الگوریتم های خوشه بندی بر اساس درخت پوشای کمینه

سها سادات مهدی

دانشگاه صنعتی شریف

مفاهیم و الگوریتم های به کار رفته

-درخت پوشای کمینه(MST) :درخت ساخته شده از گراف Gکه شامل همه ی رئوس آن باشد و مجموع وزن یال های آن کمینه باشد.

- درخت پوشای کمینه اقلیدسی (EMST) : درخت پوشای کمین ی n نقطه در فضای متریک که طول یال ، فاصله ی اقلیدسی بین دو نقطه است.

-الگوریتم های خوشه بندی بر اساس MST بعنوان الگوریتم هایی برای یافتن خوشه های بی قاعده شناخته شده اند. از جمله کاربرد های این الگوریتم ها میتوان خوشه بندی رنگ های عکس های وب را نام برد . هدف از این خوشه بندی کاهش هزاران رنگ موجود به تعداد کمی (نماینده) از آنها که به وضوح تفاوت را بازتاب کنند است.

- پس از ساخته شدن MST به ازای ورودی داده شده ، الگوریتم های خوشه بندی دو دسته میشوند ، دسته ای که تعداد خوشه ی مطلوب به آنها داد میشود و دسته ای که تعداد خوشه ی نهایی در حین الگوریتم تعیین میشود.

- SEMST (Standard EMST clustering algorithm)

ساده ترین راه برای خوشه بندی به ازای تعداد خوشه داده شده (k) ، مرتب سازی نزولی یال های درخت و حذف k-1 سنگین ترین یال است . این الگوریتم را SEMST مینامند.

- ZEMST (Zahn's EMST clustering algorithm)

از الگوریتم هاییست که تعداد خوشه در حین فرایند الگوریتم تعیین میشود. در این الگوریتم ها یال هایی که معیارهای تناقض از پیش تعیین شده را داشته باشند از درخت حذف میشود . ما معیار های تناقضی که توسط Zahn پیشنهاد شده اند را استفاده میکنیم ، به همین دلیل این الگوریتم را ZEMST مینامیم.

الگوریتم های خوشه بندی ارائه شده در این مقاله بر پایه EMST

❖الگوریتم خوشه بندی EMST سلسله مراتبی - HEMST (hierarchical EMST clustering algorithm)

با ورودی مجموعه نقاط S در و تعداد خوشه ی مطلوب k روش سلسله مراتبی با ساخته شدن MST از مجموعه نقاط S آغاز میشود.وزن هر یال فاصله ی اقلیدسی نقاط دو سر آن است. سپس ، وزن میانگین $\overline{w}$ تمامی یال های EMST و انحراف معیار آن σ محاسبه میشود; هر یال با وزن $w>\overline{w}+\sigma$ از درخت حذف میشود. این فرآیند به مجموعه ای از زیر درخت های جدا شده $S_T = \{T_1, T_2 \dots\}$ منتهی می شود .

با هر زیر درخت به شکل یک خوشه با مرکز جرم رفتار میشود .چنانچه تعداد زیر درخت های $|S_T| < k$ ، k بلندترین یال باقیمانده از کلیه ی S_T حذف میشوند تا k زیر درخت جدا شده تولید شود.اگر $|S_T| > k$ یک نقطه ی نماینده برای هر زیر درخت شناسایی میشود. نقطه ی نماینده ی T_i برای خوشه $T_i \in S_T$ بعنوان نقطه $p \in T_i$ که نزدیکترین به مرکز جرم است ، تعریف میشود. به عبارت دیگر c_i) = min $p_j \in T_i$ d(p_j, c_i).

پس از اینکه همه ی نقاط نماینده پیدا شدند هر نقطه در هر زیر درخت با نقطه ی نماینده اش جایگزین میشود . به این ترتیب تعداد نقاط از S به S_Tکاهش یافت. یک EMST از نقاط نماینده خوشه ها ساخته میشود و پروسه ی تقسیم بندی مشابه تکرار میشود تا زمانی که $|S_T| = k$ شود خوشه بندی با تولید k خوشه پایان یافته است.

❖الگوریتم خوشه بندی کاهش بیشینه ی انحراف استاندارد MSDR (maximum standard deviation reduction clustering algorithm)

این الگوریتم ابتدا EMST را از مجموعه نقاط S میسازد . سپس انحراف معیار یال ها را در EMST محاسبه میکند ، و یالی را حذف میکند تا مجموعه ای از دو زیردرخت جدا شده داشته باشد به گونه ای که انحراف معیار کلی بیشینه شود. این حذف یال کرار میشود تا زیر درخت های بیشتری تولید شود تا زمانی که انحراف معیار به آستانه برسد.

