کلیه حقوق مادی و معنوی این مقاله متعلق به دبیرخانه دائمی کنفرانس داده کاوی ایران است و درج آن فقط در سایت رسمی کنفرانس داده کاوی ایران مجاز بوده و هرگونه انتشار غیر از آن، خلاف عرف، اخلاق و قانون و مستوجب پیگیری است.

خوشه‌بندی افزایشی با استفاده از الگوریتم کرم شب‌تاب

خدیجه نظری[۱]، بابک نصیری[۲]، محمدرضا میبدی[۳]

چکیده

امروزه بسیاری از سازمان‌ها مجموعه‌های داده بسیار بزرگی دارند که مجموعه داده آنها به صورت پویا تغییر می‌کند. در بسیاری از سیستم‌های دسته‌بندی این یک مسئله بزرگ است، از این رو تغییرات داده ممکن است منجر به نتایج ضعیفی در بازآموزی صحیح گردد. هدف از این مقاله ایجاد دانش در مورد مدل افزایشی برای تغییر پایگاه داده به صورت پویا است. از یک دسته از عامل‌های خاص همچون دسته کرم‌های شب‌تاب استفاده می‌کنیم و رفتار طبیعی آنها تقلید می‌شود تا بتدریج شکل اختیاری خوشه‌ها تشکیل شود. مشخص کردن خوشه‌ها از قبل غیرضروری است. آزمایشات نشان می‌دهد که نتایج روش افزایشی تقریبا به خوبی کیفیت روش خوشه‌بندی ایستا است، اما برای مجموعه‌های داده بزرگ از روش ایستا سریع‌تر است.

کلمات کلیدی

کرم شب‌تاب، خوشه‌بندی، کاربرد وب کاوی، خوشه‌بندی افزایشی

Incremental Clustering using by the firefly algorithm

Khadije Nazari, Babak Nasiri, Mohammad Reza Meybodi

ABSTRACT

Today, many organizations have sets of very large data that datasets dynamically change. This is a big issue in many classification systems; therefore, changes may lead to poor results in the correct training. The purpose of this paper is an incremental knowledge model for a dynamically changing database. We use a special class of agent such as groups of firefly and their normal behavior are imitated to arbitrary shape of clusters gradually form. Clustering is unnecessary to specify in advance. Experiments show that the results of the incremental approach are almost as good as the quality of the static clustering methods but faster than static for large data sets.

Keywords: Firefly, Clustering, Web mining, Incremental clustering.

[۱] دانشکده برق، رایانه و فناوری اطلاعات، دانشگاه آزاد اسلامی، قزوین، ایران khadije.nazari.sasi@gmail.com
[۲] دانشکده برق، رایانه و فناوری اطلاعات، دانشگاه آزاد اسلامی، قزوین، ایران nasiri.babak@qiau.ac.ir
[۳] دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران mmeybodi@aut.ac.ir

۱-مقدمه

با گسترش هر چه بیشتر وب گسترده جهانی[۱]، اطلاعات موجود بر روی اینترنت به صورت مکرر به روز رسانی می‌شود و الگوهای ناوبری کاربران به صورت پیوسته تغییر می‌کند. (شامل اضافه شدن صفحات جدید، اضافه شدن کاربران جدید، تغییر صفحات کاربران قدیمی است).[۱] اینترنت به سبب پویایی خاصی که دارد زمینه فعالیت در کاربرد کاوی وب[۲](WUM) را فراهم نموده است. به طور کلی ورودی کاربرد کاوی وب یک مجموعه از URL ها است، که توسط یک کاربر خاص در یک زمان محدود ملاقات می‌شود که یک جلسه نامیده می‌شود.[۲] یک مجموعه از URLها و فراوانی و تکرارهای آن که مربوط به علاقه‌مندی یک کاربر خاص است که با ورود به اینترنت آنها را ملاقات می‌نماید پروفایل مربوط به آن کاربر خاص نامیده می‌شود.[۲] پروفایل‌های کشف شده به عنوان فراوانی مجموعه آیتم‌ها و الگوها در نظر گرفته می‌شوند و خلاصه‌ای از داده ورودی فراهم می‌نماید، پروفایل نزدیکترین حالت به داده ورودی اصلی در نظر گرفته می‌شود.[۲] در سال‌های اخیر روش‌های بسیاری بر روی WUM پیشنهاد شده است که می‌توان آن را به دو دسته کلی تقسیم‌بندی کرد:[۱]

