یک روش ترکیبی مبتنی بر شبکه ایمنی مصنوعی و اتوماتای یادگیر برای خوشه بندی داده ها

بابک نصیری[۱]؛ محمدرضا میبدی[۲]

چکیده

خوشه بندی یکی از وظایف اصلی در داده کاوی بشمار می رود. وقتی تعداد نمونه ها و ابعاد داده ها بسیار باشند، استفاده از اتوماتای یادگیر بمنظور خوشه بندی، بسیار زمانگیر و پر هزینه خواهد بود. از این رو در این مقاله یک رهیافت دو مرحله ای برای خوشه بندی داده ها مبتنی بر شبکه ایمنی مصنوعی و اتوماتای یادگیر پیشنهاد شده است. در ابتدا با استفاده از شبکه ایمنی مصنوعی، حجم داده های مورد آنالیز کاهش می یابد. کاهش حجم داده ها بصورت سطری با کم کردن نمونه ها در مجموعه داده صورت می گیرد. شبکه ایمنی مصنوعی، نمونه هایی که می بایست در مجموعه داده باقی بمانند را انتخاب می کند. سپس در مرحله بعد، با استفاده از اتوماتای یادگیر تطبیق پذیر، خوشه بندی داده ها بصورت پویا به روش جدیدی انجام می شود. نتایج بدست آمده بر روی مجموعه داده های مختلف در مقایسه با نتایج حاصل از سه روش خوشه بندی DBSCAN، EM و K-MEANS حکایت از قابل مقایسه بودن روش پیشنهادی در مقایسه با سایر روشها دارد.

کلمات کلیدی

خوشه بندی، سیستم ایمنی مصنوعی، اتوماتای یادگیر، شبکه ایمنی مصنوعی، داده کاوی.

An hybrid Approach based Artificial Immune Network and Learning Automata for Data Clustering

Babak Nasiri[۱]; Mohammad reza Meybodi[۲]
[۱] Computer Engineering and Information Technology Department, Azad Islamic University, Qazvin, Iran
[۲] Computer Engineering and Information Technology Department, Amirkabir University of Technology ,Tehran, Iran

Abstract

Clustering is considered one of the main tasks in data mining. When the number of samples and dimensions are very high, using learning automata for Clustering would be very time-consuming and costly. So this paper has suggested a two-step approach for Clustering data based artificial immune network and learning automata. First, using the artificial immune network, data size decreases. Data reduction is performed horizontally with reducing samples in data set. Artificial immune network select samples that should be remaining in the data set. Then the next stage, using the adaptive learning automata, Data Clustering is performed dynamically with the new method. Results on different data set shows that the proposed approach is comparable with other clustering methods such as DBSCAN, EM and K-MEANS.

Keywords

Clustering, Artificial Immune System, Learning Automata, Artificial Immune Network, Data Mining.

۱. مقدمه

در سال های اخیر داده کاوی در صنعت و تحقیقات آکادمیک بسیار مورد توجه قرار گرفته است و حجم داده عظیمی برای تبدیل شدن به اطلاعات و دانش های مفید مورد جمع آوری و آماده سازی قرار گرفته اند. اطلاعات و دانش بدست آمده از فرایند داده کاوی می تواند در

۱ دانشکده برق، کامپیوتر و فناوری اطلاعات،دانشگاه آزاد اسلامی، قزوین، ایران، nasiri_babak@yahoo.com
۲ دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه امیر کبیر، تهران، ایران، mmeybodi@aut.ac.ir

کاربردهای مختلفی نظیر مدیریت تجارت، پیش بینی ، آنالیز بازار و اکتشاف دانش مورد استفاده قرار گیرد. مهمترین وظایف داده کاوی را همانا می توان طبقه بندی، خوشه بندی، پیش بینی، کاوش قوانین پیوندی و تشخیص آنومالی نامید. خوشه بندی یعنی, تقسیم داده ها به گروه هایی از اشیا مشابه. هدف از خوشه بندی، تقسیم داده ها به K قسمت است بنحوی که عناصر هر قسمت، بیشترین شباهت را به هم داشته و با عناصر خارج از آن بسیار متفاوت باشند. خوشه بندی نه تنها بعنوان یک وظیفه در داده کاوی مطرح است بلکه بعنوان یکی از مراحل پیش پردازش داده ها برای فرایند اکتشاف دانش و داده کاوی نیز مورد استفاده قرار می گیرد.

