IDMC'07
20-21 Nov.2007
Amir Kabir University

روشی مبتنی بر اتوماتای سلولی برای خوشه‌بندی داده‌ها

حسین مرشدلو محمدرضا میبدی

چکیده

امروزه با توجه به حجم بسیار بالا و نیز روند رو به رشد میزان داده‌هایی که بشر با آن سرو کار دارد، نیاز به روشهایی همچون داده‌کاوی برای تحلیل این حجم بزرگ از اطلاعات احساس می‌گردد. از طرفی با توجه به نرخ بسیار بالای تولید داده‌ها و اطلاعات و نیاز به تحلیل سریع آنها احتیاج به روشهاي داده‌کاوی داریم که بتوانند با سرعت بالایی داده‌ها را تحلیل نمایند. یکی از معمول‌ترین روشهای افزایش سرعت، موازی سازی اعمال می‌باشد. با توجه کاربرد روزافزون آتوماتای سلولی در کاربردهای مختلف و نیز توان موازی‌سازی بسیار بالای آن، بنظر می‌رسد استفاده از آتوماتای سلولی برای داده‌کاوی زمینه‌ای بسیار جالب و قابل توجه باشد. در این مقاله روشی برای خوشه‌بندی، که یکی از مهم ترین شاخه‌های داده‌کاوی می‌باشد، ارائه شده است که در آن از مفاهیم آتوماتای سلولی و کلونی مورچه‌ها استفاده گردیده است. در روش ارائه شده در این مقاله هر داده متعلق به مجموعه داده‌ای که باید خوشه-بندی گردد، به یک مورچه منتسب می‌گردد و این مورچه‌ها هستند که بطور همزمان در فضای آتوماتای سلولی جابجا شده و سعی می‌کنند تا با مورچه‌هایی که دارای داده‌های مشابه می‌باشند، اجتماع تشکیل دهند و بدین طریق عمل خوشه‌بندی را انجام دهند. آزمایشات انجام شده برای تست روش پیشنهاد شده، توانایی و قابلیت بالای این روش را برای خوشه‌بندی داده‌ها نشان می‌دهد.

کلمات کلیدی: اتوماتای سلولی، خوشه‌بندی، کلنی مورچه ها

A Cellular Automata based Data Clustering Method

H. Morshedlou M. R. Meybodi
Soft Computing Lab.
Computer Engineering and Information Technology Department
Amirkabir University of Technology, Tehran, Iran

Abstract

Nowadays, regarding huge data volumes and increasing rate of data generation, the need for data analysis methods like data mining is increased. Furthermore, with the real time requirement and size of data that should be processed, these methods must be able to act very fast. One of the well-known techniques to increase speed is parallelism. With respect to increasing use of cellular automata in various domains and its high power in parallelism, it seems that using cellular automata for data mining is an appealing approach. This paper presents a clustering method named CAC, which is based on the cellular automata and ant-colony concepts. In this method, each data point from the dataset that should be clustered is assigned to an ant. The ants move in grid space simultaneously and find other ants with similar data to form a community. Clustering operation is done by constitution of these communities. Experiments conducted on standard datasets using this method show the capability and applicability of our approach for data clustering.

Keywords: **Cellular Automata, Clustering, Ants Colony**

۱. مقدمه

فرایند خوشه‌بندی سعی دارد که یک مجموعه داده را به چندین خوشه تقسیم نماید بطوریکه داده‌های قرار گرفته در یک خوشه با یکدیگر شبیه بوده و با داده‌های خوشه‌های دیگر متفاوت باشند. در حال حاضر روشهای متعددی برای خوشه‌بندی داده‌ها وجود دارد که بر اساس نوع داده‌ها، شکل خوشه‌ها، فاصله داده‌ها و غیره عمل خوشه‌بندی را انجام می‌دهند. یک دسته از انواع روشهای موجود برای خوشه-بندی، روشهای مبتنی بر کلونی مورچه ها می‌باشند. ایده این روشها ساده می‌باشد. در این روشها مورچه‌ها توسط یکسری از عاملها مدل می-گردند که بصورت تصادفی در محیط خود حرکت می‌کنند. این محیط برای مثال می‌تواند یک شبکه مربعی شکل دارای شرایط مرزی باشد. داده‌هایی که باید خوشه‌بندی گردند، در این محیط پراکنده هستند و عاملها یا مورچه‌ها قادرند این داده‌ها را برداشته، از نقطه‌ای به نقطه دیگر در محیط مذکور جابجا سازند و در نقطه مقصد بر روی زمین قرار دهند. در واقع عمل برداشتن و قرار دادن داده‌ها بر روی زمین بر اساس معیارهای شباهت و چگالی داده موردنظر در همسایگی ناحیه‌ای که در آن قرار دارد یا قرار داده می‌شود، تعیین می‌گردد[۵،۷]. مورچه-ها در این روشها بیشتر تمایل دارند، داده‌هایی را جابجا سازند که آن داده یا از داده‌های همسان خود دور می‌باشد و یا توسط داده‌های غیرهمسان احاطه شده است. در واقع بدین طریق عمل خوشه‌بندی داده‌ها انجام می‌گردد. یکی از مزایای این روش این است که می‌توان داده‌های چندبعدی را به کمک یک فضای دوبعدی خوشه‌بندی کرد[۲]. در زمینه خوشه‌بندی مبتنی بر کلونی مورچه‌ها کارهای متعددی انجام شده است که برای نمونه می‌توان به موارد [۳،٤،٦،۸] اشاره نمود. روشهای قدیمی خوشه‌بندی با استفاده از کلونی مورچه‌ها مانند [٦،۸] داده‌هایی را که باید خوشه‌بندی گردند، بعنوان اشیائی در نظر می‌گیرند که باید توسط مورچه‌ها حمل شوند و به نقاط مختلف منتقل گردند. اشکال اصلی این روشها این است که تمام داده‌ها نمی‌توانند همزمان حرکت کنند و لذا زمان بیشتری برای خوشه‌بندی مورد نیاز خواهد بود. در این مقاله نیز روشی مبتنی بر اتوماتای سلولی بنام CAC (Cellular Automata based Clustering) برای خوشه‌بندی داده‌ها ارائه می-گردد که در آن از مفاهیمی مشابه مفاهیم روشهای مبتنی بر کلونی مورچه‌ها برای خوشه‌بندی، استفاده شده است. در این روش با تعریف مناسب مجموعه‌ای از قواعد و حالتهای مناسب برای اتوماتای سلولی حرکت و جابجایی مجموعه‌ای از داده‌ها که در فضای سلولها پخش شده-اند به طریقی انجام می‌گیرد که داده‌های مشابه در سلولهای مجاور هم مجتمع شوند و داده‌های هر دو اجتماع از اجتماعات تشکیل شده با هم متفاوت باشند.