۱. تکرار الگوها بر پایه روش‌ها
۲. خوشه‌بندی[۳] بر پایه روش‌ها

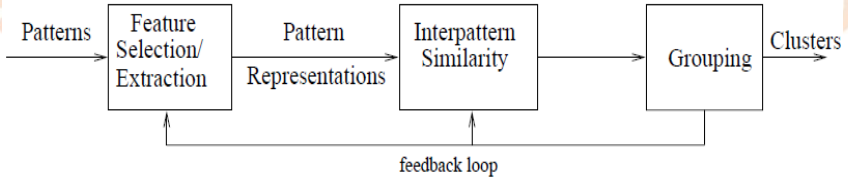
به طور کلی بیشتر روش‌های ارائه شده ایستا هستند و برای داده‌هایی که به صورت پویا تغییر می‌کنند کارایی چندانی ندارند. در این پژوهش ما یک الگوریتم خوشه‌بندی افزایشی با استفاده از کرم‌های شب‌تاب ارائه می‌دهیم که در زمینه کاربرد کاوی وب به کار می‌رود. در بخش‌های بعدی مقاله ابتدا تعریفی ساده از خوشه‌بندی ارائه می‌دهیم، سپس الگوریتم کرم شب‌تاب را معرفی می‌کنیم. در انتها به بررسی خوشه‌بندی افزایشی الگوریتم کرم شب-تاب و نتایج آزمایشات و شبیه‌سازی‌ها می‌پردازیم.

۲-خوشه‌بندی

خوشه‌بندی مسئله پیدا کردن گروههایی در مجموعه داده مطابق با بعضی خواص و ویژگی‌ها است که این ویژگی‌ها در بعضی زمینه‌ها دارای یک مفهوم است.[۱،۲،۳] یک فرایند خوشه‌بندی شامل مراحل زیر است:

۱. نمایش الگو (این مرحله شامل استخراج ویژگی است)
۲. تعریف یک معیار مجاورت یا تعریف همسایگی
۳. خوشه‌بندی
۴. استخراج داده (اگر لازم باشد)
۵. بررسی و ارزیابی خروجی (اگر لازم باشد)

شکل (۱) مراحل خوشه‌بندی را نشان می‌دهد:[۳]



شکل(۱) مراحل خوشه‌بندی

[۱] Word Wide Web
[۲] Web Usage Mining
[۳] Clustering

خوشه‌بندی در زمینه تشخیص الگو، پردازش تصویر، بازیابی اطلاعات و کاربرد کاوی وب کاربردهای فراوانی دارد علاوه بر این خوشه‌بندی در سایر رشته-های علمی همچون زیست شناسی، روان‌پزشکی، روان‌شناسی، باستان‌شناسی، زمین‌شناسی، جغرافیا و بازار کاربرد دارد. عباراتی نظیر یادگیری بدون ناظر، طبقه‌بندی عددی، بردار تدریجی و یادگیری توسط مشاهده معادل با خوشه‌بندی است که در علوم مختلف مورد استفاده قرار می‌گیرد.[۳]

یکی از روش‌های متفاوت که برای خوشه‌بندی مورد استفاده قرار می‌گیرد، هوش جمعی(SI) می‌باشد. هوش جمعی یک نمونه الگوی هوش مصنوعی است که به طور اساسی از پویایی و تعاملات چندین اجتماع همچون کلونی مورچه‌ها، دسته پرندگان و دسته ماهی‌ها الهام گرفته شده است. SI بر پایه رفتارهای اجتماعی، همکاری و ساخت یافتگی عامل‌های غیرمتمرکز و خودسامانده به وجود آمده است. اگرچه این عامل‌ها ظرفیت بسیار محدودی در حل مسائل دارند، اما با همکاری هم وظایف بسیار پیچیده‌ای را انجام می‌دهند. ویژگی‌های هوش دسته‌ای به صورت زیر است:[٤]

۱. همکاری: عامل‌ها در دسته با یکدیگر همکاری می‌کنند و با سایر عامل‌ها و محیط تعامل دارند.
۲. هوش جمعی: در حالی که اساسا عامل‌های موجود در یک دسته غیرهوشمند هستند، عامل‌ها با تعامل با یکدیگر و کارکردن تحت مکانیزم دسته یک سیستم هوشمند به وجود می‌آورند.
۳. الهام از طبیعت
۴. کنترل غیرمتمرکز

عمومی‌ترین الگوریتم‌های هوش جمعی که برای خوشه‌بندی استفاده می‌شود عبارتند از:

۱. خوشه‌بندی کلونی مورچه‌ها
۲. خوشه‌بندی دسته ذرات
۳. خوشه‌بندی بر پایه دسته‌ای از عامل‌ها

اخیرا از الگوریتم کرم شب‌تاب نیز برای خوشه‌بندی استفاده شده است که جزئیات آن در منبع [۵] قابل مشاهده است. در ادامه به بررسی این الگوریتم و کاربرد آن در زمینه خوشه‌بندی می‌پردازیم.