با مجموعه نقاط S داده شده ، این الگوریتم نقاط را به گونه ای گروه بندی میکند که هر جفت نقطه در هر دسته به طور مستقیم یا غیر مستقیم در فضای متریک به یکدیگر نزدیک باشند.دو نقطه بطور مستقیم نزدیک هستند اگر فاصله ی آن دو کوچک باشد و بطور غیرمستقیم نزدیک هستند اگر از هم دور باشند اما نقطه ای در همان گروه وجود داشته باشد که هر دو نقطه به آن نزدیک باشند.این به ما کمک میکند تا خوشه هایی که اشکال هندسی ای پیچیده تر از کروی دارند شناسایی شوند. جزئیات بیشتر این الگوریتم را در سودوکود مشاهده میکنید

Table 1. The pseudocode of HEMST.

Algorithm: HEMST (k)
Initialize $n_c \leftarrow 1$ //number of clusters
Let S be the point set
Let e be an edge in the EMST constructed from S
Let w_e be the weight of e
Let σ be the standard deviation of the edge weights
Let $S_T = \emptyset$ be the set of disjoint subtrees of the EMST
Repeat
  Construct an EMST from S
  Compute the average weight $\overline{w}$ of all the edges
  Compute the standard deviation σ of the edges
  For each $e \in EMST$
    If $w_e > \overline{w} + \sigma$
    Remove e from EMST
    $n_c \leftarrow n_c + 1$
    $S_T = S_T \cup \{T'\}$ //T' is the new disjoint subtree
  // If the number of clusters n_c is less than k,
  remove $n_c - k$ longest edges so that $n_c = k$ //
  If $n_c < k$
    While $n_c \neq k$
      Remove the current longest edge
      $n_c \leftarrow n_c + 1$
      $S_T = S_T \cup \{T'\}$ //T' is the new disjoint subtree
    Return k clusters
  // If the number of clusters n_c is greater than k //
  If $n_c > k$
    Compute the centroid c_i of each $T_i \in S_T$
    Find the representative $r_i \in T_i$ closest to c_i
    $S = \cup_{T_i \in S_T} \{r_i\}$
until $n_c = k$
Return k clusters

Table 2. The pseudocode of MSDR.

Algorithm: MSDR ()
Let S be the point set
Let T_0 be the EMST constructed from S
Let S_K be the set of disjoint subtrees of T_0
Let e be an edge in S_K
Let $\sigma(S_K)$ be the overall StdDev of all edges in S_K
Let $\sigma(T_i)$ be the StdDev of edges in subtree $T_i \in S_K$
Let $\Delta\sigma(S_K)[i] = 0$ be the maximum StdDev reduction after the removal of an edge e at each iteration i
Let $\epsilon = 0.0001$
$S_K = \{T_0\}$
$\sigma(S_K) = \sigma(T_0)$
$i = 0$
Repeat
  $i \leftarrow i + 1$
  $temp = \sigma(S_K)$
  // Choose an edge that leads to max StdDev reduction once it is removed from S_K //
  For each $e \in S_K$
    Assume e is removed from S_K thus $S_K = \cup_{j=1}^{j=i} T_j$
    $\sigma(S_K) = \frac{\sum_{e' \in S_K} (T_j) \sigma(T_j)}{\sum_{e' \in S_K} |T_j|}$
    // Compute StdDev reduction //
    If $\Delta\sigma(S_K)[i] < \sigma(S_K) - temp$
      $\Delta\sigma(S_K)[i] = \sigma(S_K) - temp$
    Remove e from S_K that corresponds to $\Delta\sigma(S_K)[i]$
    $\sigma(S_K) = temp - \Delta\sigma(S_K)[i]$
until $|\Delta\sigma(S_K)[i] - \Delta\sigma(S_K)[i-1]| < |\epsilon \cdot (\Delta\sigma(S_K)[i] + 1)|$
$f(j) = PolyRegression(\cup_{j-1}^{j} \Delta\sigma(S_K)[j])$
// No. of clusters corresponds to the 1st local minimum //
$K = \min\{j \in [1, i]\}$ that satisfies $f'(j) = 0$ & $f''(j) > 0$
Return $S_K = \{T_1, \dots, T_K\}$

نتایج تجربی به دست آمده

دو آزمایش انجام شده ، در اولی سه مسئله ی خوشه بندی در نظر گرفته شده و دو الگوریتم ارائه شده با الگوریتم های EM (بیشینه کردن امید ریاضی) و k-means مقایسه میشوند . در آزمایش دوم الگوریتم های ارائه شده با SEMST و ZEMST در در خوشه بندی رنگ عکس مقایسه میکنیم.

