$$J(y) \cong [\mathrm{E}\{\mathrm{G(y)}\} - E\{G'(V)\}]^2 \qquad (۴)$$

در رابطهٔ (۴)، تابع E امیدریاضی و V متغیرگوسی با میانگین صفر و واریانس یک است. تابع G به‌صورت رابطه (۵) تعریف شده و G' برابر مشتق تابع G می‌باشد. $1 \le c \le 2$ ثابت است که درعمل برابر یک درنظرگرفته می‌شود.

$$G(u) = \tanh(cu) \qquad (۵)$$

در این پژوهش ازالگوریتم FastICA برای به‌دست‌آوردن مؤلّفه غیرگوسی معنادار از داده‌ها استفاده شده است. این الگوریتم به‌صورتی تنظیم شده که تنها یک مؤلّفهٔ غیرگوسی را به‌دست می‌آورد. در الگوریتم FastICA اصلی از یک بردار وزن تصادفی[1] برای آغاز الگوریتم استفاده شده است، اما از آن‌جاکه در این پژوهش به‌دنبال تنها یک مؤلّفه غیرگوسی می‌باشیم و می‌خواهیم آن مؤلّفه بر جهت پراهمیتی از داده‌ها با تراکم داده‌ای بالا و توزیع با خاصیت عدم‌گوسین بالا قرارگیرد، از اولین مؤلّفه اصلی، که در پیش‌پردازش الگوریتم FastICA استفاده می‌شود، به‌عنوان بردار وزن اولیه استفاده می‌کنیم.

۳-۱-۲-افکنش

هدف از مرحلهٔ افکنش، افکنش داده‌های ماتریس $(M_i)_j$ روی مؤلّفه $(a_i)_j$ به‌دست‌آمده در مرحله قبل است. هدف از افکنش داده‌ها روی چنین مؤلّفه‌ای، تبدیل داده‌های چندبعدی با استفاده از افکنش آنها روی یک مؤلّفه به داده‌های تک‌بعدی است؛ سپس داده‌ها از محلی بر داده‌های افکنش‌شده و توسط ابرصفحه‌ای عمود بر آن، به دو زیرخوشه تجزیه می‌شوند. افکنش داده‌ها روی مؤلّفه $(a_i)_j$ طبق رابطهٔ زیر صورت می‌گیرد.

$$\mathrm{F_j(x_b)} = (\mathrm{a_i})_j^{\,T}(x_b) \qquad (۶)$$

در رابطهٔ (۶)، x_b بردار ویژگی b(b=1..nodesize)ام در ماتریس $(M_i)_j$، T علامت ترانهاده و $F_j(x_b)$ حاوی افکنش‌های تک‌بعدی بردارهای ویژگی چندبعدی x_b است.

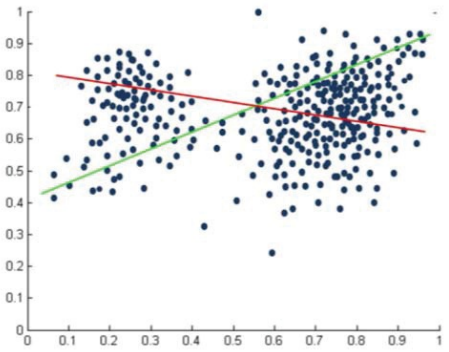
[1] Weight vector

$F_j(x_b)$ برداری است که هر مؤلّفه آن یک مقدارعددی[2] می‌باشد، که مقدار افکنش هریک از بردارهای داده x_b را نشان می‌دهد. nodesize تعداد نقاط داده‌ای درون $(M_i)_j$ است. از آن‌جاکه مؤلّفهٔ به‌دست‌آمده در زیر مرحلهٔ قبل (در این پژوهش این مؤلّفه، مؤلّفه غیرگوسی معنادار نام‌گذاری شده است) در جهتی قراردارد که می‌تواند ساختار داده‌ها را به‌خوبی نشان دهد؛ بنابراین افکنش داده‌ها روی این مؤلّفه می‌تواند حاوی اطلاعات زیر باشد:

مکان‌هایی که تراکم داده‌ها در آنها کمتر است، محل مناسبی برای تجزیه خوشه مربوطه است.

مکان‌هایی که تراکم داده‌ها در آنها بیشتر است، نزدیک به مرکزثقل زیرخوشه‌های اصلی فرضی است.

درشکل (۳-۳) مثالی از خوشه‌های نامتوازن غیرمجزا در فضای دوبعدی نشان داده شده است. در این شکل، ساختار خوشه‌ای داده‌ها که از خوشه‌های نامتوازن و غیرمجزا تشکیل شده مشاهده می‌شود. همچنین، اولین مؤلّفهٔ اصلی با رنگ سبز و مؤلّفهٔ غیرگوسی معنادار با رنگ قرمز نشان داده شده است.



(شکل(۳-۳)): مثالی از خوشه‌های نامتوازن و غیرمجزا در فضای دوبعدی

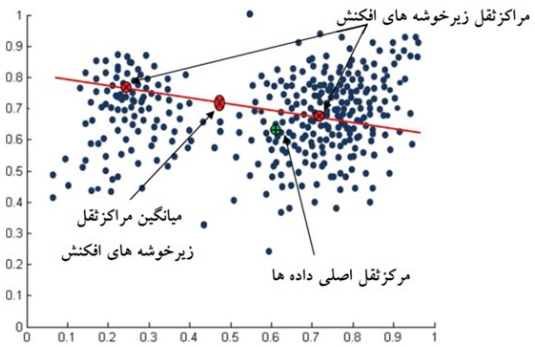
۳-۱-۳- خوشه‌بندی داده‌های افکنش‌شده

هدف از این مرحله به‌کارگیری یک الگوریتم خوشه‌بندی مانند k-means (MacQueen, 1967) روی داده‌های افکنش‌شده F_j به‌دست‌آمده در مرحله قبل، برای به‌دست‌آوردن دو زیرخوشه افکنش (IDX1,IDX2) و مراکز ثقل آنها (CP1,CP2) است.

افکنش داده‌ها روی مؤلّفه $(a_i)_j$ می‌تواند محل‌هایی با تراکم کمتر و محل‌هایی با تراکم بالا را نشان دهد. در این پژوهش برای این‌که محل‌های موردنظر حدس زده شود از

[2] Scalar

خوشه‌بندی k-means، با درنظرگرفتن k=2روی بردار F_j حاوی داده‌های تک‌بعدی افکنش، استفاده شده است. به این ترتیب که دو مرکز ثقل به‌دست‌آمده همان محل‌های متراکم داده‌ها و میانگین این مراکزثقل به‌عنوان محلی که به‌طور معمول تراکم داده‌ای کمتری دارد درنظرگرفته می‌شود. اگر ساختار خوشه‌های طبیعی داده‌ها به‌گونه‌ای باشد که خوشه‌ها متوازن و به‌طور کامل مجزا باشند، محل میانگین زیرخوشه‌های افکنش به احتمال زیاد در محل مرکز ثقل داده‌ها یا نزدیک به آن خواهد بود؛ اما اگر خوشه‌ها نامتوازن و غیرمجزا باشند، محل میانگین مراکزثقل خوشه‌های افکنش، محل کم‌تراکم‌تری را نسبت به محل مرکزثقل اصلی داده‌ها نشان می‌دهند. در شکل(۳-۴)، مرکزثقل اصلی داده‌ها و میانگین خوشه‌های افکنش به‌همراه مؤلّفهٔ غیرگوسی معنادار نشان داده شده است. همان‌طورکه مشاهده می‌شود، محل میانگین مراکزثقل خوشه‌های افکنش، که روی مؤلّفه غیرگوسی معنادار قراردارد، محل کم‌تراکم‌تری را نسبت به مرکزثقل اصلی داده‌ها نشان می‌دهد.



(شکل(۳-۴)): مقایسه مرکز ثقل اصلی داده‌ها ومیانگین مراکز ثقل خوشه‌های افکنش در مثال شکل (۳-۳)

۳-۲-افرازبندی

در مرحلهٔ افرازبندی، که شبه‌کد آن در ادامه آمده است، اطلاعات I_i به‌دست‌آمده از مرحله پیش‌افرازبندی و مجموعه M_i دریافت می‌شود. M_i حاوی برگ‌های تشکیل‌شده تا تکرار iاست. در مرحله افرازبندی ابتدا با استفاده از اطلاعاتی که در مرحله پیش‌افرازبندی به‌دست‌آمده یک معیار انتخاب برای گزینش یکی از برگ‌های تولیدشده تا تکرار i ارائه، سپس یکی از برگ‌ها براساس معیار ارائه‌شده انتخاب و خوشه مربوط به آن براساس اطلاعات به‌دست‌آمده از افکنش داده‌ها به دو زیرخوشه تجزیه می‌شود. مرحله افرازبندی شامل دو مرحله است که در شبه‌کد زیر به‌ترتیب با توابع Selection()و Split() نشان داده شده‌اند:

انتخاب خوشه: هدف از این مرحله، انتخاب یکی از برگ‌ها است که داده‌های ماتریس مربوط به آن توزیع ساخت‌یافته‌تری را دارد.