۳- الگوریتم کرم شب‌تاب

کرم‌های شب‌تاب[۱] حشره‌های روشنایی نامیده‌می‌شوند و در جهان حدود ۲۰۰۰ گونه متفاوت از این حشره وجود دارد. محل زندگی کرم‌های شب‌تاب بیشتر در نواحی هاره‌ای و مناطق معتدل است. این حشرات از طریق تابش نور با یکدیگر ارتباط برقرار می‌کنند، که مناظر سحرانگیزی را در آسمان شب ایجاد می‌کنند. الگوریتم کرم شب‌تاب[۲] اولین بار توسط شن‌شی‌یانگ[۳] از دانشگاه کمبریج در سال ۲۰۰۸ معرفی شد، به طور کلی این الگوریتم جزء الگوریتم‌های ملهم از طبیعت[۴] دسته‌بندی می‌شود.[۶،۸،۹،۱۰،۱۱،۱۲] شن‌شی‌یانگ سه قانون کلی را در مورد الگوریتم کرم شب‌تاب مطرح ساخت:

۱. همه کرم‌های شب‌تاب هم‌جنس[۵] در نظر گرفته می‌شوند در اینصورت هر کرم شب‌تاب ، کرم‌های دیگر را بدون در نظر گرفتن جنسیت آن جذب می‌کند.
۲. میزان جذابیت یک کرم شب‌تاب[۶] با درخشندگی[۷] آن کرم شب‌تاب تعیین می‌گردد بدین صورت که بین دو کرم شب‌تاب ، کرم با درخشندگی درخشندگی یا نور کمتر به طرف کرم پرنورتر یا با درخشندگی بیشتر حرکت می‌کند ، در این حالت کرم شب‌تاب کم نورتر از جذابیت کمتری نسبت به کرم شب‌تاب پرنورتر برخوردار است.در صورت افزایش فاصله از تابع هدف میزان جذابیت و درخشندگی کاهش می‌یابد. اگر در میان مجموعه کرم‌های شب‌تاب هیچ‌کدام درخشان‌تر نباشد حرکت کرم‌ها تصادفی خواهد بود.
۳. میزان درخشندگی یک کرم شب‌تاب توسط تابع هدف تعیین می‌گردد .[۷،۸،۹،۱۰،۱۱،۱۲]

---

[۱] Firefly
[۲] Firefly algorithm(FA)
[۳] Xin-she yang
[٤] Bio-inspired
[٥] unisex
[٦] Attractiveness
[۷] brightness

شبه کد الگوریتم کرم شب‌تاب به صورت زیر ارائه می‌گردد:

```
Objective function f(x), x = (x۱, ..., xd)^T
Generate initial population of fireflies xi (i = ۱, ۲, ..., n)
Light intensity Ii at xi is determined by f(xi)
Define light absorption coefficient
while (t <MaxGeneration)
for i = ۱ : n all n fireflies
for j = ۱ : i all n fireflies
if (Ij > Ii), Move firefly i towards j in d-dimension; end if
Attractiveness varies with distance r via exp[−γr]
Evaluate new solutions and update light intensity
end for j
end for i
Rank the fireflies and find the current best
end while
Postprocess results and visualization
```

در الگوریتم کرم شب‌تاب دو مسئله مهم وجود دارد : نوسانات شدت نور[۱] و فرمول‌بندی کردن میزان جذابیت یک کرم شب‌تاب.[۷،۱۱،۱۲]

میزان جذابیت یک کرم شب‌تاب (β): این پارامتر بر نیرومندی یک کرم شب‌تاب در جذب سایر کرم‌های شب‌تاب دلالت دارد ، پارامتر را باید طوری در نظر گرفت که فاصله کرم شب‌تاب با نقطه هدف به طور یکنواخت کاهش یابد و در جهت سادگی ما می‌توانیم β را با I که نشان دهنده شدت نور یا درخشندگی است متناسب در نظر بگیریم شدت نور تمایل آنرا به یکپارچه شدن با تابع هدف کدشده نشان می‌دهد ($I(x) \propto F(x)$) معمولا شدت نور را عکس تابع هزینه در نظر می‌گیرند . شدت نور در منبع را با (I(s نشان می‌دهند و طبیعی است که با فاصله از منبع نور از شدت نور کاسته خواهد شد در ساده‌ترین شکل شدت نور را می‌توان به این صورت فرمول (۱) بیان کرد:

$$I(r) = \frac{I(r)}{r^2} \qquad (۱)$$

که در فرمول (۱) (I (s شدت نور در فاصله r از منبع است.همچنین این نکته را نیز باید در نظر گرفت که شدت نور توسط رسانه[۲] یا فضا جذب می‌شود. هوا نور را جذب می‌کند و به همین دلیل با افزایش فاصله نور ضعیف‌تر و ضعیف‌تر می‌شود،این عامل باعث می‌شود تا کرم‌های شب‌تاب تنها در یک فاصله محدود مشهود هستند.این فاصله در حدود چند صد متر مربع است که برای ارتباط کرم‌های شب‌تاب در شب مناسب است، پس در فرمول فوق درجه جذب[۳] (γ) نیز موثر است . رسانه‌های مختلف ضریب جذب متفاوتی دارند و با توجه به مسئله مقدار ضریب جذب از آن استنتاج می‌شود، در حالت کلی برای یک رسانه با ضریب جذب (γ) شدت نور (I) به این صورت فرمول (۲) بیان می‌شود:

$$I(r) = I_0 e^{-\gamma r^2} \qquad (۲)$$

در عبارت فرمول (۲) .I شدت نور در منبع است .[۶،۷،۸،۱۰،۱۱،۱۲] اگر در یک مسئله کاربردی خاص نیاز به تابع‌ای داشتیم که به تدریج و یکنواخت کاهش پیدا کند، می‌توانیم از عبارت

$$I(r) = \frac{I_0}{1+\gamma r^2} \qquad (۳)$$

استفاده می‌کنیم. حال می‌توانیم میزان جذابیت یک کرم‌شب‌تاب را توسط فرمول (۴) بیان کنیم:

$$\beta(r) = \beta_0 e^{-\lambda r^2} \qquad (۴)$$

β_0 معادل میزان اهمیت کرم شب‌تاب در r = ۰ است و در بعضی از مسائل از فرمول (۵) استفاده می‌کنند:

$$\beta(r) = \frac{\beta_0}{1+\lambda r^2} \qquad (۵)$$

---

[۱] variation of light intensity
[۲] media
[۳] light absorption coefficient

کنفرانس داده کاوی ایران

چون محاسبه توابع نمایی مستلزم سربار زمانی بیشتری است، در بسیاری از مسائل از فرمول (۵) استفاده می‌کنند. به طور کلی میزان جذابیت یک کرم شب‌تاب در دید شاهدان و یا توسط سایر کرم‌های شب‌تاب تعیین می‌گردد. به منظور حرکت کرم شب‌تاب i با درجه جذابیت و درخشندگی کمتر به طرف کرم شب‌تاب j با درجه جذابیت و درخشندگی بیشتر از معادله حرکت (۶) استفاده می‌شود:

$$X_i = x_i + \beta_0 e^{-\gamma r_{ij}^2}(x_j - x_i) + \alpha\left(rand - \frac{1}{2}\right) \qquad (۶)$$

در طرف راست معادله (۶) عبارت دوم تحت تاثیر پارامتر ضریب جذب و میزان جذابیت اولیه کرم شب‌تاب قرار دارد ، درحقیقت سرعت جذب یک کرم شب‌تاب را به طرف نقطه بهینه مشخص می‌کند و در عبارت سوم α یک پارامتر تصادفی است که نوسانات حرکت توصیف می‌کند . وجود نوسانات در مسیر حرکت یک کرم شب‌تاب سبب می‌شود تا به طور مستقیم به طرف نقطه بهینه حرکت نکند در اینصورت اگر بهینه دیگری وجود داشته باشد احتمال پیدا شدن چنین بهینه‌هایی افزایش می‌یابد و rand مولد یک عدد تصادفی است ، که به طور یکنواخت در [1 و 0] توزیع شده است. در پیاده-سازی‌ها اغلب $\beta_0 = ۱$ و $\alpha \in [۰ , ۱]$ در نظر می‌گیرند.