در زمینه کارهای انجام شده در زمینه خوشه‌بندی مبتنی برکلونی مورچه‌ها با استفاده از آتوماتای سلولی می‌توان به موارد [۱،۲] اشاره کرد. در [۲] هر مورچه بعنوان یک شی داده‌ای فرض می‌شود که در یک سلول از سلولهای آتوماتای سلولی قرار گرفته است و می‌تواند در این محیط حرکت نماید. این مورچه با توجه به وضعیت فعلی خود و تناسب با اطرافیانش برای عمل بعدی (حرکت یا سکون) تصمیم می‌گیرد. روش ارائه شده در [۲] با روشی که ما در این مقاله پیشنهاد داده‌ایم دارای تفاوتهایی می‌باشد. برای مثال در [۲] از یک معیار مشابهت گسسته دوحالتی برای تعیین مشابه بودن یا نبودن استفاده می‌شود. این موضوع ممکن است سبب شود مورچه بین دو شی که تفاوت کمی با هم دارند و میزان شباهت آنها با مورچه نزدیک به مقدار آستانه‌ای است، تفاوت زیادی قائل شود. به این جهت ما در روش پیشنهادی خود (CAC) از یک معیار پیوسته استفاده می‌کنیم. همچنین تاثیرپذیری تناسب یک مورچه از وضعیت تناسب همسایگان خود از دیگر موارد دیگری است که سبب بهبود سرعت و نیاز به تکرار کمتر در روش CAC گردیده است. از موارد مشابهت روش ارائه شده در [۲] با روش CAC که در این مقاله پیشنهاد شده است می‌توان به استفاده از فرمون برای بهبود کارایی و کاهش تکرارهای لازم برای خوشه‌بندی اشاره کرد. البته قابل ذکر است که فرمون گذاری در روش CAC با فرمون گذاری در [۲] کمی تفاوت دارد که این تفاوت نیز در عمل و آزمایشات انجام شده، باعث بهبود دقت حرکت مورچه و سریعتر شدن عمل ردیابی مورچه‌های مشابه گردیده است.

در روش ارائه شده در [۱] نیز به ازای هر شیء یک مورچه در نظر گرفته شده است که وظیفه جابجایی آن شیء را بر عهده دارد و یک مدل رفتاری بنام ASM نیز برای مورچه‌ها تعریف شده است و روش پیشنهادی برای خوشه‌بندی بر روی این مدل رفتاری ارائه شده است. البته قابل ذکر است که روش ارائه شده در [۱] استفاده چندانی از مفاهیم روشهای مبتنی بر کلونی مورچه‌ها نکرده است. رویکرد ارائه شده در [۱] با روش پیشنهادی ما تفاوتهای عمده‌ای دارد. برای مثال از جمله مهمترین آنها می‌توان به مواردی همچون نحوه محاسبه شباهت دو مورچه، نحوه محاسبه تناسب یک مورچه، استفاده از روش فرمون‌گذاری و غیره اشاره کرد. نحوه تعیین مشابهت دو مورچه در روش CAC بگونه‌ای است که نسبت به نویز در داده‌ها حساسیت بسیار کمتری داشته و کارایی بسیار بهتری بخصوص در مورد خوشه‌بندی داده‌های نویزدار دارد. همچنین در [۱] میزان تناسب یک مورچه تنها به همسایگانش بستگی دارد اما تناسب یک مورچه در روش CAC از موارد متعددی همچون همسایگان، تناسب آنها، فرمون‌های موجود در سلول‌های همسایه و حالت فعلی خود مورچه تاثیر می‌گیرد. تفاوت مهم دیگر روش CAC با روش ارائه شده در [۱]، حرکت هدفمند در هنگام جابجایی مورچه می‌باشد. بعبارتی مشکل اصلی دیگری که در روش ارائه شده در [۱] وجود دارد، این است که هنگامی که مورچه‌ای تصمیم به جابجایی می‌گیرد، انتخاب سلول بعدی چندان هدفمند نیست که این موضوع سبب طولانی شدن زمان کلی خوشه‌بندی می‌گردد. در روش CAC با استفاده از فرمون‌گذاری این مشکل تا حد زیادی برطرف شده است.