تجزیه: هدف این مرحله، تجزیهٔ خوشهٔ انتخاب‌شده به دو زیرخوشه است.

```
Partitioning (Mi ,Ii):
Begin
(Mi)s = Selection (Mi, Ii)
{(MRi)s, (MLi )s} = Split ((Mi)s, Ii)
Returns ((MRi)s, (MLi)s)
End
```

۳-۲-۱-انتخاب خوشه

در این مرحله، خوشه‌ای که باید در آن تکرار به دو زیرخوشه تجزیه شود، از میان برگ‌هایی که تا آن تکرار ایجاد شده‌اند، انتخاب می‌شود. در مرحلهٔ انتخاب خوشه، براساس اطلاعات مجموعه‌های M_i و I_i، خوشه مربوط به یکی از برگ‌ها انتخاب و در مرحلهٔ تجزیه به دو زیرخوشهٔ راست و چپ تجزیه می‌شود. در این پژوهش از اطلاعات حاصل از افکنش داده‌های مربوط به زیرخوشه‌ها در مرحله پیش‌افرازبندی استفاده شده و معیاری برای انتخاب برگی که حاوی داده‌هایی با بیشترین ساخت‌یافتگی باشد ارائه شده است.

انتخاب خوشه بعدی در هر تکرار یکی از چالش‌های عمده روش‌های سلسله‌مراتبی تقسیم‌کننده به‌حساب می‌آید؛ زیرا در این مرحله هدف پیداکردن خوشه‌ای است که بیشترین ساخت‌یافتگی را داشته باشد. دراکثر تحقیقات پیشین مانند PDDP، خوشه‌ای که بیشترین پراکندگی از مرکزثقل را داشته باشد به‌عنوان خوشهٔ موردتجزیه در هر تکرار انتخاب می‌شد. رابطه(۷)، پراکندگی[1] از مرکزثقل را نشان می‌دهد.

$$Scattvalue = \frac{1}{N}\sum_{i=1}^{N}\left\|x_i - w\right\|^2 \qquad (۷)$$

در این رابطه، Scattvalue مقدار پراکندگی از مرکزثقل، Nتعداد نقاط داده‌ای خوشه مربوطه، wمرکزثقل خوشه و $\|\ \|$ نرم اقلیدسی است.

اما همان‌طورکه در مرجع (Savaresi et al., 2002) مطرح شده، پراکندگی از مرکزثقل خوشه نمی‌تواند معیار خوبی برای نشان دادن ساختار خوشه‌بندی‌شدهٔ داده‌ها باشد.

[1] Scatter

بنابراین، در این پژوهش از اطلاعات $(I_i)_j$ به‌دست‌آمده از افکنش داده‌های مربوط به هر برگ در تکرار iام، روی مؤلّفه غیرگوسی معنادار و زیرخوشه‌ها ومراکزثقل حاصل از به‌کارگیری خوشه‌بندی 2-means روی افکنش‌ها استفاده شده و معیاری برای انتخاب خوشه (تعریف‌شده در رابطه(۸)) ارائه شده است. هرچه مقدار به‌دست‌آمده از معیار ارائه‌شده برای خوشه‌ای بزرگ‌تر باشد، ساختار داده‌های مربوط به آن خوشه ساخت‌یافته‌تراست. بنابراین در تکرارiام روش پیشنهادی، در زیرمرحلهٔ انتخاب خوشه، خوشه‌ای که این معیار را بیشینه نماید از میان برگ‌های موجود درمجموعه M_i انتخاب می‌شود.

$$selvalue = \left| CP1 - CP2 \right| \Big/ \max_{c=1,2}(diameter(IDX_c)) \qquad (۸)$$

در رابطهٔ (۸)selvalue مقدار معیار انتخاب، CP1,CP2 مراکزثقل زیرخوشه‌های افکنش مربوط به برگ مورد بررسی و$diameter(IDX_C)$، قطر خوشه افکنش‌cام (c=1,2) را نشان می‌دهد. قطر یک خوشهٔافکنش از رابطه زیر به‌دست‌می‌آید:

$$diameter(IDX) = \max(F_p) - \min(F_p) \qquad (۹)$$

در رابطهٔ (۹)، IDX خوشه افکنش و F_p مقدار داده افکنشpام p=1..psize را در زیرخوشه افکنش نشان می‌دهد، طوری‌که psize برابر تعداد افکنش‌هاست.

ایده معیار انتخاب ارائه‌شده این است که ما اعتقاد داریم اگر زیرخوشه‌های افکنش که از خوشه‌بندی داده‌های افکنش‌شده روی مؤلّفه غیرگوسی معنادار به‌دست‌می‌آیند استقلال بیشتری داشته باشند، یا به‌عبارتی فاصلهٔ مرکزثقل‌های آنها از هم بیشتر و قطر آنها کوچک‌تر باشد، زیرخوشه‌های طبیعی در ساختار داده نیز استقلال بیشتری دارند؛ بنابراین خوشه موردنظر ساختار خوشه‌بندی‌شده‌تری را نشان می‌دهد. شکل(۳-۵) مثالی را در فضای دوبعدی نشان می‌دهد که در آن نقاط داده‌ای، افکنش‌ها و مراکز ثقل زیرخوشه‌های افکنش نشان داده شده است.

۳-۲-۲-تجزیه

در مرحلهٔ تجزیه، ماتریس مربوط به برگ انتخاب‌شده در تکرار iبا $((M_i)_s)$ به همراه مؤلّفه غیرگوسی معنادار $((a_i)_s)$ و همچنین بردار حاوی داده‌های افکنش‌شده برگ انتخاب‌شده (F_s) گرفته می‌شود و زیرخوشهٔ مربوط به برگ انتخاب‌شده از محل میانگین مراکزثقل زیرخوشه‌های افکنش، توسط ابرصفحه‌ای عمودبر $(a_i)_s$ به دو زیرخوشه راست و چپ تجزیه می‌شود و $(a_i)_s$برای هر دو زیرخوشه راست و چپ ذخیره می‌شود.

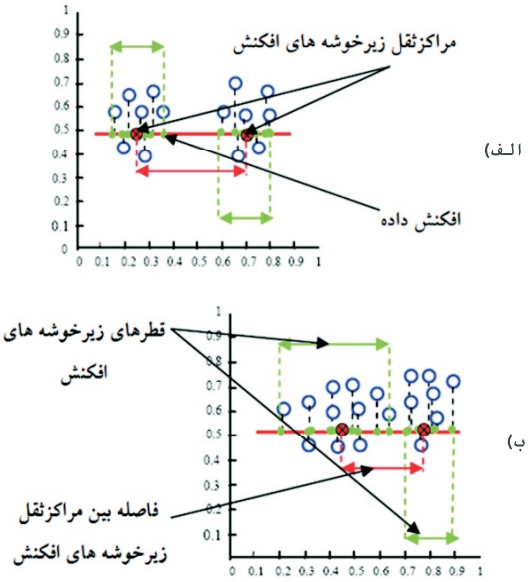
دلیل تجزیهٔ خوشه از محل میانگین، در بخش ۳-۱-۳ بیان شد. شکل(۳-۶) ابرصفحهٔ جداکننده[1] را درمورد شکل(۳-۳) نشان می‌دهد. درمورد داده‌هایی که توزیع داده‌ها در آنها به‌گونه‌ای است که از خوشه‌های غیرمتوازن و غیرمجزا تشکیل شده‌اند، انتخاب مرکز ثقل داده‌ها، به‌عنوان محلی که ابرصفحهٔ جداکننده از آن بگذرد، انتخاب جالبی نیست؛ زیرا خوشه‌های طبیعی را قطع می‌کند (بخش الف) شکل ۳-۶). اما میانگین مراکز ثقل زیرخوشه‌های افکنش، محل مناسب‌تری است. تجزیه داده‌ها توسط ابرصفحه عمودبر مؤلّفه غیرگوسی معنادار و گذرنده از میانگین مراکز ثقل زیرخوشه‌های افکنش، داده‌ها را به دو زیرخوشهٔ فشرده تجزیه می‌کند که درنهایت پس از مرزبندی، مستطیل‌های کمینه فشرده‌تری را منجر می‌شوند (بخش ب) شکل ۳-۶).

تجزیهٔ خوشه توسط ابرصفحهٔ جداکننده عمودبر مؤلّفه غیرگوسی معنادار و گذرنده از میانگین مرکز ثقل زیرخوشه‌های افکنش، نقاط داده‌ای درون خوشه انتخاب‌شده برای تجزیه، که افکنش آنها در راست محل میانگین مرکز ثقل‌های افکنش باشد به زیرخوشهٔ راست نسبت داده شده و در ماتریس $(MR_i)_s$ قرارداده می‌شوند. نقاط داده‌ای که افکنش متناظر با آنها در چپ قراردارد به زیرخوشه چپ نسبت داده شده و در $(ML_i)_s$ قرار داده می‌شوند. درحقیقت، برای تمام x_bها، رابطه (۱۰) اعمال و درصورتی‌که مقدار به‌دست‌آمده مثبت باشد، x_b در $(MR_i)_s$ و درغیر این‌صورت در $(ML_i)_s$ قرارداده می‌شود.