تاکنون روش های مختلفی برای خوشه بندی داده ها توسعه داده شده اند که می توان آنها را به دسته های زیر تقسیم نمود : خوشه بندی *جزء بندی*[1]*، خوشه بندی سلسله مراتبی*[2]*، خوشه بندی مبتنی بر تراکم*[3] ، *خوشه بندی مبتنی بر توری*[4] و ...[۱] . استفاده از روش های هوشمند برای خوشه بندی داده ها نیز در سالهای اخیر بسیار مورد توجه قرار گرفته است که از بین آنها می توان به خوشه بندی مبتنی بر شبکه عصبی مصنوعی، الگوریتم ژنتیک [۲]، الگوریتم مورچگان [۳]، جستجوی Tabu [۴] و بعضی الگوریتم های ترکیبی نظیر SOM و K-means [۵]، SOM و GA [۶] اشاره نمود. همچنین اخیرا روش هایی مبتنی بر اتوماتای سلولی[۷]، اتوماتای یادگیر [۸] و اتوماتای یادگیر سلولی [۹] برای خوشه بندی داده ها ارائه شده است.

هدف از خوشه بندی پاسخ دادن به دو سوال اصلی زیر است: چند خوشه در داده ها موجود می باشد و این خوشه ها در کجا قرار گرفته اند. اگر تعداد خوشه ها قبل از خوشه بندی مشخص شود، خوشه بندی را ایستا و در غیر اینصورت خوشه بندی را پویا می نامیم.

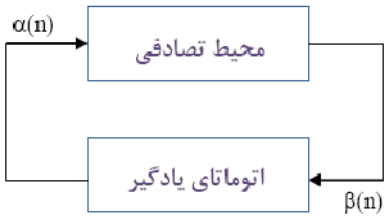
اکثر الگوریتم های ارائه شده خوشه بندی را بصورت ایستا انجام می دهند و زمانیکه با حجم انبوهی از داده ها رو به رو باشیم دچار مشکل می شوند. چنانچه بخواهیم از این الگوریتم ها برای خوشه بندی پویا بهره ببریم بسیار زمانگیر خواهند بود. در این مقاله یک رهیافت دو مرحله ای برای خوشه بندی داده ها بصورت پویا ارائه شده است. ابتدا از یک روش نمونه گیری مبتنی بر سیستم ایمنی مصنوعی (شبکه ایمنی مصنوعی) برای کاهش حجم داده ها استفاده شده است. برای این کار نمونه های موجود در مجموعه داده نقش آنتی ژن را بازی می کنند و با تولید یک مجموعه تصادفی از آنتی بادی ها کار شروع می شود. سپس در یک فرایند تکاملی آنتی بادی هایی که بیشترین شباهت را با آنتی ژن ها داشته باشند در حافظه نگهداری می شوند. این آنتی بادی ها در نهایت نمونه هایی هستند که برای خوشه بندی انتخاب شده اند.

در مرحله بعد پس از نمونه گیری داده ها و کاهش حجم آن، نوبت به خوشه بندی داده ها بصورت پویا با استفاده از اتوماتای یادگیر می رسد. برای این کار به هر نمونه یک اتوماتای یادگیر اختصاص داده می شود سپس با توجه به میزان شباهت بین نمونه ها (فاصله بین نمونه ها) و شعاع همسایگی – که از قبل مشخص شده - همسایه های هر نمونه مشخص می شود. هر همسایه حکم یک عمل برای هر نمونه را خواهد داشت. سپس با یک روش تکراری، بردارهای احتمال هر نمونه آموزش می یابند تا هر نمونه، همخوشه ای های خود را پیدا کند.