برای تعریف فاصله بین دو کرم شب‌تاب یعنی r لزومی به استفاده از فاصله اقلیدوسی نیست و می‌توان معیارهای دیگری نیز به کار برد ، مثلا برای مسائل زمانبندی، r می‌تواند به عنوان تاخیر زمانی یا فاصله زمانی تعریف شود و برای شبکه‌های پیچیده همچون اینترنت ، r می‌تواند به عنوان درجه خوشه‌بندی و میزان نزدیکی رئوس تعریف شود ، در واقع هر پارامتری که بتواند به طور موثر کمیت موجود در مسائل بهینه‌سازی را توصیف نماید می-تواند به عنوان پارامتر r تعریف شود . همچنین باید این امر را مد نظر قرار دهیم که ، کرم‌های شب‌تاب باید به‌طور یکنواخت در فضای جستجو توزیع شوند تا بتوانیم همه بهینه‌های سراسری و محلی موجود در فضای جستجو را کشف کنیم ، این کار مشابه مقدار اولیه دادن در شبیه‌سازی مونت کارلو انجام می‌شود . با توجه به توزیع یکنواخت کرم‌های شب‌تاب در فضای جستجو می‌توان با ۵۰ تا ۱۰۰ تکرار جواب بهینه را به‌دست آورد. در مسائل کاربردی جمعیت کرم‌های شب‌تاب در حدود ۵۰ – ۴۰ عدد در نظر گرفته می‌شود ولی برای مسائل پیچیده‌تر می‌توان از تعداد بیشتری کرم شب‌تاب استفاده کرد.[۶،۸،۱۲]

۴-خوشه‌بندی افزایشی با استفاده از الگوریتم کرم‌شب‌تاب

امروزه بسیاری از سازمان‌ها مجموعه داده خیلی بزرگی دارند که مجموعه داده آنها به صورت پویا تغییر می‌کند. این تغییرات داده ممکن است منجر به نتایج ضعیفی در بازآموزی داده گردد و به همین دلیل از روش خوشه‌بندی افزایشی استفاده می‌شود. [۱۳] در روش خوشه‌بندی افزایشی مجموعه داده به صورت یکجا در اختیار ما قرار ندارد، بلکه در فواصل زمانی معین مقداری از مجموعه داده برای خوشه‌بندی و پردازش در اختیار ما قرار می‌گیرد. باید توجه داشته باشیم که ممکن است با اضافه شدن قسمتی از مجموعه داده ساختار خوشه‌بندی تغییر نماید، خوشه‌ها حرکت کرده و یا حتی در یک زمان خوشه ناپدید شده و دوباره ظاهر گردد.[۱۴]

روش‌های خوشه‌بندی صرف نظر از اشیاء در داخل گروه‌ها یا کلاس‌ها بر پایه یادگیری بدون ناظر توسعه یافته است. در تکنیک بدون ناظر مجموعه داده آموزشی در ابتدا بر پایه اطلاعات عددی در داده (مانند مراکز خوشه‌ها) گروه بندی می‌شود و سپس توسط تحلیلگر کلاس‌های اطلاعاتی تطبیق داده می‌شود. مجموعه داده‌هایی که ردگیری می‌کنیم شامل اطلاعات کلاس‌ها برای هر داده است. بنابراین هدف اصلی پیدا کردن مراکز خوشه‌ها توسط کمینه کردن تابع هدف است، مجموع فاصله الگوها مراکز آن است. برای N شی در مسئله داده شده هدف کمینه کردن مجموع مربع فاصله اقلیدسی بین هر الگو و اختصاص دادن هر کدام از الگوها به یکی از K مرکز خوشه است. تابع هدف خوشه بندی، مجموع مربع خطا از طریق فرمول (۷) محاسبه می‌شود:

$$J(K) = \sum_{k=1}^{k}\sum_{i\in c_k}(x_i - c_k) \qquad (۷)$$

در فرمول (۷) K تعداد خوشه‌ها، برای n الگو x (I=۱،۲،……..,n) مکان iامین الگو و c_k (k=۱،۲،… …..,K) kامین مرکز خوشه است که توسط فرمول (۸) محاسبه می‌شود:

$$c_k = \sum_{i\in c_k}\frac{x_i}{n_k} \qquad (۸)$$

در معادله (۸) n_k تعداد الگوها در خوشه k ام است.

تجزیه و تحلیل خوشه به این شکل می‌باشد که مجموعه داده‌ها به خوشه‌ها اختصاص می‌یابد به طوریکه الگوها بر اساس بعضی معیارهای شباهت در یک خوشه گروه بندی می‌شوند. برای ارزیابی شباهت بین الگوها معمولا از معیار اندازه‌گیری استفاده می‌شود. مراکز خوشه متغییرهای تصمیم‌گیری هستند

کنفرانس داده‌کاوی ایران

که به وسیله کمینه کردن مجموع فاصله اقلیدسی بر روی همه مثال های آموزشی در فضای n بعدی به دست می‌آید. تابع هزینه (هدف) برای الگوی i توسط معادله (۹) محاسبه می‌شود:

$$f_i = \frac{1}{D_{Train}} \sum_{j=1}^{D_{Train}} d(x_j, p_i^{CL_{known(x_j)}}) \qquad (۹)$$