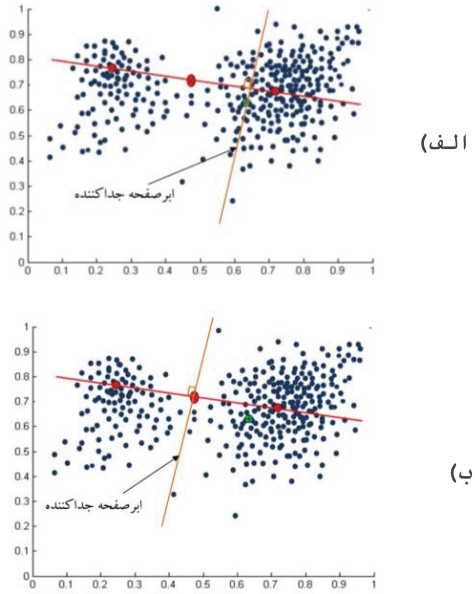
$$F_s(x_b) - (a_i)_s^{\,T} c_{mean} \qquad (۱۰)$$

در رابطهٔ (۱۰)، x_b بردار داده bام، $F_s(x_b)$ افکنش داده bام در خوشه انتخاب‌شده و c_{mean} میانگین مراکز ثقل زیرخوشه‌های افکنش است.

[1] Separating hyper-plane



(شکل (۳-۵)): مثالی از داده‌های افکنش‌شده روی مؤلّفه غیرگوسی معنادار در فضای دوبُعدی الف) خوشه‌ای که در آن داده‌ها دارای توزیع غیرساخت‌یافته می‌باشند، مقدار selvalue=2.8 ب) خوشه‌ای که در آن داده‌ها دارای توزیع ساخت‌یافته می‌باشند، مقدار selvalue=1.5



(شکل(۳-۶)): تجزیه توسط ابرصفحه جداکننده. الف) ابرصفحه جداکننده عمودبر مؤلّفه غیرگوسی معنادار و گذرنده از مرکزثقل اصلی داده‌ها. ب) ابرصفحه جداکننده عمودبر مؤلّفه غیرگوسی معنادار و گذرنده از میانگین مرکزثقل زیرخوشه‌های افکنش

۳-۳- مرزبندی

همان‌طورکه در شکل(۳-۱) فرآیند ایجاد ساختار شاخص‌گذاری چندبعدی پیشنهادی نشان داده شد، در هر تکرار، بعداز مرحله افرازبندی، زیرخوشه‌های راست و چپ تولیدشده در زیرمرحله افرازبندی، همراه $(a_i)_s$ به مرحلهٔ مرزبندی زیرخوشه‌ها فرستاده می‌شوند تا در این مرحله، توسط مستطیل‌های مرزبندی کمینه گسترده‌شده در امتداد $(a_i)_s$، مرزبندی شوند. شبه‌کد مربوط به مرزبندی در زیر آمده است.

```
Bounding ((MRi)s, (MLi )s, (ai)s)
Begin
    {(ai)s,(NRi)s, (NLi )s}=Change-reference mark ((MRi)s,
    (MLi )s, (ai)s)
    {(BRi)s, (BLi)s} =Minimum bounding rectangle
    construction ((NRi)s, (NLi )s)
    Returns((BRi)s,(BLi)s ,(MRi)s, (MLi )s)
End
```

مرحله مرزبندی شامل دو مرحله است که در شبه‌کد به‌ترتیب با توابع Change-reference mark() و Minimum bounding rectangle construction() مشخص می‌شوند:

تغییر مبدأ مختصات: هدف از این مرحله، تغییر مبدأ مختصات درجهت $(a_i)_s$، برای جلوگیری از هم‌پوشانی مستطیل‌های مرزبندی کمینه است.

تشکیل مستطیل‌های مرزبندی کمینه: هدف مرزبندی زیرخوشه‌های راست و چپ در مبدأ مختصات جدید است.

۳-۳-۱- تغییر مبدأ مختصات

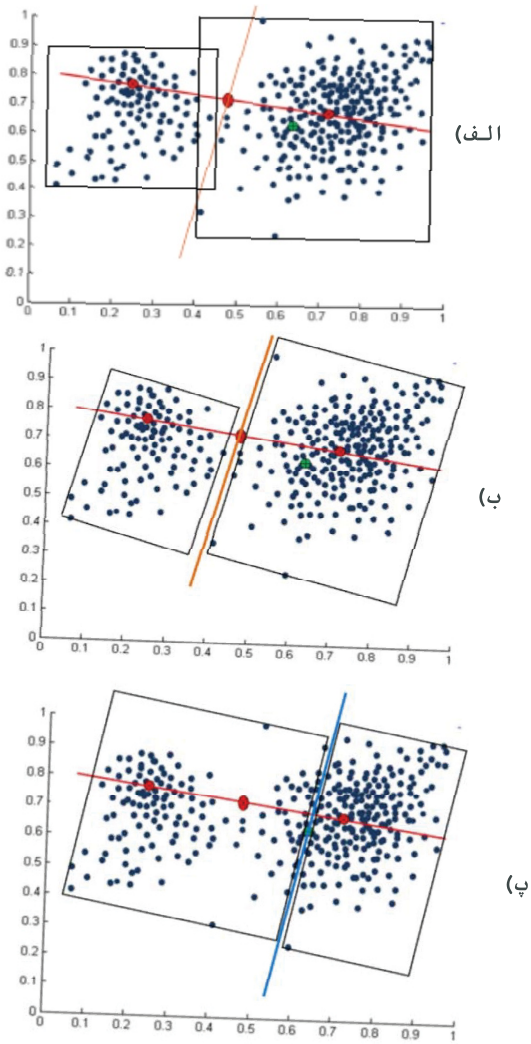
همان‌طورکه درقبل توضیح داده شد، برای جلوگیری از هم‌پوشانی بین گره‌ها در ساختار شاخص‌گذاری پیشنهادی، در این مرحله بعداز تولید زیرخوشه‌های راست و چپ در هر تکرار، ماتریس داده‌های مربوط به هریک از زیرخوشه‌های راست و چپ در مختصات جدید، که در امتداد با $(a_i)_s$ قرارداده شده است، به‌دست می‌آیند.

مسأله تغییر جهت مبدأ مختصات برای جلوگیری از هم‌پوشانی گره‌ها در ساختارهای شاخص‌گذاری ایده جدیدی نیست. پیش‌تر درNOHIS-tree برای جلوگیری از هم‌پوشانی مستطیل‌های مرزبندی کمینه، از تغییر جهت مبدأ مختصات و سپس تشکیل مستطیل‌های مرزبندی کمینه در آن جهت استفاده شده است(Taileb et al., 2007).

۳-۳-۲- تشکیل مستطیل‌های مرزبندی کمینه

در این مرحله، زیرخوشه‌های راست و چپ در مختصات جدید ($(NR_i)_s$ و $(NL_i)_s$) گرفته می‌شوند تا توسط مستطیل‌های مرزبندی کمینه (MBRs) مرزبندی شوند. از آن‌جاکه $(NR_i)_s$ و $(NL_i)_s$ شامل ماتریس داده‌های مربوط به زیرخوشه‌های راست و چپ در مختصات جدید می‌باشند، مستطیل‌های مرزبندی کمینه درجهت مؤلّفه غیرگوسی معنادار مربوط به گره والد، یا به‌عبارتی مؤلّفه غیرگوسی معنادار خوشهٔ انتخاب‌شده در تکرار i ($(a_i)_s$) تشکیل می‌شوند. سپس اطلاعات به‌دست‌آمده مربوط به مستطیل‌های مرزبندی کمینه زیرخوشه‌های راست و چپ ($(BR_i)_s$,$(BL_i)_s$) ذخیره می‌شود. برای مرزبندی روش استفاده شده در (Taileb et al., 2007) به‌کارگرفته شده است.

درشکل(۳-۷) نشان داده شده که چگونه تغییر مبدأ مختصات و تشکیل مستطیل‌های مرزبندی کمینه در راستای مبدأ جدید از هم‌پوشانی بین گره‌ها جلوگیری به‌عمل می‌آورد. همان‌طورکه در شکل(۳-۷)-ب) مشاهده می‌شود، تشکیل مستطیل‌های مرزبندی کمینه در راستای مبدأ جدید از هم‌پوشانی گره‌ها جلوگیری می‌کند. درشکل (۳-۷)-الف) و ب) تجزیه داده‌ها توسط ابرصفحه عمودبر مؤلّفه غیرگوسی معنادار و گذرنده از میانگین مراکز ثقل زیرخوشه‌های افکنش که با رنگ نارنجی نشان داده شده، صورت‌گرفته است. در شکل(۳-۷)-پ) تجزیه داده‌ها توسط ابرصفحه عمودبر مؤلّفه غیرگوسی معنادار و گذرنده از مرکزثقل اصلی داده‌ها که با رنگ آبی نشان داده شده، صورت‌گرفته است. با این‌که هم‌چنان عدم هم‌پوشانی بین مستطیل‌های مرزبندی کمینه در راستای $(a_i)_s$ تضمین شده، اما از آن‌جاکه تجزیه داده‌ها در این حالت از محل مرکزثقل اصلی داده‌ها صورت‌گرفته است، تجزیه داده‌ها منجربه خوشه‌های مجزا نمی‌شود و بنابراین، در مجموع، حجم مستطیل‌های مرزبندی کمینه تشکیل‌شده در شکل(۳-۷)-پ) نسبت به شکل (۳-۷)-ب) بیشتر است؛ زیرا مستطیل‌های مرزبندی شکل(۳-۷)-ب) فشرده‌تر می‌باشند.