ادامه مقاله بصورت زیر سازماندهی شده است. در بخش دوم، اتوماتای یادگیر و در بخش سوم سیستم ایمنی مصنوعی و شبکه ایمنی مصنوعی به اختصار تشریح شده است. بخش چهارم به معرفی الگوریتم پیشنهادی اختصاص یافته است. در بخش پنجم نتایج حاصل از پیاده سازی روش پیشنهادی بر روی مجموعه داده های مختلف مورد ارزیابی قرار گرفته است و بخش آخر به نتیجه گیری و پیشنهادهایی برای بهبود تخصیص دارد.

۲. اتوماتای یادگیر

اتوماتای یادگیر یک مدل انتزاعی است که بطور تصادفی یک عمل از مجموعه متناهی اعمال خود را انتخاب کرده و بر محیط اعمال می کند. محیط، عمل انتخاب شده توسط اتوماتای یادگیر را ارزیابی کرده و نتیجه ارزیابی خود را توسط یک سیگنال تقویتی به اتوماتای یادگیر اطلاع می دهد. سپس اتوماتای یادگیر با اطلاع از عمل انتخاب شده و سیگنال تقویتی ، وضعیت داخلی خود را بروز کرده و عمل بعدی خود را انتخاب می کند. هدف نهایی این است که اتوماتا یاد بگیرد تا از بین اعمال خود بهترین عمل را انتخاب کند. بهترین عمل، عملی است که احتمال دریافت پاداش از محیط را به حداکثر برساند. کارکرد اتوماتای یادگیر در تعامل با محیط، درشکل (۱) مشاهده میشود.



شکل (۱): ارتباط بین اتوماتای یادگیر با محیط

محیط را می توان توسط سه تایی $E=\{\alpha,\beta,C\}$ نشان داد که در آن $\alpha=\{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ مجموعه ورودیها، $\beta=\{\beta_1, \beta_2, \ldots, \beta_m\}$ مجموعه خروجیها و $C=\{C_1, C_2, \ldots, C_r\}$ مجموعه احتمالات جریمه می باشد. هرگاه β مجموعه دو عضوی باشد محیط از نوع P می باشد. در چنین محیطی $\beta_1 = 1$ بعنوان پاسخ نامطلوب یا شکست، $\beta_2 = 0$ بعنوان پاسخ مطلوب یا موفقیت در نظر گرفته می شوند. در محیط از نوع Q، مجموعه β دارای تعداد متناهی عضو می باشد و در محیط از نوع S، تعداد اعضاء مجموعه β نامتناهی است. C_i نشان دهنده احتمال نامطلوب بودن سیگنال تقویتی محیط در پاسخ به عمل α_i می باشد. در یک محیط ایستا مقادیر C_i ها ثابت هستند، حال آنکه در یک محیط غیر ایستا این مقادیر در طی زمان تغییر می کنند. بر اساس اینکه تابع بروزرسانی وضعیت اتوماتای یادگیر (که با اطلاع از عمل انتخاب شده و سیگنال تقویت β، وضعیت بعدی اتوماتای یادگیر را محاسبه کند) ثابت یا متغیر باشد، اتوماتای یادگیر به دو دسته اتوماتای یادگیر با ساختار ثابت و اتوماتای یادگیر با ساختار متغیر تقسیم می گردند.