در فرمول (۹) D_{Train} تعداد مجموعه داده آموزشی است که برای نرمالیزه کردن جمع استفاده می‌شود که در محدوده فاصله‌ای بین [۰،۰،۱،۰] قرار دارد و $p_i^{CL_{known(x_j)}}$ کلاسی تعریف می‌شود که نمونه مطابق پایگاه داده به آن تعلق دارد. توجه نمائید که در الگوریتم FA متغییرهای تصمیم‌گیری، مراکز خوشه‌ها هستند. تابع هدف در الگوریتم کرم‌شب‌تاب توسط فرمول (۹) مشخص می‌شود. برای یک مجموعه داده n تعداد نقاط داده هایی، d بعد مسئله و c تعداد کلاس را نشان می‌دهد. یک نقطه داده‌ای تنها به یکی از c کلاس تعلق دارد. مراکز خوشه با استفاده از معادله (۹) به دست آید.[۵] در ادامه به بررسی نتایج آزمایشات می‌پردازیم.

۵-نتایج آزمایشات

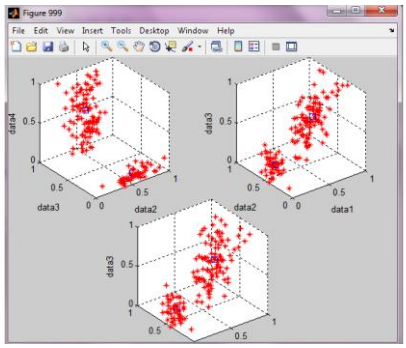
شبیه‌سازی‌ها بر روی سیستمی با ۲ گیگابایت حافظه RAM و CPU cori۵ که ویندوز ۷ بر روی آن نصب بوده انجام شده است. در ابتدا داده‌ها نرمال-سازی می‌شود و تعداد کرم‌های شب‌تاب به کار رفته در این شبیه‌سازی برابر ۲۰ است، مقدار متغییرهای $\alpha = 0.25$ ، $\beta = 0.2$ و $\gamma = 1$ است. در این شبیه‌سازی از Dataset های Dermatology, Diyabet, Balance Scale, Iris, Datalog استفاده شده است. Datalog را می‌توان از وب سایت http://maya.cs.depaul.edu/~classes/ect۵۸۴/resource.html دانلود کرد و سایر Dataset ها را می‌توان از پایگاه داده UCI دانلود نمود. نتایج شبیه‌سازی نشان می‌دهد که برای Datasetهایی که تعداد نمونه‌های کمتری دارند و تعداد ویژگی‌های آن نیز کم است، الگوریتم کرم‌شب‌تاب ایستا از سرعت بالاتری برخوردار است اما با افزایش تعداد نمونه‌های یک Dataset الگوریتم کرم شب‌تاب افزایشی سرعت بالاتری دارد و در هر دو حالت کیفیت خوشه‌بندی الگوریتم کرم‌شب‌تاب افزایشی بهتر است. سرعت اجرایی الگوریتم‌های کرم شب‌تاب ایستا و پویا بر روی Dataset های متفاوت در جدول(۱) نشان داده شده است.

جدول(۱) مقایسه سرعت اجرایی الگوریتم خوشه‌بندی ایستا و افزایشی کرم شب‌تاب

الگوریتم / مجموعه داده	Iris	Balance Scale	Dermatology	Deyabet	Datalog
خوشه‌بندی ایستا کرم شب‌تاب	۳,۹۶	۹,۶	۱۱,۳	۲۶,۶	۱۲۰,۳
خوشه‌بندی افزایشی کرم شب‌تاب	۱۲,۵۲	۱۵,۳	۱۵,۲۹	۱۵,۳۹	۲۵,۲

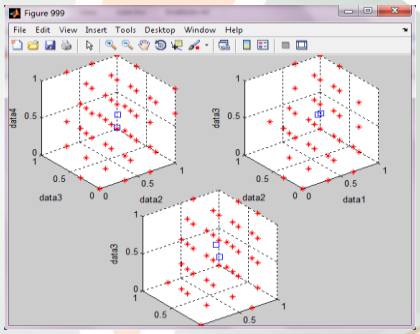
مجموعه داده Iris چهار ویژگی دارد که برای رسم نمودار سه بعدی سه ویژگی آن انتخاب و نمایش داده شده است. با در نظر گرفتن سه ویژگی سه نمودار ایجاد می‌شود که در شکل (۲) نشان داده شده است.