(شکل(۳-۷)): تشکیل مستطیل‌های مرزبندی کمینه. الف) تشکیل مستطیل‌های مرزبندی کمینه در راستای مبدأ اصلی. ب) تشکیل مستطیل‌های مرزبندی کمینه در راستای مبدأ جدید. پ) تجزیه داده‌ها از محل مرکزثقل اصلی داده‌ها و تشکیل مستطیل‌های مرزبندی کمینه در راستای مبدأ جدید

۳-۴ ایجاد ساختار درختی

در این مرحله، در هر تکرار از فرآیند ایجاد ساختار شاخص‌گذاری پیشنهادی، زیرخوشه‌های راست و چپ $(MR_i)_s$, $(ML_i)_s$ به‌همراه ماتریس حاوی اطلاعات مرزبندی زیرخوشه‌ها ((BR_i) و (BL_i)) به ساختار درختی TS_i اضافه می‌شوند. شبه‌کد مربوطه در ادامه آمده است.

```
Build tree structure ((BRi)s, (BLi)s, (MRi)s, (MLi )s)
Begin
    If number of rows of ((MRi)s/(MLi )s) <Minpts
        let ((MRi)s/(MLi )s) be outlier
    Else let it be leaf
    Add (MRi)s, (MLi )s to the binary tree structureand save
    (BRi)s, (BLi)s, LNOi = LNOi+1
    Returns (TSi, LNOi)
End
```

همان‌طورکه در شبه‌کد بالا ملاحظه می‌شود، در فرآیند مرحله ساخت ساختار درختی، درصورتی‌که تعداد اعضای داده‌های هریک از آنها کوچک‌تر از مقدار تنظیم‌شده برای پارامتر Minpts باشد، آن زیرخوشه به‌عنوان داده پرت علامت‌زده می‌شود و درغیراین‌صورت به‌عنوان برگ علامت زده می‌شوند و در تکرار بعدی الگوریتم در مجموعه M_i قرارمی‌گیرد.

۴-پیاده‌سازی و آزمون

در این بخش نتایج پیاده‌سازی و آزمون روش پیشنهادی ارائه شده است.

۴-۱-پیاده‌سازی

در این بخش به جزئیات پیاده‌سازی، شامل معیارهای ارزیابی، روش آزمون، مجموعه داده‌ها و محیط پیاده‌سازی پرداخته شده است.

۴-۱-۱-معیارهای ارزیابی

در این پژوهش، دو معیار Recall و زمان پاسخ[1] به‌منظور ارزیابی روشپیشنهادی درنظرگرفته شده‌اند.کیفیت خوشه‌های نهایی که موردجستجو قرارمی‌گیرند با استفاده از معیارRecall اندازه‌گیری شده است. معیارRecall، نسبت تعداد داده‌های مرتبط با بردار پرس‌وجو را که بازیابی شده‌اند به کل داده‌های مرتبط با بردار پرس‌وجوی مطرح‌شده در پایگاه‌داده، اندازه‌گیری می‌کند، و طبق رابطهٔ زیر تعریف می‌شود (Li et al., 2002, Yu and Zhang, 2003, Mitchell, 1997).

$$\text{Recall} = \frac{|\alpha \cap \beta|}{|\alpha|} \qquad (۱۱)$$

در رابطهٔ (۱۱)، α مجموعه داده‌های مرتبط بازیابی‌شده، β مجموعه داده‌های بازیابی‌شده و | |، نشان‌دهندهٔ تعداد است.

[1] Response time

از آن‌جاکه هدف اصلی ساختارهای شاخص‌گذاری تصویر، بالابردن سرعت بازیابی تصاویر است، بازیابی داده‌ها ازطریق ساختار شاخص‌گذاری پیشنهادی در این پژوهش، با معیار زمان پاسخ اندازه‌گیری شده‌اند. این معیار، درحقیقت زمانی است که صرف جستجوی پایگاه‌داده ازطریق ساختار شاخص‌گذاری برای پیداکردن k نزدیک‌ترین داده‌های درون پایگاه‌داده نسبت به داده پرس‌وجو می‌شود، و طبق رابطهٔ (۱۲) تعریف می‌شود (Li et al., 2002, Taileb et al., 2007, Yu and Zhang, 2003).

$$R = r_1 - r_2 \qquad (۱۲)$$

در رابطهٔ (۱۲)، R زمان پاسخ، r_1 زمان خروج نتایج نهایی از بخش جستجوی شباهت و r_2 زمان ورود پرس‌وجو به بخش جستجوی شباهت است.

۴-۱-۲-روش آزمون

همه آزمون‌ها با استفاده از روش Cross-validation انجام گرفته‌اند. بدین منظور هر آزمون ده‌بار انجام و هر بار، بیست بردار ویژگی از پایگاه‌داده ویژگی استفاده‌شده برای آزمون موردنظر، در مجموعه آزمون قرارداده شده و برای هریک، کل پایگاه‌داده، با جستجوی ترتیبی[2] و معیار فاصلهٔ اقلیدسی، پیمایش شده و ۲۰-نزدیک‌ترین بردارهای مربوط به هر بردار موجود در مجموعه آزمون پیدا و سپس در مجموعه داده‌های مرتبط (مجموعه α) قرارداده شده و در مرحلهٔ بعد ساختارهای شاخص‌گذاری مورد ارزیابی با تصاویر باقی‌مانده ساخته شده است. از داده‌های مجموعه آزمون به‌عنوان بردارهای پرس‌وجو و به‌منظور پرس‌وجو از پایگاه‌داده استفاده و میانگین نتایج حاصل برای بیست‌بردار پرس‌وجو به‌دست‌آمده ودرنهایت نتایج میانگین ده‌آزمون گزارششده است (Savaresi et al., 2002, Yu and Zhang, 2003).

۴-۳-۱-مجموعه داده‌ها

برای ارزیابی‌ها و مقایسه روش پیشنهادی با دیگر روش‌های مطرح در این حوزه، از مجموعه داده به‌کاررفته در آزمون‌های مطرح در (Taileb et al., 2007)، که پایگاه‌داده حاوی ۱۱۹۳۶۴۷ بردار ویژگی استخراج‌شده از تصاویر رنگی می‌باشد، استفاده شده است. بردارهای ویژگی تصاویر موجود در این پایگاه‌داده، ویژگی‌های محلی براساس نقاط مطلوب استخراج‌شده از ۴۹۹۶ تصویر مختلف رنگی می‌باشند که در

[2] Sequential-scan

تحقیقات پیشین توسط محققین، به‌منظور ارزیابی روش‌ها، مورد استفاده قرارگرفته‌اند. در آزمون‌های انجام‌شده از ۴ پایگاه‌داده حاوی ۵۰،۰۰۰ بردار ویژگی در ابعاد مختلف استفاده شده است. پایگاه‌داده اول شامل ۵۰،۰۰۰بردار ویژگی ۲۵بعدی، پایگاه‌داده دوم شامل ۵۰،۰۰۰ بردار ویژگی ۴۰بعدی، پایگاه‌داده سوم شامل ۵۰،۰۰۰بردار ویژگی ۶۰بعدی و پایگاه‌داده چهارم شامل ۵۰،۰۰۰ بردار ویژگی ۸۰بعدی است.

۴-۱-۴-محیط پیاده‌سازی

برای پیاده‌سازی و آزمون روش پیشنهادی و سایر روش‌های مورد مقایسه، از محیط MATLAB 7.8.0.347 در یک دستگاه رایانه شخصی با امکانات CPU.4GHZ و RAM4GB استفاده شده است.

۴-۲-آزمون

دراین بخش، جزئیات مربوط به آزمون روش پیشنهادی مطرح شده است.

۴-۲-۱-تنظیم پارامترهای پیش‌نیاز

در این بخش، به‌منظور تنظیم پارامترهای پیش‌نیاز طرح شاخص‌گذاری پیشنهادی، آزمون‌هایی انجام شده و نتایج آن مورد تحلیل قرارگرفته است.