آزمایشات انجام شده بر روی داده‌های استاندارد و نتایج بدست آمده، بیانگر کارایی روش CAC برای خوشه‌بندی داده‌ها می‌باشد. در ادامه مقاله، ابتدا در بخش ۲ تعریفی از آتوماتای سلولی ارائه خواهیم کرد و در بخش ۳ به بیان جزئیات و مفاهیم بکار رفته در روش پیشنهادی خواهیم پرداخت و راه‌حل پیشنهادی را در چارچوب این تعاریف بیان خواهیم کرد. در بخش ٤ نیز نتایج حاصل از اعمال این روش بر روی مجموعه داده‌های واقعی ارائه خواهد شد و در نهایت در بخش ٥ نیز جمع بندی را ارائه خواهیم کرد.

۲. اتوماتای سلولی

اتوماتاي سلولي يك مدل رياضي براي سيستم‌هايي است كه از اجزا و مولفه‌هاي ساده تشكيل شده، پويا هستند، با مرور زمان تغيير مي‌كنند و بر اساس ارتباط محلي بين اجزاي خود فعاليت مي‌كنند. اجزاي سيستم در مدل اتوماتاي سلولي بوسيله مجموعه‌اي منظم از سلول‌ها نمايش داده مي‌شوند كه در آن هر سلول معادل يكي از اجزاي سيستم است. در هر قدم زماني به هر سلول مقداري از مجموعه‌اي نامتناهي نسبت داده مي‌شود كه نشان دهنده وضعيت سلول است. اين مقداردهي بر اساس مقدار جاري سلول، سلول‌هاي همسايه آن و قانون تعريف شده براي اتوماتاي سلولي انجام مي‌شود؛ از آنجايي كه قانون در نظر گرفته شده تنها مقدار سلول و همسايه‌هاي آن را مورد استفاده قرار مي‌دهد، قانون محلي نيز ناميده مي‌شود[۱۰][۹] .

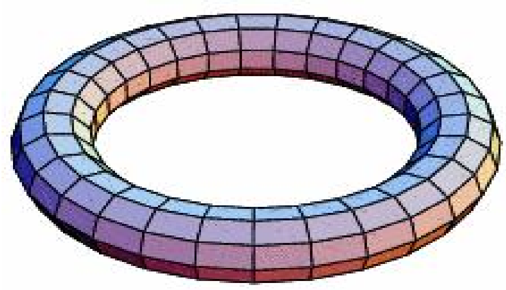
دسته‌بندي‌هاي مختلفي را مي‌توان براي اتوماتاي سلولي در نظر گرفت كه از جمله آنها مي‌توان به موارد زير اشاره كرد:

- اگر مجموعه مقادير سلول‌ها از دو عنصر تشكيل شده باشد، اتوماتا از نوع دودويي و در صورتي كه مقدار عناصر بيشتر از دو باشد اتوماتا از نوع چند مقداره خواهد بود.
- در صورتي كه قوانين شكلي قطعي داشته باشند اتوماتا از نوع قطعي و در صورتي كه احتمالي باشند اتوماتا از نوع احتمالي مي‌باشد.
- هرگاه تمامي سلول‌هاي اتوماتا همزمان مقداردهي شوند اتوماتا از نوع همگام، در مقابل اگر سلول‌ها با ترتيبي خاص و غير همزمان مقداردهي شوند اتوماتا از نوع ناهمگام خواهد بود.
- اگر سلول‌هاي اتوماتا به شكلي منظم در كنار يكديگر قرار گرفته باشند و تشكيل آرايه‌اي يك بعدي، دو بعدي و يا چند بعدي را بدهند، اتوماتا از نوع منظم و در صورتي كه سلول‌ها به شكلي نامنظم (مثلا به عنوان گره‌هاي يك گراف) قرار گرفته باشند اتوماتا از نوع نامنظم خواهد بود.

بطور رسمی یک اتوماتای سلولی را می‌توان با چهارتایی (S_d, Q, N, δ) نشان داد که S_d فضاي d بعدی سلولها می‌باشد، Q مجموعه‌ای محدود از حالات است که سلولها می‌توانند اختیار کنند، N همسایگی سلولها را مشخص می‌کند و δ مجموعه قوانین محلی می‌باشد که سلولها بر طبق این قوانین حالات خود را تغییر می‌دهند.

۳. روش پیشنهادی (CAC)