کنفرانس داده کاوی ایران



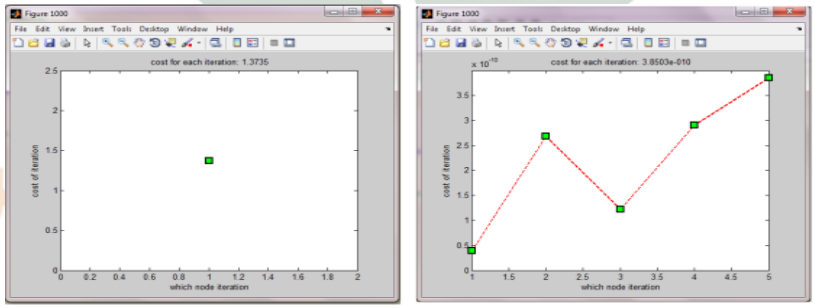
شکل(۲) خوشه‌بندی Iris dataset

مجموعه داده Dermatology در مجموع ۳۳ ویژگی دارد و به دلیل زیاد بودن ویژگی‌ها و اینکه نمودار سه بعدی تنها ۴ ویژگی اول را در نظر می‌گیرد، در این شکل مراکز خوشه به خوبی قابل رویت نیست.



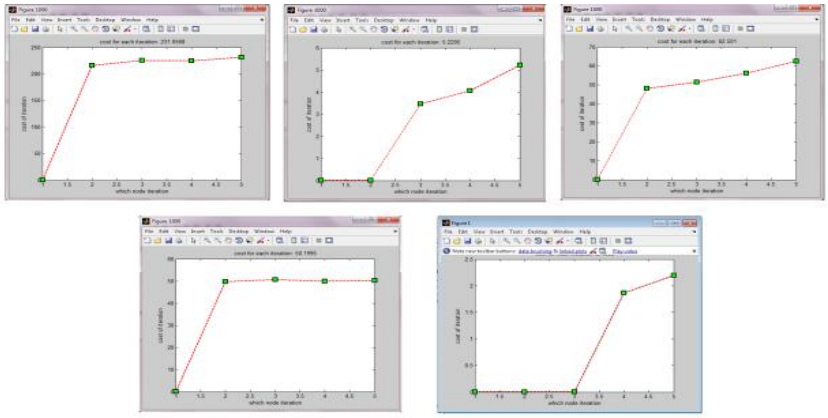
شکل(۳) خوشه‌بندی Dermatology dataset

در خوشه‌بندی افزایشی در ابتدا مجموعه داده را به ۵ قسمت تقسیم می‌کنیم و سپس یک قسمت را انتخاب و خوشه‌بندی می‌کنیم و قسمت‌های باقیمانده در مراخل بعدی اضافه شده و خوشه‌بندی به‌روزرسانی می‌شود. شکل (٤) هزینه خوشه‌بندی دو حالت ایستا و پویا را برای مجموعه داده Iris نشان داده است.



شکل (۴) مقایسه هزینه خوشه‌بندی در دو حالت ایستا و افزایشی برای مجموعه داده Iris

اگر شکل (۴) را مشاهده نمائید در نمودار سمت چپ بدلیل اینکه کل Dataset به یکباره برای خوشه‌بندی در اختیار ما قرار می‌گیرد مختصات نقطه‌ای که در شکل نشان داده شده است در یک بعد عدد ۱ را نشان می‌دهد. نمودار هزینه هر کدام از Data set های به کار رفته در حالت افزایشی در شکل (۵) نشان داده شده است.



شکل(۵) نمودار هزینه خوشه‌بندی افزایشی هر کدام از Datasetها که نمودارها از سمت چپ مربوط به مجموعه داده‌های

Dermatology- Iris- Balance scale- Diyabet- Datalog

۶-نتیجه‌گیری

با توجه به گسترش بانک‌های اطلاعاتی سیستم‌های کنونی، نیاز به روش‌های جدیدی برای خوشه‌بندی و دسته‌بندی داده وجود دارد که بتواند با توجه به حجم زیاد داده‌ها کارایی خوبی از خود نشان دهد. در این مقاله ما به بررسی خوشه‌بندی با استفاده از الگوریتم کرم شب‌تاب به روش افزایشی پرداختیم. هدف از این مقاله ایجاد دانش در مورد مدل افزایشی برای تغییر پایگاه داده به صورت پویا است. از یک دسته از عامل‌های خاص همچون دسته کرم‌های شب‌تاب استفاده می‌کنیم و رفتار طبیعی آنها تقلید می‌شود تا بتدریج شکل اختیاری خوشه‌ها تشکیل شود. مشخص کردن خوشه‌ها از قبل غیرضروری است. آزمایشات و شبیه‌سازی‌های انجام شده بیان‌گر این مطلب است که نتایج روش افزایشی تقریبا به خوبی روش خوشه‌بندی ایستا است و برای مجموعه داده‌های بزرگ روش افزایشی سرعت بالاتری نسبت به روش ایستا دارد.