اگر به شبه‌کدهای بخش۳رجوع شود، مشاهده می‌شود که در طرح پیشنهادی دو پارامتر پیش‌نیاز k وMinpts درنظرگرفته شده‌اند. پارامترk، تعداد خوشه‌های نهایی (برگ‌ها و داده‌های پرت نهایی) است. پارامترk درحقیقت تعداد برگ‌ها به‌علاوه تعداد داده‌های پرت نهایی، پس از ایجاد ساختار شاخص‌گذاری چندبعدی پیشنهادی NO-NGP-tree را مشخص می‌کند. مقدار حداقل درنظرگرفته شده برای تعداد داده‌های برگ با پارامتر Minpts مشخص می‌شود. در آزمون‌های انجام‌شده، پارامترMinpts به‌صورت درصد متوسط تعداد اعضای داده خوشه‌های نهایی درنظرگرفته شده است. به‌عنوان مثال، مقدار ۱۵، به‌معنای ۱۵درصد متوسط تعداد اعضای داده خوشه‌های نهایی است. این مقدار متوسط طبق رابطه (۱۳) به‌دست می‌آید.

(۱۳)

k/تعداد کل داده‌ها=متوسط تعداد اعضای داده خوشه‌های نهایی

برای تنظیم دو پارامتر پیش‌نیاز k وMinpts، آزمون‌هایی انجام، و متوسط زمان پاسخ برحسب ثانیه گزارش شده است. این آزمون‌ها با تغییر مقادیر پارامتر Minpts و مقدارثابت k روی هر چهار پایگاه‌داده ۲۵،۴۰،۶۰ و۸۰بعدی انجام و این آزمون برای مقادیر مختلف k تکرارشده است. جدول(۴-۱) نتایج متوسط زمان پاسخ برحسب ثانیه (S) به‌دست‌آمده از آزمون روش پیشنهادی را، با درنظرگرفتن مقادیر مختلف Minptsو k، روی چهار پایگاه‌داده ۲۵،۴۰،۶۰و۸۰بعدی نشان می‌دهد.

(جدول(۴-۱)): متوسط زمان پاسخ برای مقادیر مختلف Minptsو k روی پایگاه‌داده‌ها با ابعاد مختلف

Minpts	۵	۱۵	۲۵	۳۵	۴۵	۶۵
متوسط زمان پاسخ(S) روی پایگاه داده ها با ابعاد مختلف،k=200						
پایگاه داده ۲۵ بُعدی	۰.۱۴۶	۰.۱۲۵	۰.۱۰۴	۰.۱۱۲	۰.۱۲۵	۰.۱۴۶
پایگاه داده ۴۰ بُعدی	۰.۴۶۲	۰.۳۶۸	۰.۳۵۲	۰.۳۷۵	۰.۳۸۶	۰.۴۲۷
پایگاه داده ۶۰ بُعدی	۰.۶۶۳	۰.۵۷۸	۰.۵۳۵	۰.۵۵۸	۰.۵۹۳	۰.۶۱۳
پایگاه داده ۸۰ بُعدی	۰.۷۷	۰.۶۸۴	۰.۶۶۴	۰.۶۷۹	۰.۷۰۳	۰.۷۳۲
متوسط زمان پاسخ (S) روی پایگاه داده ها با ابعاد مختلف، k=600						
پایگاه داده ۲۵ بُعدی	۰.۰۸۴	۰.۰۶۹	۰.۰۶۲	۰.۰۶۳	۰.۰۶۸	۰.۰۷۴
پایگاه داده ۴۰ بُعدی	۰.۳۶	۰.۲۴۳	۰.۲۲۸	۰.۲۳	۰.۲۵۷	۰.۳۱۶
پایگاه داده ۶۰ بُعدی	۰.۴۷۵	۰.۳۶	۰.۳۴۵	۰.۳۷۲	۰.۴۱	۰.۴۳۸
پایگاه داده ۸۰ بُعدی	۰.۶۳۲	۰.۵۴	۰.۵۱۳	۰.۵۲۱	۰.۵۶۱	۰.۶۰۷
متوسط زمان پاسخ (S) روی پایگاه داده ها با ابعاد مختلف، k=800						
پایگاه داده ۲۵ بُعدی	۰.۱۷۳	۰.۱۰۹	۰.۰۹۸	۰.۱۰۳	۰.۱۱	۰.۱۲۴
پایگاه داده ۴۰ بُعدی	۰.۴۴۷	۰.۳۴۳	۰.۳۰۱	۰.۳۲۱	۰.۳۵	۰.۴۰۱
پایگاه داده ۶۰ بُعدی	۰.۵۲۴	۰.۴۰۶	۰.۳۶	۰.۳۸۷	۰.۴۵۱	۰.۴۶
پایگاه داده ۸۰ بُعدی	۰.۶۵	۰.۵۶۷	۰.۵۲	۰.۵۳۲	۰.۵۸	۰.۶۱۷
متوسط زمان پاسخ (S) روی پایگاه داده ها با ابعاد مختلف، k=1000						
پایگاه داده ۲۵ بُعدی	۰.۲۴۶	۰.۱۴۵	۰.۱۴	۰.۱۴۸	۰.۱۵۵	۰.۲۰۱
پایگاه داده ۴۰ بُعدی	۰.۶۱	۰.۵۱۸	۰.۴۵۳	۰.۴۹	۰.۵۲۶	۰.۵۶۳
پایگاه داده ۶۰ بُعدی	۰.۷۹۲	۰.۷۰۵	۰.۶۵۲	۰.۶۸۳	۰.۷۲۴	۰.۷۶۳
پایگاه داده ۸۰ بُعدی	۱.۰۲۶	۰.۹۲۵	۰.۸۷۳	۰.۹	۰.۹۴۶	۰.۹۷

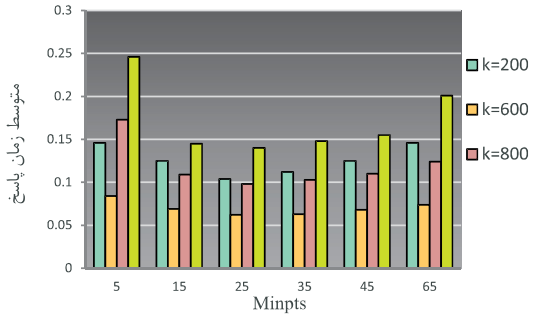
بررسی نتایج جدول (۴-۱)، نشان می‌دهد که با بالارفتن ابعاد پایگاه‌داده، متوسط زمان پاسخ برای هر چهار مقدار k افزایش یافته و با تغییر مقادیر Minpts، نتایج مختلفی حاصل شده است. بنابراین برای تحلیل دقیق‌تر نتایج، در شکل(۴-۱) از ابعاد صرف‌نظرشده و نتایج روی پایگاه‌داده ۲۵بعدی به‌صورت هیستوگرام متوسط زمان پاسخ، با تغییر مقادیر مختلفMinpts، برای مقادیر مختلفk، نشان داده شده است. در شکل(۴-۱)، مشاهده می‌شود که با تنظیمات انجام‌شده روی پارامترهای پیش‌نیاز، بهترین نتایج

زمانی حاصل می‌شود که Minpts برابر ۲۵درصد متوسط تعداد اعضای داده خوشه‌های نهایی، وk نیز برابر ششصد قرارداده شده است. همچنین وقتی Minpts روی مقادیر بزرگ (۶۵درصد) و یا مقادیر کوچک (۵درصد) تنظیم شد، متوسط زمان پاسخ بالا رفته، و هنگامی‌که Minpts روی مقادیر ۱۵ تا ۲۵ درصدتنظیم شده، نتایج بهتری حاصل شده است.

این مسأله ناشی از این حقیقت است که اگر حداقل درنظرگرفته‌شده برای تعداد اعضای خوشه‌هایی که آنها را برگ می‌نامیم درهر تکرار بسیار بزرگ باشد، از آن‌جاکه معیار انتخاب خوشهٔبعدی ارائه‌شده در روش پیشنهادی تمایل به انتخاب خوشه‌های کوچک‌تر دارد، به‌طور مداوم خوشه‌های کوچک تجزیه می‌شوند و این مسأله منجربه افزایش ارتفاع ساختار شاخص‌گذاری به‌طور نامتوازن‌تر شده و درنتیجه سرعت جستجوی پایگاه‌داده ازطریق ساختار شاخص‌گذاری پیشنهادی کاهش می‌یابد. همچنین، تنظیم مقادیر بسیار کوچک برای Minpts باعث می‌شود داده‌های پرت به‌درستی تشخیص داده نشوند و فقط خوشه‌هایی با تعداد داده‌های بسیار کم به‌عنوان داده‌های پرت علامت‌زده شوند. درنتیجه منجربه افرازبندی‌های بی‌جا و کاهش کیفیت خوشه‌های نهایی می‌شود، که درنهایت افزایش زمان پاسخ را دربر خواهد داشت. در شکل(۴-۱)، مشاهده می‌شود که زمان پاسخ برای مقادیر مختلفk و مقدار ثابت Minpts متفاوت است. این مسأله می‌تواند به این علت باشد که اگر مقدارk نزدیک به تعداد خوشه‌های طبیعی موجود در توزیع داده‌ها تنظیم شود، کیفیت خوشه‌های نهایی بهتر خواهد بود و درنتیجه از آن‌جاکه جستجو در خوشه‌های کمتری را موجب می‌شود، جستجو ازطریق ساختار شاخص‌گذاری پیشنهادی زمان کمتری خواهد برد.