در روش CAC در ابتدا هرکدام از داده‌های موجود از مجموعه داده‌ای که باید عمل خوشه‌بندی بر روی آن انجام گیرد به یک مورچه تخصیص می‌یابد و هر مورچه در یکی از سلولهای اتوماتای سلولی بطور تصادفی قرار داده می‌شود. بازه زمانی‌ای که در آن داده‌ها باید خوشه‌بندی شوند، از تعدادی دور یا تکرار تشکیل شده است که در هر تکرار، هر مورچه در این اتوماتای سلولی در یکی از دو حالت فعال یا خواب قرار دارد. اگر سلولی که مورچه در آن قرار دارد، برای آن مورچه مناسب باشد یا به عبارت دیگر در اطراف آن مورچه، مورچه‌هایی با داده‌های مشابه قرار گرفته باشند، آنگاه مورچه به حالت خواب می‌رود و در هر تکراری که این تناسب از حد مشخصی کمتر گردید، مورچه دوباره فعال می‌گردد و سلول مناسب دیگری برای خود پیدا می‌کند. برای پیدا کردن این سلول مناسب مورچه از فرمون‌های بجا مانده از دیگر مورچه‌ها نیز بهره می‌گیرد. در واقع هر مورچه سعی می‌کند که در کنار مورچه‌هایی قرار گیرد که داده‌هایی مشابه داده وی دارند تا با یکدیگر اجتماع تشکیل دهند و بدین طریق در فضای اتوماتای سلولی، مجموعه‌ای از اجتماع‌های با داده‌های غیرمشابه هم، بوجود آورند. برای جلوگیری از تشکیل اجتماع‌هایی جدا از هم که دارای داده‌های مشابهی می‌باشند، احتمال جدا شدن یک مورچه از اجتماعی که دارای مورچه‌هایی با داده‌های مشابه می‌باشد، هیچگاه صفر نیست. در اینجا S_d یا فضایی که مورچه‌ها در آن پراکنده هستند، یک آرایه دوبعدی است که گوشه‌های بالا و پایین آن بهم و گوشه‌های راست و چپ آن نیز به یکدیگر متصل هستند و نمایی همچون شکل (۱) دارد. ما اندازه این شبکه مربعی را برابر $2\times\lfloor\sqrt{n}\rfloor$ در $2\times\lfloor\sqrt{n}\rfloor$ در نظر گرفته‌ایم که در آن n برابر تعداد داده‌ها می‌باشد. حالت هر سلول در این شبکه مربعی بصورت (i, C_i) نشان داده می‌شود که در آن i برچسب یا شناسه مورچه‌ای که در این سلول قرار می‌گیرد، می‌باشد و C_i نشان‌دهنده کلاس یا خوشه‌ای است که این مورچه به آن تعلق دارد. اگر i و C_i هر دو صفر باشند، به منزله خالی بودن آن سلول است.



شکل (۱) نمایی از شبکه مربعی مورد استفاده

همسایگی تعریف شده برای اتوماتای سلولی در این مساله بصورت زیر است:

$$N(ant_i) = \{\, x \bmod w(n), y \bmod h(n) \mid |x - x_i| \le 1, |y - y_i| \le 1 \,\} \qquad (۱)$$

مقدار $w(n)$ و $h(n)$ برابر با $\lfloor \sqrt{n} \rfloor \times 2$ در نظر گرفته شده‌اند و شعاع همسایگی برای اتوماتای سلولی برابر با یک در نظر گرفته شده است. در ادامه این بخش ابتدا چند مفهوم و رابطه را تعریف خواهیم کرد و سپس δ یا مجموعه قوانین محلی حاکم بر اتوماتای سلولی را شرح خواهیم داد.

فاصله: فرض کنید که داده‌های $data_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$ و $data_j = (z_{j1}, z_{j2}, \ldots, z_{jk})$ بترتیب به دو مورچه Ant_i و Ant_j انتساب داده شده‌اند. در اینصورت $dist_{ij}$ یا فاصله که بیانگر شباهت دو مورچه نیز می‌باشد، را بصورت زیر تعریف می‌کنیم:

$$dist_{ij} = \sqrt{\sum_{x=1}^{k} \left(\left(1 - \frac{|data_i(x) - data_j(x)|^2}{d_{ij}^{\,2}}\right) \times |data_i(x) - data_j(x)|^2 \right)} \qquad (۲)$$

که در رابطه بالا d_{ij} برابر است با فاصله اقلیدسی دو داده مذکور.

$$d_{ij} = d(ant_i, ant_j) = d(data_i, data_j) = \|data_i - data_j\| \qquad (۳)$$

توجه داشته باشید که عبارت $1 - \frac{|data_i(x) - data_j(x)|^2}{d_{ij}^{\,2}}$ در فرمول (۲) بر اساس اینکه $|data_i(x) - data_j(x)|^2$ تا چه حد بر فاصله اقلیدسی دو داده تاثیر دارد، تعیین می‌گردد و لذا میزان تاثیر $|data_i(x) - data_j(x)|^2$ را بر روی محاسبه فاصله موردنظر ما تعدیل می‌کند. این مورد بخصوص در مواردی که دو نمونه از یک مجموعه داده، که دارای چندین خصوصیت می‌باشند و تنها در یک یا دو خصوصیت به دلایلی همچون وجود نویز با هم تفاوت دارند، مفید است. از آنجا که در عمل خوشه‌بندی بدلیل عدم آگاهی از میزان بهره اطلاعاتی (Information Gain) هر خصوصیت، تخصیص وزن یکسان به همه خصوصیتها معقولانه‌ترین کار بحساب می‌آید، در مواردی هم که خصوصیات، وزن و تاثیر یکسانی در محاسبه فاصله اقلیدسی دو داده دارند، استفاده از روش فرمول (۲) برای محاسبه فاصله دو داده اشکالی بوجود نخواهد آورد.