۷-مراجع

[۱] J. Shen, Y. Lin, and Z. Chen, "***Incremental Web Usage Minig based on Active Ant Colony***", Wuhan University Journal of Natural Sciences, Vol. ۱۱, pp. ۱۰۸۱- ۱۰۸۵, ۲۰۰۶.

[۲] E. Saka, O. Nasraoui, ***" On Dynamic Data Clustring and Visualization Using Swarm Intelligence"***, Data Engineering Workshops(ICDEW), International Conference IEEE, pp. ۳۳۷- ۳٤۰, ۲۰۱۰.

[۳] A.K. Jain, M.N. Murty, and P.J.Flynn, " ***Data Clustring: A Review***", ACM Comouting Surveys, Vol. ۳۱, pp. ۲٦٤- ۳۲۳, ۱۹۹۹.

[۴] E. Saka, and O. Nasraoui, " ***Improvements in Flock-Based Collaborative Clustering Algorithm***", Springer Berlin Heidelberg, M. Christine, J. Lakhmi, Vol. ۱, ۲۰۰۹.

[۵] J. Senthilnath, S.N. Omkar, and V. Mani, " ***Clustring Using Firefly Algorithm: Performance Study***", Swarm and Evolutionary Computation, Vol. ۱, ۲۰۱۱.

[۶] S. Lukasik, and S. Zak, " ***Firefly Algorithm for Continuous Constrained Optimization Task***", in Proceeding of the International Conference on Computer and Computational Intelligence (ICCCI), N. T.Nguyen, R. Kowlczyk, and S. M. Chen, Vol. ۵۷۹۲, pp.۹۷- ۱۰٦, Springer, ۲۰۰۹.

[۷] X. S. Yang, " ***Firefly Algorithm, Stochastic Test Functions and Design Optimisation***", International Journal of Bio-Inspired Computation, Vol. ۲, PP. ۷۸- ۸٤, ۲۰۱۰.

[۸] X. S. Yang, " ***Firefly Algorithms for Multimodal Optimization***", in Stochastic Algorithms: Foundations and Applications, SAGA, Lecture Notes in Computer Sience, Vol. ۵۷۹۲, PP. ۱٦۸- ۱۷۸, ۲۰۰۹.

[۹] X. S. Yang, " ***Firefly Algorithm, Levy Flights and Global Optimization***", in Research and Development in Intelligent Systems XXVI, Springer London, PP. ۲۰۹- ۲۱۸, ۲۰۱۰.

[۱۰] X. S. Yang, and S. Deb, " ***Eagle Strategy Using Levy Walk and Firefly Algorithms for Stochastic Optimisation***", in Nature Inspired Cooperative Strategies for Optimization (NICSO), Springer, CSI Vol. ۲۸٤, PP. ۱۰۱-۱۱۱, ۲۰۱۰

[۱۱] M. K. Sayadi, R. Ramezanian, and N. Ghaffari-Nasab, " ***A Discrete Firefly Meta-Huristic with Local Search for Makespan Minimization in Permutation Flow Shop Scheduling Problems***", International Journal of Industrial Engineering Computations, Vol. ۱, PP.۱- ۱۰, ۲۰۱۰

[۱۲] X. S. Yang, " ***Engineering Optimization: An Introduction with Metaheuristic Applications***", Wiley and Sons, New Jersey, ۲۰۱۰.

[۱۳] B. Liu, and B. Mckay, “ ***Incremental Clustring Based on Swarm Intelligence***”, W. Tzai-Der, L. Xiaodong, C. shu-Heng, W. Xufa, A. Hussein, L. Hitoshi, C. Guo-liang and Y. Xin, Lecture Notes in Computer Sience, Springer Berlin, Heidelberg, Vol. ٤٢٤٧.

[۱۴] S. Young, I. Arel, T. Karnowski, and D. Rose, “ ***A Fast and Stable Incremental Clustering Algorithm***”, in Seventh International Conference Information Technology. IEEE, PP. ٢٠٤- ٢٠٩, ٢٠١٠.

کنفرانس داده کاوی ایران