۴-۲-۲- نتایج حاصل از ارزیابی کیفیت خوشه‌های نهایی جستجوشده

در این بخش نتایج حاصل از ارزیابی کیفیت خوشه‌های نهایی جستجوشده، با استفاده از معیار Recall و درمقایسه با سه روش زیر نشان داده شده است:



(شکل(۴-۱)): نتایج روی پایگاه‌داده ۲۵بعدی به‌صورت هیستوگرام متوسط زمان پاسخ با تغییر مقادیر مختلف Minpts برای مقادیر مختلفk

NGP-tree: این ساختار شاخص‌گذاری با حذف مرحله تغییر مبدأ مختصات از فرآیند ایجاد ساختار شاخص‌گذاری پیشنهادی No-NGP-tree، و با هدف مقایسه اثر عدم هم‌پوشانی گره‌ها در ساختار شاخص‌گذاری پیشنهادی ایجاد شده است.

PDDP-tree: با اضافه‌کردن MBRها به روش خوشه‌بندی سلسله‌مراتبی تقسیم‌کننده PDDP ایجاد شده است.

NOHIS-tree: ساختار شاخص‌گذاری مبتنی‌بر خوشه‌بندی سلسله‌مراتبی تقسیم‌کننده PDDP است.

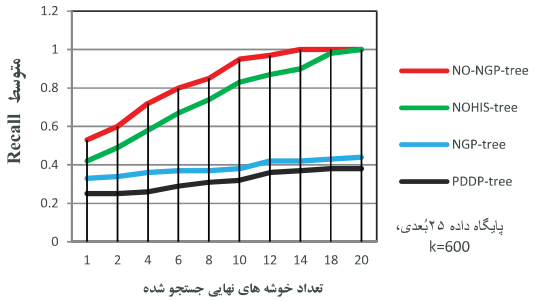
لازم به ذکر است در چهار روش بررسی‌شده، در بخش جستجوی شباهت، از الگوریتم جستجوی شباهت پیشنهادشده در (Taileb et al., 2007) استفاده شده است. آزمون‌های این بخش با تنظیم پارامتر Minpts برابر ۲۵درصد، و با درنظرگرفتن مقادیر مختلف برای پارامترk، برای روش‌هایNO-NGP-treeو NGP-tree انجام شده است. هدف از انجام این ارزیابی، با درنظرگرفتن مقادیر مختلف پارامتر k، نشان‌دادن تأثیر تنظیم این پارامتر و مقایسه روش پیشنهادی در این حالات با روش‌های دیگر است.

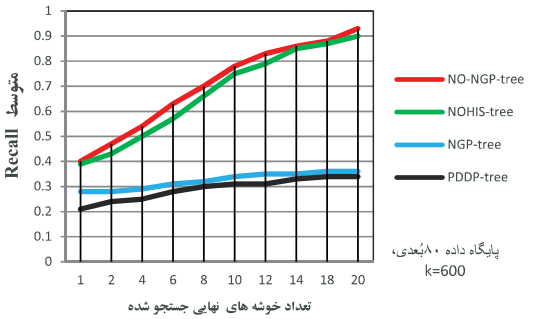
برای نشان دادن بهتر نتایج، نمودارهای متوسط Recall نسبت به تعداد خوشه‌های نهایی جستجوشده روی پایگاه‌داده‌های ۲۵ و ۸۰بعدی رسم شده است. شکل(۴-۲)، نمودارهای متوسط Recall نسبت به تعداد خوشه‌های نهایی جستجوشده روی پایگاه‌داده ۲۵ و۸۰بعدی، درحالی‌که پارامترk برابر ۶۰۰، ۸۰۰، ۱۰۰۰، و پارامتر Minpts برابر ۲۵درصد را قرارداده شده است نشان می‌دهد.

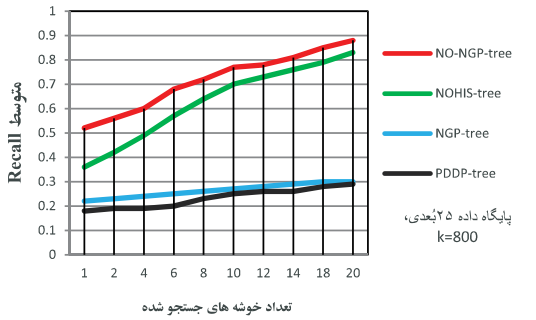
درمورد NO-NGP-tree، جستجو پس از جستجوی به‌طور متوسط چهارده خوشه نهایی متوقف می‌شود، و

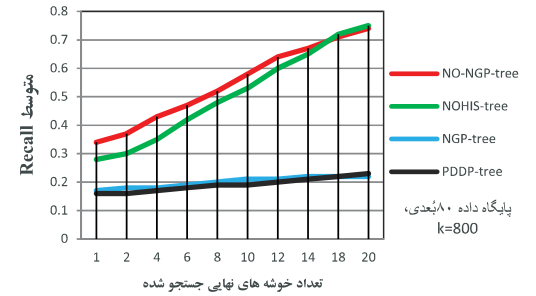
درمورد NOHIS-tree جستجو بعد از جستجوی به‌طور متوسط بیست‌خوشه نهایی متوقف می‌شود. همان‌طورکه در شکل(۴-۲) مشاهده می‌شود متوسط Recall با افزایش تعداد خوشه‌های جستجوشده افزایش می‌یابد وRecall بعداز جستجوی به‌طور تقریبی چهارده و بیست‌خوشه به‌ترتیب در روشNO-NGP-tree و NOHIS-treeیک شده است. همچنین درمورد روش‌های PDDP-treeو NGP-tree که در ساختار آنها گره‌ها هم‌پوشانی دارند، متوسطRecall به‌کندی افزایش یافته و درنهایت درNGP-tree پس از جستجوی به‌طور تقریبی ۲۵۷خوشه، و درPDDP-tree پس از جستجوی به‌طور تقریبی۲۷۳خوشه، جستجو متوقف می‌شود. دلیل افزایش کندتر متوسط Recall، درمورد دو روش که در ساختار آنها گره‌ها هم‌پوشانی دارند؛ را می‌توان عدم دقت محاسبه فاصله MINDIST (فاصله بین مستطیل مرزبندی کمینه و بردار پرس‌وجو (Taileb et al., 2007))، زمانی‌که گره‌های میانی هم‌پوشانی دارند، توجیه کرد، همچنین هر اندازه مستطیل‌های مرزبندی کمینه فشرده‌تر باشند، دقت محاسبه فاصله MINDIST بیشتر خواهد بود. همان‌طورکه در شکل(۴-۲) مشاهده می‌شود، روش NO-NGP-tree نتایج نسبتاً بهتری را نسبت به NOHIS-tree تولید کرده است. دلیل این نتایج بهتر را می‌توان باتوجه به افرازبندی‌های بهینه‌ای که درطول ساخت ساختار شاخص گذاری دارد و منجربه ایجاد مستطیل‌های مرزبندی کمینه به‌نسبه فشرده‌ای می‌شود، توجیه کرد.

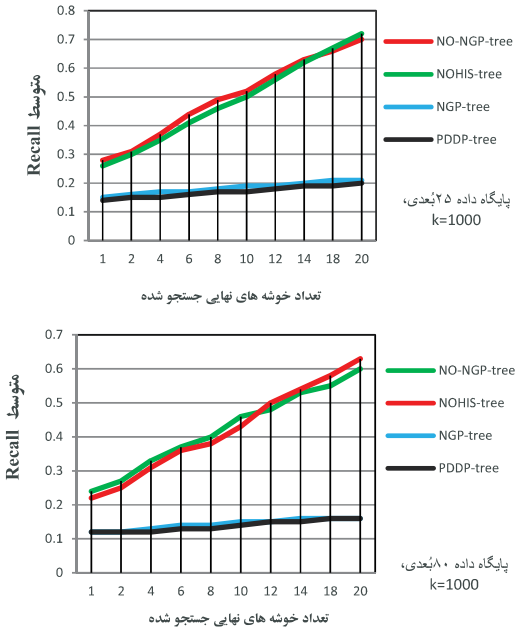
همان‌طورکه در شکل(۴-۲) مشاهده می‌شود، متوسط Recall نسبت به نتایج حاصله روی پایگاه‌داده با ابعاد ۲۵، در هر چهار روش کندتر رشد کرده است. همچنین، مشاهده می‌شود، با بالارفتن مقدار پارامتر پیش‌نیاز k، ساختار پیشنهادیNO-NGP-tree و NGP-tree افزایش متوسط Recall کمتری را نشان می‌دهند. دلیل کاهش متوسط Recall این است که هرچه مقدارk ازتعداد خوشه‌های واقعی موجود در توزیع داده‌ها فاصله می‌گیرد، کیفیت خوشه‌های نهایی به‌دست‌آمده پایین می‌آید، همچنین این کاهش کارایی در بُعد۸۰ بیشتر خود را نشان می‌دهد، به‌طوری‌که بعداز جستجوی دوازده‌خوشه، متوسط Recall برای NOHIS-tree مقدار بالاتری را نسبت به متوسط Recall برای NO-NGP-tree نشان می‌دهد. دلیل این مسأله، کاهش کارایی الگوریتم استفاده‌شده در مرحله پیش‌افرازبندی برای یافتن یک مؤلّفهٔ‌غیرگوسی معنادار از داده‌ها در پایگاه‌داده‌های با ابعاد بالاتر است.