تابع تناسب یا **fitness**: تابع تناسب یا fitness در واقع مشخص کننده این واقعیت است که یک مورچه قرار گرفته در محیط، تا چه حد با همسایگانش شباهت دارد و در واقع میزان این تناسب را بیان می‌کند. میزان بیشتر این مقدار نشانگر متناسب‌تر بودن وضعیت فعلی مورچه در بین همسایگانش است و باعث کم شدن احتمال حرکت و منتقل شدن مورچه از این سلول می‌گردد. برای نشان دادن تابع تناسب از فرمول (٤) استفاده می‌کنیم که در آن $f_{t+1}(ant_i)$ برابر با میزان تناسب مورچه i ام در تکرار $t+1$ ام با همسایگانش است. توجه کنید که عبارت $f^{\frac{1}{5}}{}_t(ant_j)$ در فرمول مذکور بیانگر وضعیت تناسب مورچه همسایه در تکرار t ام است. به عبارتی مناسب بودن وضعیت تناسب مورچه‌های همسایه‌ای که داده‌هایی شبیه به داده این مورچه دارند، تاثیر مثبت در وضعیت تناسب این مورچه و مناسب نبودن وضعیت آنها تاثیر منفی در وضعیت تناسب این مورچه می‌گذارد.

$$f_{t+1}(ant_i) = \min\{\ 1, \max\{\ 0, \frac{1}{8} \sum_{ant_j \in N(ant_i)} (m \times f^{\frac{1}{5}}{}_t(ant_j))^x \times (1 - \frac{dist_{ij}}{\alpha_i})\}\}$$

$$if\ (dist_{ij} > \gamma) \quad x = 0 \qquad (۴)$$

$$if\ (dist_{ij} < \gamma \quad and \quad f_t(ant_j) \le \beta) \quad x = 1, m = 1$$

$$if\ (dist_{ij} < \gamma \quad and \quad f_t(ant_j) > \beta) \quad x = 1,\ m = \frac{3}{2}$$

در لحظه شروع، f_t مربوط به تمام مورچه‌ها برابر یک می‌باشد. γ مقدار آستانه‌ای است که شباهت یا عدم شباهت دو مورچه را مشخص می‌کند و β نیز مقدار آستانه‌ای تناسب مورچه با محیط اطرافش می‌باشد، که مقادیر تناسب کمتر از این مقدار با احتمال بالایی باعث فعال شدن مورچه خواهند شد. مقدار α_i نیز در رابطه بالا از روی فرمول (۵) حساب می‌شود.

$$\alpha_i = \frac{1}{n-1} \sum_{j=1}^{n} dist_{ij} \qquad (۵)$$

احتمال فعال شدن یک مورچه: در اینجا از تابع $p_a(ant_i)$ برای نشان دادن احتمال فعال شدن مورچه در تکرار فعلی استفاده می‌کنیم. در واقع ما تابع مذکور را بصورت زیر تعریف می‌کنیم:

$$p_a(ant_i) = \min\{p, (\frac{\beta}{\beta + f_t} + \frac{f_t'}{f_t' + m \times f_t})\} \qquad (۶)$$

همانطور که در بالا دیده می‌شود، فرمول ذکر شده، احتمال فعال شدن مورچه را بر اساس دو معیار تعیین می‌کند. معیار اول میزان تناسب مورچه در رابطه با همسایگان فعلی‌اش می‌باشد که با قسمت $\frac{\beta}{\beta + f_t}$ بیان می‌گردد. همانطور که مشاهده می‌شود در صورتیکه مقدار تناسب f_t از مقدار آستانه‌ای β بسیار بزرگتر باشد، آنگاه قسمت $\frac{\beta}{\beta + f_t}$ در فرمول احتمال بالا نزدیک به صفر خواهد بود و احتمال فعال شدن مورچه از جانب این بخش بسیار کم است و اگر چنانچه مقدار تناسب f_t از مقدار آستانه‌ای β بسیار کوچکتر باشد (به معنی نامناسب بودن وضعیت فعلی مورچه)، آنگاه قسمت $\frac{\beta}{\beta + f_t}$ نزدیک به یک خواهد بود و احتمال فعال شدن مورچه از این طریق بسیار زیاد است.

بخش $\frac{f_t'}{f_t' + m \times f_t}$ نیز تاثیر فرمون‌های موجود در سلول‌هایی از همسایگی مورچه که خالی هستند و دارای فرمون متعلق به مورچه‌هایی بغیر از خود این مورچه و همسایگان فعلی‌اش می‌باشند را در احتمال فعال شدن مورچه اعمال می‌کند. f_t' در واقع میزان تناسب مورچه با مورچه‌هایی است که در سلول همسایه فرمون بجا گذاشته‌اند. مقدار f_t' از طریق فرمول (۷) حساب می‌شود. همانطور که در فرمول (٦) مشاهده می‌شود، با گذشت زمان بیشتر احتمال تاثیر فرمون کمتر می‌شود. همچنان هرقدر میزان f_t نسبت به f_t' بزرگتر باشد (در نتیجه مورچه در وضعیت مناسبتری قرار داشته باشد) تاثیر فرمونها (f_t') در فعال شدن مورچه مذکور کمتر خواهد بود.

$$f'(ant_i) = \max\{\ 0, \frac{1}{8} \sum_{ant_j \in Pheromone\ (ant_i)} \frac{1}{\sqrt{t_{current} - t_{ph_j}}} \times (1 - \frac{dist_{ij}}{\alpha_i})\} \qquad (۷)$$

همانطور که ذکر شد، حالت هر سلول در اتوماتای سلولی توسط (i, C_i) نشان داده می‌شود که اگر C_i و i هر دو صفر باشند، بمعنی خالی بودن آن سلول است و در غیراینصورت i نشانگر شناسه مورچه‌ای است که در آن سلول واقع شده است و C_i نیز نشان دهنده خوشه‌ای است که آن مورچه به آن تعلق دارد. در صورتیکه مورچه در حالت خواب باشد، برچسب C_i آن مشابه برچسب اکثریت همسایه‌ها است و اگر فعال باشد، برچسب برابر شناسه آن مورچه خواهد بود.