مقایسه HEMST با k-means و EM با MSDR

مسئله ی اول (شکل 3) 2 خوشه ی نهایی مطلوب ماست که هر خوشه به صورت یک خط منحنی نمایش داده میشود . مسئله ی دوم (شکل 4) 2 خوشه را شامل میشود که یکی داخل دیگری محاط شده است. مسئله ی سوم (شکل 5) شامل دو خوشه با چگالی ناهمگن میشود.میدانیم HEMST و k-means عدد خوشه ی از پیش تعین شده نیاز دارند.همانطور که پیداست در شکل 3 k-means خوشه ها(ی مطلوب ما) را به دو قسمت تقسیم میکند و نتیجه ای از ترکیب آن دو به ما میدهد.HEMST و EM یک خوشه را تجزیه میکنند و تنها MSDR بطور موفقیت آمیز نتیجه ی مطلوب را داده است.

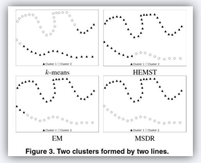

Figure 3. Two clusters formed by two lines.

به طور مشابه در شکل 4 k-means اجزای درونی و بیرونی را به دو قسمت تقسیم میکند در حالی که HEMST و EM بخش درونی را تجزیه نمیکنند اگر چه که بخش بیرونی در نهایت تجزیه شده و باز هم MSDR به زور صحیح نتیجه مطلوب را ارائه میدهد.

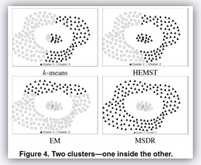

Figure 4. Two clusters—one inside the other.

در شکل 5 k-means و HEMST تلاش بر گروه بندی بخش چگال تر در یک خوشه دارند ، EM تنها یک خوشه میدهد و MSDR 3 خوشه تولید میکند بطوریکه بخش چگال تر یک خوشه و دو بخش طرفین دو خوشه ی دیگر را تشکیل میدهند.

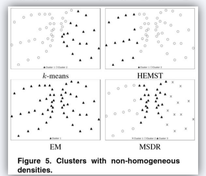

Figure 5. Clusters with non-homogeneous densities.

با توجه به آزمایش ها و چند نمونه ای که مشاهده کردید الگوریتم MSDR در شناسایی خوشه های مطلوب موفق تر عمل کرده است.

خوشه بندی رنگی با SEMST و ZEMST

1 - خوشه بندی با k داده شده

شکل 7 نتایج HEMST و SEMST روی عکس GIF شامل 128 رنگ متمایز را نشان میدهد. با توجه به شکل تفاوت نتایج دو الگوریتم به ازای k های یکسان قابل مشاهده است. عکس اصلی شماره 7 شامل 69 رنگ متمایز و عکس اصلی شماره 9 شامل 189 رنگ متمایز است.

نتایج تمامی آزمایشات انجام شده نشان میدهد الگوریم HESMST کارایی بسیار بالاتری نسبت به SEMST در خوشه بندی مطلوب رنگ ها و در واقع تمیز دادن آنها به ازای kهای یکسان دارد.

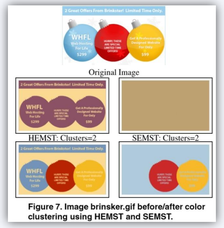

Figure 7. Image brinsker.gif before/after color clustering using HEMST and SEMST.

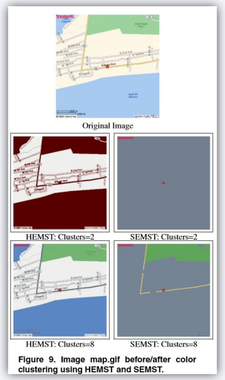

Figure 9. Image map.gif before/after color clustering using HEMST and SEMST.

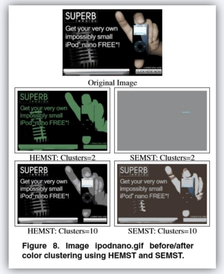

Figure 8. Image ipodnano.gif before/after color clustering using HEMST and SEMST.

2-خوشه بندی بدون k داده شده

آزمایش های مشابه برای مسائل بدون k ورودی ، که توسط دو الگوریتم ZEMST و MSDR تست شدند ، انجام گرفتند جزئیات به کار رفته در آزمایش ZEMST در شکل 11 نشان داده شده است.

مشاهدات عمومی نشان میدهد الگوریتم MSDR میتواند در حین حفظ کردن اشیائ شکل ، تعداد بسیار کمتری رنگ (خوشه - k) نسبت به الگوریتم ZEMST تولید کند.(شکل 10 و 11)

Figure 10. The original image giftcard.gif and the output of our MSDR algorithm.

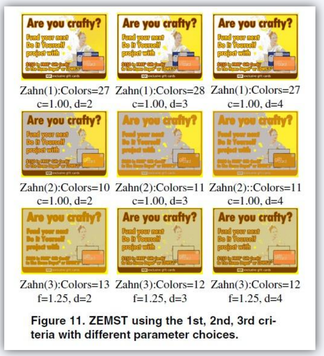

Figure 11. ZEMST using the 1st, 2nd, 3rd criteria with different parameter choices.

مراجع

[1] O. Grygorash, Y. Zhou, Z. Jorgensen. Minimum Spanning Tree Based Clustering Algorithms.