(شکل(۴-۲)): متوسط Recall نسبت به تعداد خوشه‌های نهایی جستجوشده، Minpts=25

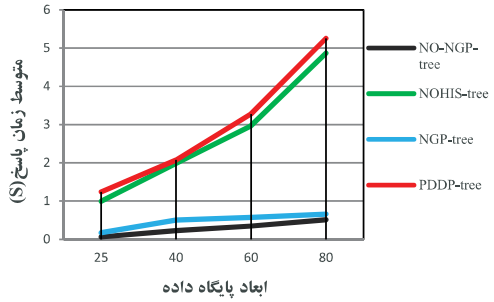
۴-۳- نتایج حاصل از ارزیابی سرعت جست‌وجو از طریق طرح شاخص‌گذاری پیشنهادی

در این بخش نتایج حاصل از ارزیابی سرعت جستجو ازطریق طرح شاخص‌گذاری پیشنهادی، با استفاده از معیار زمان پاسخ و در مقایسه با طرح‌های شاخص‌گذاری NOHIS-tree، NGP-tree،PDDP-tree و جستجوی ترتیبی نشان داده شده است.

آزمون‌های این بخش با تنظیم پارامترهای پیش‌نیاز در ساختارهای NO-NGP-treeوNGP-tree،روی بهترین حالت به‌دست‌آمده در بخش تنظیم پارامترهای موجود انجام شده است. بنابراین در تمام آزمون‌های این بخش Minpts برابر ۲۵درصد و kبرابر ۶۰۰ تنظیم شده است. همچنین در همه روش‌های بررسی‌شده، در بخش جستجوی شباهت، از الگوریتم جستجوی پیشنهادشده در (Taileb et al., 2007) استفاده شده است. جدول(۴-۲)، متوسط زمان پاسخ برحسب ثانیه را برای پایگاه‌داده ۶۰،۲۵،۴۰ و ۸۰بعدی نشان می‌دهد. شکل(۴-۳) نمودارهای متوسط زمان پاسخ را برای مقایسه طرح‌های شاخص‌گذاری NO-NGP-tree، NOHIS-tree،NGP-tree وPDDP-tree، روی پایگاه‌داده‌های ۲۵،۴۰، ۶۰و۸۰بعدی نشان می‌دهد.

(جدول(۴-۲)): متوسط زمان پاسخ روی پایگاه‌داده‌های با ابعاد مختلف

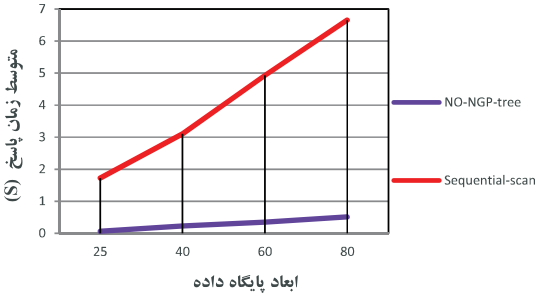
طرح شاخص‌گذاری	متوسط زمان پاسخ(S)			
	پایگاه داده ۲۵بُعدی	پایگاه داده ۴۰بُعدی	پایگاه داده ۶۰بُعدی	پایگاه داده ۸۰بُعدی
NO-NGP-tree	۰.۶۲	۰.۲۲۸	۰.۳۴۵	۰.۵۱۳
NGP-tree	۰.۹۹۲	۱.۹۸۶	۲.۸۶۹	۴.۸۷
NOHIS-tree	۰.۱۷۲	۰.۵۰۴	۰.۵۷۲	۰.۶۶۰
PDDP-tree	۱.۲۴۳	۲.۵۰۳	۳.۲۷۶	۵.۲۵۲
Sequential-scan	۱.۷۲۳	۳.۱۰۲	۴.۹۲۸	۶.۶۵۲



(شکل(۴-۳)): متوسط زمان پاسخ روی پایگاه‌داده‌های۲۵،۴۰، ۶۰و۸۰بعدی برای مقایسه طرح‌های شاخص‌گذاری NO-NGP-tree، NOHIS-tree،NGP-tree و PDDP-tree

در شکل(۴-۳)، طرح شاخص‌گذاری پیشنهادی NO-NGP-tree که با رنگ مشکی نشان داده شده است، متوسط زمان پاسخ کمتری را نسبت به NOHIS-tree، NGP-tree وPDDP-tree در همه ابعاد دارد. همان‌طورکه مشاهده می‌شود زمان پاسخ برای هر چهارروش با افزایش ابعاد افزایش یافته است. در این روش‌ها افزایش ابعاد پایگاه‌داده، منجربه افزایش حجم مستطیل‌های مرزبندی کمینه گره‌ها می‌شود. بدین ترتیب، متوسط فاصلهٔ بردار پرس‌وجو از مستطیل‌های مرزبندی کمینه افزایش و درنتیجه دقت محاسبه فاصله MINDIST در گره‌ها کاهش می‌یابد. کاهش دقت محاسبه فاصله منجربه افزایش تعداد پیمایش‌های ساختار شاخص‌گذاری، افزایش تعداد خوشه‌های نهایی جستجوشده تا پیش از اتمام جستجو، و بنابراین موجب افزایش زمان پاسخ می‌شود. همچنین در شکل (۴-۳) مشاهده می‌شود دو روش NGP-tree وPDDP-tree که در ساختارشان از هم‌پوشانی گره‌ها جلوگیری نشده است، نسبت به روش‌های NO-NGP-tree و NOHIS-tree متوسط زمان پاسخ بیشتری را دارند. دلیل بیشتر بودن زمان پاسخ این روش‌ها وجود هم‌پوشانی بین گره‌هاست که موجب افزایش پیمایش‌های ساختار شاخص‌گذاری هنگام جستجو

می‌شود. دراین پژوهش ادعا شده که دلیل کاهش زمان پاسخ در طرح شاخص‌گذاری پیشنهادی NO-NGP-tree نسبت به طرح شاخص‌گذاری NOHIS-tree، افرازبندی‌های بهینه و تولید زیرخوشه‌های متراکم‌تر و درنتیجه کاهش حجم مستطیل‌های مرزبندی کمینه در ساختار سلسله‌مراتبی ارائه‌شده NO-NGP-tree است. کاهش حجم مستطیل‌های مرزبندی کمینه موجب افزایش دقت محاسبه MINDIST، و درنتیجه باعث کاهش زمان پاسخ می‌شود. به همین دلیل زمان پاسخ برای طرح شاخص‌گذاری NGP-tree از PDDP-tree کمتر شده است. شکل(۴-۴) نمودار زمان پاسخ برای مقایسه طرح شاخص‌گذاری NO-NGP-treeو Sequential-scan، در پایگاه‌داده‌های ۲۵،۴۰، ۶۰ و ۸۰بعدی را نشان می‌دهد.



(شکل(۴-۴)): متوسط زمان پاسخ روی پایگاه‌داده‌های ۲۵،۴۰، ۶۰و۸۰بعدی برای مقایسه طرح‌های شاخص‌گذاری NO-NGP-tree و Sequential-scan

دلیل مقایسه طرح شاخص‌گذاری پیشنهادی با جستجوی ترتیبی این است که اکثر طرح‌های شاخص‌گذاری سنتی در فضاهای با ابعاد بالا کارایی خود را به‌گونه‌ای ازدست می‌دهند که گاهی جستجوی ترتیبی داده‌ها بهتر عمل می‌کند(Markov et al., 2008, Chakrabarti and Mehrotra, 1999). همان‌طورکه در شکل(۴-۴) مشاهده می‌شود، طرح شاخص‌گذاری پیشنهادی NO-NGP-tree با فاصلهٔ زیادی توانسته است زمان پاسخ را نسبت به جستجوی ترتیبی بهبود دهد.

۵-نتیجه‌گیری

در اکثر کاربردهای بازیابی تصویر، تصاویر توسط بردارهای ویژگی با ابعاد بالا مورداستفاده قرارمی‌گیرند. نتایج تحقیقات نشان داده که ساختارهای شاخص‌گذاری چندبعدی رایج با افزایش ابعاد داده‌ها کارایی خود را ازدست می‌دهند. در این پژوهش با بهره‌گیری از الگوی روش‌های خوشه‌بندی سلسله‌مراتبی تقسیم‌کننده، ساختار شاخص‌گذاری چندبعدی جهت مدیریت پایگاه‌داده‌های حاوی بردارهای ویژگی تصویر درابعاد بالا ارائه شد. در بخش چهار، پارامترهای پیش‌نیاز ساختار شاخص‌گذاری چندبعدی پیشنهادیNO-NGP-tree با مقادیر مختلف تنظیم شده، ساختار شاخص‌گذاری چندبعدی پیشنهادی موردارزیابی قرارگرفت و نتایج به‌دست‌آمده گزارش شد. همچنین، کیفیت خوشه‌های نهایی حاصل از روش خوشه‌بندی سلسله‌مراتبی تقسیم‌کنندهٔ پیشنهادی، که ساختار پیشنهادی براساس آن NO-NGP-tree شکل گرفته است، موردارزیابی قرارداده شد. درنهایت با تنظیم پارامترهای پیش‌نیاز در بهترین حالت، زمان بازیابی ازطریق ساختار شاخص‌گذاری چندبعدی پیشنهادی موردارزیابی قرارگرفته و نتایج مطلوبی حاصل شد.