هر مورچه هنگام عبور از سلول‌های اتوماتای سلولی، فرمون یا آثاری از خود بجا می‌گذارد که شامل مواردی مانند شماره شناسه آن مورچه، مقادیر خصوصیات، جهت حرکت مورچه به سلول بعدی و زمانی که مورچه این سلول را ترک کرده است، می‌باشد. شماره شناسه برای این است که یک مورچه، فرمون مربوط به خود یا همسایگان فعلی‌اش را دنبال نکند. مقادیر خصوصیات برای این استفاده می‌شوند که مورچه‌های دیگر بتوانند میزان تفاوت خود را با مورچه‌ای که فرمون را بجا گذاشته است، اندازه بگیرند و زمان برای این بکار می‌رود که مورچه‌های دیگر بدانند چه مدت زمانی از لحظه ترک این سلول توسط مورچه مورد نظر گذشته است. عبارت $\frac{1}{\sqrt{t_{Current} - t_{ph_j}}}$ در فرمول (۷) در واقع بنوعی تاثیر تبخیر را در محاسبه تناسب موردنظر اعمال می‌کند. وجود فرمون از دو جنبه بر رفتار مورچه تاثیر دارد. اول اینکه وقتی مورچه فعال گردید، هنگام انتخاب مکان بعدی برای انتقال، سلول‌هایی را انتخاب می‌کند که احتمالا به مورچه‌های شبیه به خودش نزدیکتر است و همچنین مزیت دیگری که دارد این است که مورچه بدنبال مورچه‌های غیرمشابه نمی‌رود و در نتیجه بطور کلی این باعث می‌گردد که برای عمل خوشه‌بندی با دقت معینی، تعداد دور یا تکرار کمتری لازم باشد.

هر مورچه که وضعیت تناسب خوبی نداشته باشد، بعد از فعال شدن، در بین سلول‌هایی از همسایگی که در آنها مورچه‌ای وجود ندارد، سعی می‌کند بهترین سلول را برای انتقال پیدا کند. طریقه پیدا کردن بهترین سلول همسایه برای انتقال به این صورت است که مورچه تمام سلول‌های خالی اطراف را در نظر می‌گیرد. اگر سلول‌های مذکور دارای فرمون باشند مورچه از روی فرمون آنها مقدار فرمول (۸) را برای هر یک از سلولهای خالی محاسبه می‌کند.

$$v = (t_{current} - t_{ph_j})^{\frac{1}{3}} \times \sqrt{\sum_{x=1}^{k} \left(1 - \frac{|data_i(x) - data_j(x)|^2}{d_{ij}^{\,2}}\right) \times |data_i(x) - data_j(x)|^2} \qquad (۸)$$

اگر سلولهای همسایه‌ها فرمون نداشته باشند، یا فرمون خود مورچه یا فرمون همسایگانش در آنها وجود داشته باشد، مورچه فرض می‌کند که مقداری که توسط فرمول (۸) برای سلول‌های دارای فرمون محاسبه می‌شود، برای این سلول‌ها برابر با α_i یا میانگین فاصله این مورچه از بقیه مورچه‌ها باشد. مقدار α_i را می‌توان از روی فرمول (۵) بدست آورد. در نهایت نیز کمترین مقدار بدست آمده از بین این مقادیر برای سلول‌های خالی در همسایگی را پیدا کرده و با احتمال p $(p \geq 0.90)$، سلول متناظر را برای حرکت بعدی انتخاب و با احتمال $(1-p)$ نیز یکی از سلول‌های خالی اطراف بصورت تصادفی انتخاب می‌شود. پس از انتخاب سلول مناسب مورچه ابتدا فرمونی در سلول فعلی ثبت می‌کند که حاوی اطلاعاتی است که قبلا بیان گردید و سپس به سلول جدید منتقل می‌گردد و تمام متغیرهای ذکر شده بر این اساس بروز می‌شوند.

٤. نتایج آزمایشات

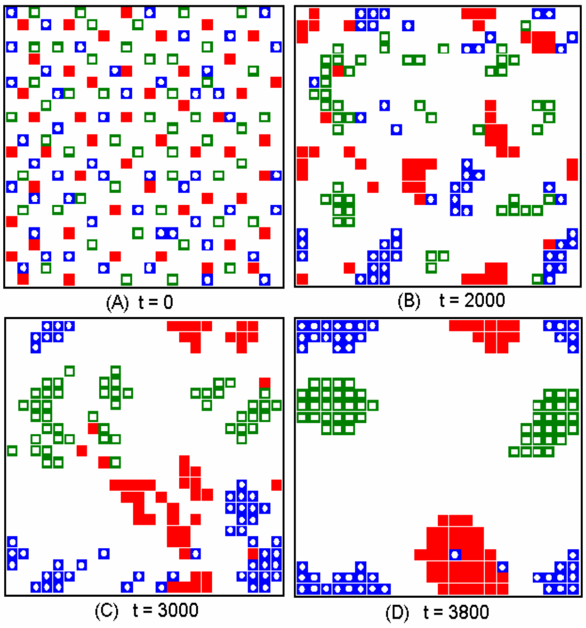
در این بخش نتایج حاصل از انجام آزمایشات را بر روی دو مجموعه داده واقعیIris و Glass بیان خواهیم نمود، اما قبل از آن اجازه دهید دو مجموعه داده مذکور را معرفی نماییم. هر دوی این مجموعه داده‌ها در [۱۱] در دسترس می‌باشند.

Iris: مجموعه داده Iris حاوی ۱۵۰ نمونه یا رکورد می‌باشد که هر نمونه یا رکورد در آن دارای چهار خصوصیت است. تعداد خوشه‌های موجود در این مجموعه داده سه عدد است و هر کدام از این خوشه‌ها یا کلاسها حاوی پنجاه نمونه یا رکورد است که برای هر نمونه یا رکورد، خصوصیت class مشخص می‌کند که آن نمونه به چه کلاسی تعلق دارد.

Glass: مجموعه داده Glass حاوی ۲۱٤ نمونه یا رکورد می‌باشد که هر نمونه یا رکورد در آن دارای ده خصوصیت (با احتساب شماره *id*) می‌باشد. تعداد خوشه‌های موجود در این مجموعه داده هفت عدد می‌باشد. برای هر نمونه یا رکورد، خصوصیت class مشخص می‌کند که آن نمونه یا رکورد به چه کلاسی تعلق دارد.

ما برای تست روش ارائه شده، دو مجموعه داده ذکر شده را توسط این روش خوشه‌بندی نمودیم. در شکل ۲ نمایی از اجرای این روش را بر روی داده‌های Iris در تکرارهای مختلف مشاهده می‌کنید. نتایج حاصل از اجرای آزمایش بر روی این دو مجموعه داده را با نتایج روشهای مشابه A4C و SA4C در [۱] مقایسه نمودیم. جدول (۱) نتایج حاصل از اعمال این روش را بر روی مجموعه داده‌های Iris و Glass نشان می‌دهد. در این جداول نتایج حاصل از اعمال این روش با دو تابع فاصله مختلف نشان داده شده است. شکلهای ۳ و ٤ نیز نتایج بدست آمده را بر روی نمودار نمایش می‌دهند. در حالتی که تابع فاصله، اقلیدسی است، فاصله اقلیدسی داده‌ها بعنوان معیار شباهت و نزدیکی داده‌ها مورد استفاده قرار گرفته است و برای محاسبه میزان تناسب نیز از فاصله اقلیدسی استفاده شده است. در روش CAC عادی، از فرمول (۲) بعنوان تابعی برای مشخص کردن میزان تفاوت و شباهت داده‌ها استفاده گردیده است. آزمایشات زیر با مقدار پارامترهای

$p=0.95$, $m = 2$ و $\beta = 0.2$ انجام گرفته است. توجه کنید که نتایج جداول، حاصل از ۱۰۰ بار اجرای الگوریتم می‌باشد. برای نمونه "میانگین خطا" میانگین تعداد خطاها در ۱۰۰ بار اجرای روش مذکور بوده است.



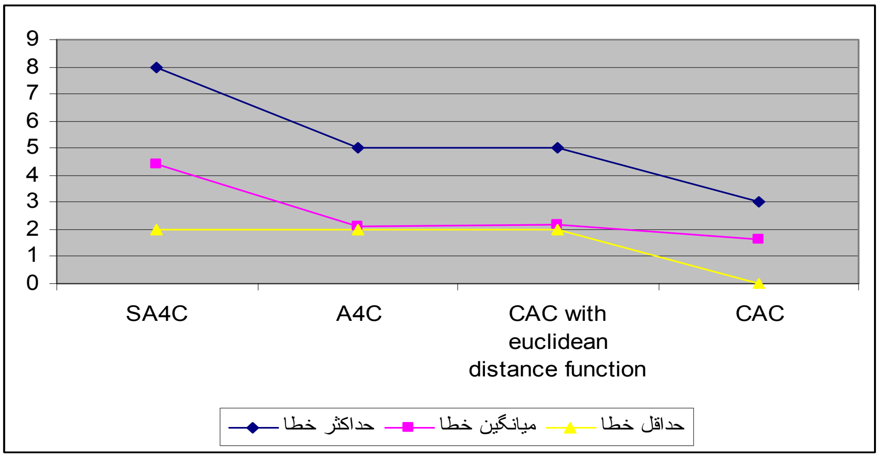
شکل(۲) نمایی از اجرای روش **CAC** بر روی **Iris**