برخی از فعالیت‌هایی که به‌منظور بهبود و توسعه ساختار پیشنهادی NO-NGP-tree می‌توانند در آینده درنظرگرفته شوند عبارتنداز: ۱-به‌کارگیری توابع هدف دیگر به‌جای تقریب آشفتگی منفی در فرآیند پیداکردن مؤلّفه غیرگوسی معنادار. مقایسه نتایج به‌دست‌آمده می‌تواند در فعالیت‌های آینده درنظرگرفته شود؛۲-ساختار شاخص‌گذاری چندبعدی پیشنهادی ازیک ساختار درختی باینری نامتوازن تبعیت می‌کند، بنابراین تجزیه داده‌ها به چندین زیرخوشه به‌جای تنها دو زیرخوشه در هر تکرار الگوریتم می‌تواند تاحدّ مطلوبی کارایی ساختار شاخص‌گذاری پیشنهادی را بهبود دهد؛۳-ارائهٔ روشی برای تخمین پارامتر پیش‌نیازk، با استفاده از روش‌های انتخاب مدل؛ ۴- تحلیل و ارزیابی عملکرد ساختار پیشنهادی در مجموعه داده‌هایی با بردارهای ویژگی در ابعاد بسیار بالاتر.

۶- منابع

Asbaghi sh., Keyvanpour M.R., Amiri A., "Learning-Based Approach for Semantic Image Retrieval by Using a Dynamic Semantic Network", DEXA, 23rd International Workshop on Database and Expert Systems Applications, pp. 107-111, 2008.

Bohm C., Berchtold S., Keim D. A., "Searching in high-dimensional spaces-Index structure for Improving the performance of Multimedia Databases", ACM Computing surveys, Vol. 33, No. 3, pp. 322-373, 2001.

Keyvanpour M.R. and Tavoli R., "Feature Weighting for Improving Document Image Retrieval System Performance", International Journal of Computer Science Issues, Vol 9, Issue 3, No 3, pp. 125-130, 2012.

Lew M. S., Sebe N., Djeraba C. and Jain R., "Content-based Multimedia Information Retrieval: State of the Art and Challenges", ACM Transactions on Multimedia Computing, Communications and Applications, 2006.

Li C., Chang E., Garcia-Molina H. and Wiederhold G., "Clustering for approximate similarity search in high-dimensional spaces", IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 4, pp. 792–808, 2002.

MacQueen J., "Some Methods for Classification and Analysis of Multivariate Observations", In Proceeding of 5th BerkeleySymp. Mat. Statist, Prob.1, pp. 281-297, 1967.

Markov K., Ivanova K., Mitov I. and Karastanev S., "Advance of the Access Methods", International Journal of Information Technologies and Knowledge, Vol. 2, 2008.

Mitchell T. M., "Machine Learning", McGraw-Hill, 1997.

Rui Y. and Huang T. S., "Image Retrieval: current techniques, promising, Directions and Open Issues", Journal of Visual communication and Image representation, Vol. 10, Issue. 1, pp. 39-62, 1999.

Savaresi S. M., Boley D., Bittanti S. and Gazzaniga G., "Cluster Selection in Divisive Clustering Algorithms", In Proceedings of SDM, 2002.

Srinivasa Rao Ch., Srinivas Kumar S. and Chandra Mohan B., "Content based Image Retrieval Using Exact Legendre Moments And Support Vector Machine", The international journal of Multimedia & Its applications,Vol. 2, No. 2, 2010.

Taileb M., Lamrous S. and Touati S., "Non-overlapping Hierarchical Index Structure for similarity search", International Journal of Computer Science, Vol. 3, No. 1, 2007.

Xu H., Xu D. and Lin E., "An Applicable Hierarchical Clustering Algorithm for Content-Based Image Retrieval", Springer-Verlag Berlin Heidelberg, 2007.

Yu D. and Zhang A., "ClusterTree: Integration of Cluster Representation and Nearest-NeighborSearch for Large Data Sets with High Dimensions", IEEE transaction on Knowledge and data Engineering, VOL. 15, NO. 3, MAY/JUNE 2003.

Boley D., "Principal Direction Divisive Partitioning", Data Mining and Knowledge Discovery 2, 325–344, 1998.

Chakrabarti K. and Mehrotra Sh., "The hybrid tree: a structure for high dimensional feature spaces", IEEE international confererance on data engineering, 1999.

Datta R., Joshi Jiali D. and Wang J. Z., "Image Retrieval: Ideas, Influences and Trends of the New Age", The Pennsylvania State University, ACM Transactions on Computing Surveys, 2008.

Gaede V. and Günther O., "Multi-dimensional Access Methods", ACM Computing Surveys, Vol. 30, No. 2, 1998.

Guha S., Rastogi R. and Shim K., "CURE: an efficient clustering algorithm for large databases", Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, ACM Press, pp. 73-84, 1998.

Hyvärinen A. and Oja E., "A fast fixed-point algorithm for independent component analysis", NeuralComputation, Vol. 9, NO.7, pp. 1483–1492, 1997.

Hyvärinen A. and Oja E., "Independent Component Analysis: Algorithms and Applications", Neural Networks, Vol. 13, No. (4-5), pp. 411-430, 2000.

Jackson J. E., "A User's Guide to Principal Components", New York: John Wiley and Sons, Inc, 1991.

Jolliffe I. T., "Principal Component Analysis", Second edition, New York: Springer-Verlag New York, Inc, 2002.

Keyvanpour M.R. and Asbaghi sh., "A new approach for interactive semantic image retrieval using the high level semantics", Proceedings of the ACM Symposium on Applied Computing (SAC), pp. 1175-1179, 2008.

Keyvanpour M.R. and Izadpanah N., "Analytical Classification of Multimedia Index Structures by Using a Partitioning Method-Based Framework", The International Journal of Multimedia & Its Applications, Vol. 3, No. 1, February 2011.

Keyvanpour M.R. and Izadpanah N., "Classification and Evaluation of High-dimensional Image Indexing Structures", IEEE 3rd international conference on Machin Learning and Computing (ICMLC), 2011.

Keyvanpour M.R. and Tavoli R., "Document Image Retrieval: Algorithms, Analysis and Promising Directions", International Journal of Software Engineering and Its Applications, Vol 7, No. 1, pp. 93-106, 2013.

Zhang T., Ramakrishnan R. and Livny M., "BIRCH: An Efficient Data Clustering Method for VeryLarge Databases", In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Canada, pp. 103-114, 1996.

محمدرضا کیوان‌پور: دورهٔ کارشناسی خود را در سال ۱۳۷۶ در رشتهٔ مهندسی کامپیوتر (نرم‌افزار) در دانشگاه علم و صنعت به‌پایان رساند. همچنین مدرک کارشناسی ارشد خود را در رشتهٔ مهندسی کامپیوتر (نرم‌افزار) در سال ۱۳۷۹ و مدرک دکترای خود را در سال ۱۳۸۶ در رشتهٔ مهندسی کامپیوتر (نرم‌افزار) از دانشگاه تربیت مدرس اخذ کرد.وی هم‌اکنون به‌عنوان عضو هیأت علمی دانشکدهٔ فنی و مهندسی دانشگاه الزهرا مشغول به فعالیت است. زمینه‌های علمی موردعلاقهٔ وی پردازش تصویر و داده‌کاوی است.

نشانی رایانامهٔ ایشان عبارت است از:

keyvanpour@Alzahra.ac.ir

سعیده رنجبران:دورهٔ کارشناسی خود را در سال ۱۳۸۵ در رشتهٔ مهندسی کامپیوتر (نرم‌افزار) در دانشگاه آزاد اسلامی قزوین به‌پایان رساند. وی اکنون درحال گذراندن دورهٔ کارشناسی ارشد در رشتهٔ مهندسی کامپیوتر (نرم‌افزار) درهمان دانشگاه است. زمینه‌های علمی موردعلاقهٔ وی رمزنگاریو داده‌کاوی است.

نشانی رایانامهٔ ایشان عبارت است از:

se.ranjbaran@gmail.com

نجوا ایزدپناه:دورهٔ کارشناسی خود را در سال ۱۳۸۴در رشتهٔ مهندسی کامپیوتر (نرم‌افزار) در دانشگاه علم و فرهنگبه‌پایان رساند. همچنین مدرک کارشناسی ارشد خود را در رشتهٔ مهندسی کامپیوتر (نرم‌افزار) در سال ۱۳۹۰ از دانشگاه آزاد اسلامی قزوین اخذ کرد. زمینه‌های علمی موردعلاقهٔ وی شاخص‌گذاری تصویر و داده‌کاوی است.

نشانی رایانامهٔ ایشان عبارت است از:

n.izadpanah@qiau.ac.ir