۵. جمع‌بندی

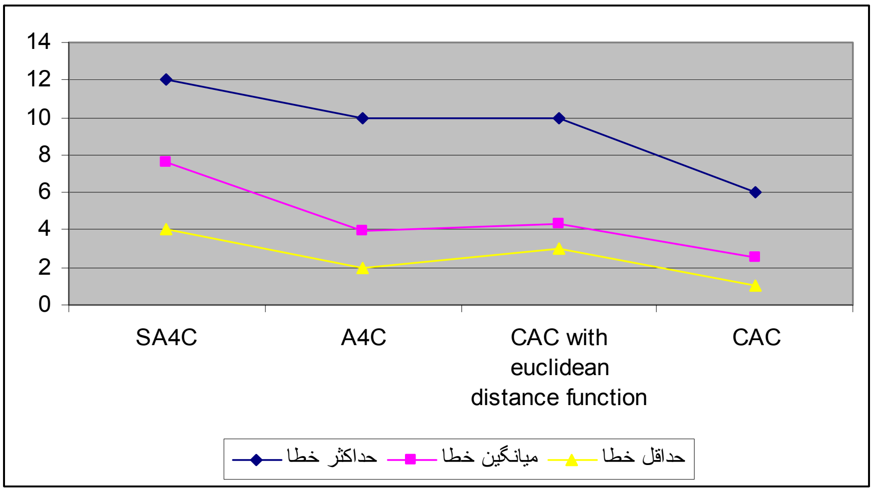
در این مقاله روشی بنام CAC برای خوشه‌بندی مبتنی بر آتوماتای سلولی ارائه شد که در آن با تخصیص هر داده به یک مورچه و تشکیل اجتماع‌هایی از مورچه‌های مشابه، عمل خوشه‌بندی انجام می‌گیرد. استفاده از یک روش تعیین فاصله و شباهت غیرحساس به نویز در روش CAC باعث گردیده که روش CAC نسبت به روشهای مشابه خود بهتر عمل نماید که نتایج آزمایشات انجام شده بر روی مجموعه داده‌های استاندارد با استفاده از این روش تعیین فاصله و بدون آن، کارایی این روش را برای مجموعه داده‌هاي واقعی تایید می‌کند. همچنین استفاده از مفهوم فرمون، تاثیرپذیری وضعیت تناسب مورچه‌ها از وضعیت تناسب همسایگان خود و جابجایی‌های هدفمند در روش CAC باعث گردیده تا مورچه‌ها در ردیابی مورچه‌های شبیه خود بهتر و دقیقتر عمل نمایند و در نتیجه تعداد تکرارهای لازم برای رسیدن عمل خوشه‌بندی به دقت خاصی کاهش پیدا کند. نتایج حاصل از اعمال روش CAC بر روی مجموعه داده‌های استاندارد، نشان می‌دهد که این روش از کارایی مناسب و خوبی برای خوشه‌بندی داده‌ها برخوردار است.

جدول (۱) نتایج حاصل از اعمال روشهای مختلف بر روی داده‌های **Iris** و **Glass**

	Iris				Glass			
روشها	SA4C	A4C	روش CAC با تابع فاصله اقلیدسی	روش CAC	SA4C	A4C	روش CAC با تابع فاصله اقلیدسی	روش CAC
تعداد تکرار	۵۰۰۰	۵۰۰۰	۴۰۰۰	۴۰۰۰	٥۰۰۰	٥۰۰۰	٤۰۰۰	٤۰۰۰
دفعات اجرا	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
حداقل خطا	۲	۲	۲	۰	٤	۲	۳	۱
حداکثر خطا	۸	۵	۵	۳	۱۲	۱۰	۱۰	٦
میانگین خطا	۴/۳۹	۲/۱۳	۲/۱۸	۱/۶۲	۷/۵۹	۳/۹٤	٤/۳۲	۲/٥۱



شکل (۳) نتایج اجرای روشهای مختلف برروی **Iris**



شکل (۴) نتایج اجرای روشهای مختلف برروی **Glass**

٦. مراجع

[1] Chen, L., Xu, X., Chen, Y. and He, P., "*A Novel Ant Clustering Algorithm Based on Cellular Automata*", Intelligent Agent Technology (IAT 2004), 2004, pp.148-154.
[2] Vande Moere, A., James Clayden, J., "*Cellular Ants: Combining Ant-Based Clustering with Cellular Automata*", Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence table of contents, Pages: 177 - 184, 2005.
[3] Ramos, V. and Abraham, A., "*Evolving a Stigmergic Self-Organized Data-Mining*", Conf. on Intelligent Systems, Design and Applications (ISDA-04), 2004, pp.725-730.
[4] Schockaert, S., De Cock, M., Cornelis, C. and Kerre, E.E., "*Fuzzy Ant Based Clustering*", Lecture Notes in Computer Science, 3172. Pp.342-349.
[5] Handl, J., Knowles, J., and Dorigo, M. (2003). "*On the performance of ant-based clustering*". In Design and Application of Hybrid Intelligent Systems, Vol. 104 of Frontiers in Artificial Intelligence and Applications (pp. 204-213). Amsterdam, the Netherlands; IOS Press.
[6] Lumer, E.D. and Faieta, B., "*Diversity and Adaptation in Populations of Clustering Ants*", From Animals to Animates: Conference on Simulation of Adaptative Behaviour, 1994, pp.501-508.
[7] Handl, J. and Meyer, B. "*Improved ant-based clustering and sorting in a document retrieval interface*", PPSN VII, LNCS 2439, 2002.
[8] Kuntz, A., Layzell, P. and Snyers, D., "*A Colony of Ant-Like Agents for Partitioning in VLSI Technology*", European Conf.e on Articial Life, 1997, MIT, pp.417-424.
[9] Wolfram, S., *Universality and complexity in cellular automata*, Physica D, no. 10, pp. 1-35, January 1984.
[10] Wolfram, S., *Cellular Automata*, Los Alamos Science, vol. 9, pp. 2-21, Fall 1983.
[11] http://kdd.ics.uci.edu