اتوماتای یادگیر با ساختار متغیر را میتوان توسط چهارتایی $\{\alpha,\beta,P,T\}$ نشان داد که در آن $\alpha=\{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ مجموعه اعمال اتوماتای یادگیر، $\beta=\{\beta_1, \beta_2, \ldots, \beta_m\}$ مجموعه ورودیهای اتوماتای یادگیر، $P = \{P_1, P_2, \ldots, P_r\}$ بردار احتمال انتخاب هریک از عملها و T، $P(k+1) = T[\alpha(k), \beta(k), P(k)]$ الگوریتم یادگیری اتوماتای یادگیر میباشد. الگوریتم های یادگیری متنوعی برای اتوماتای یادگیر ارائه شده است که در ادامه یک الگوریتم یادگیری خطی برای اتوماتای یادگیر بیان می گردد. فرض کنید اتوماتای یادگیر در مرحله nام اقدام α_i خود را انتخاب نموده و محیط ارزیابی خود را توسط سیگنال تقویتی $\beta(n)$ به اتوماتای یادگیر اعلام می کند. با استفاده از الگوریتم یادگیری خطی، اتوماتای یادگیر، بردار احتمال انتخاب اقدام های خود را مطابق رابطه (۱) تنظیم می کند.

$$
\begin{aligned}
P_i(n+1) &= P_i(n) + a.(1-\beta(n)).(1-P_i(n)) \\
&\quad - b.\beta(n).P_i(n) \\
P_j(n+1) &= P_j(n) + a.(1-\beta(n)).P_j(n) \\
&\quad + \frac{b.\beta(n)}{r-1} - b.\beta(n).P_i(n) \qquad \text{if } j \neq i
\end{aligned}
\tag{۱}
$$

که a پارامتر پاداش و b پارامتر جریمه می باشد. اگر a و b با هم برابر باشند، الگوریتم L_{R-P}[5]، اگر b از a خیلی کوچکتر باشد، الگوریتم $L_{R\in P}$[6] و اگر b صفر باشد، الگوریتم L_{R-I}[7] نام دارد[۱۰و۱۱].

۳. سیستم ایمنی مصنوعی

سیستم ایمنی مصنوعی الهام گرفته از سیستم ایمنی طبیعی (سیستم ایمنی موجود در بدن موجودات زنده) می باشد. سیستم ایمنی طبیعی وظیفه محافظت از بدن را در برابر موجودات خارجی نظیر باکتری ها، ویروس ها و ... برعهده دارد. برای این منظور می بایست ابتدا سلول های خودی از غیر خودی تشخیص[8] داده شود، سپس در برابر سلول های غیر خودی از خود واکنش نشان دهد. این سیستم همچنین دارای حافظه بوده و آنتی بادی هایی را که قبلا برای از بین بردن آنتی ژن ها بکار رفته اند را نگهداری می کند تا در صورت حمله مجدد از طرف این آنتی ژن ها سریعا به آنها پاسخ گوید. الگوریتم های موجود در سیستم ایمنی مصنوعی را می توان بطور کلی به چهار دسته مختلف تقسیم نمود. این الگوریتم-ها عبارتند از: انتخاب منفی[9]، انتخاب تولیدمثل[10]، تئوری شبکه‌های ایمنی[11] و تئوری خطر[12] که هر کدام بخشی از سیستم ایمنی طبیعی را مدل کرده‌اند. از این الگوریتم‌ها برای حل مسائل مختلفی نظیر بهینه‌سازی، دسته‌بندی، خوشه‌بندی، تشخیص نفوذ و ... استفاده شده است[۱۲].

۱.۳. شبکه ایمنی مصنوعی

تئوری شبکه‌های ایمنی اولین بار توسط Jerne در سال ۱۹۷۴ مطرح گردید و بصورت شبکه پیچیده‌ای از پاراتوپ‌ها[۱۳] که مجموعه‌ای از ایدیوتوپ‌ها[۱۴] را شناسایی می‌کنند و مجموعه ای از ایدیوتوپ‌ها که توسط پاراتوپ‌ها شناسایی می شوند، تعریف می‌شود. بنابراین هر یک از عناصر شبکه قابلیت شناسایی و شناخته‌شدن را به صورت همزمان دارا می‌باشند. این خاصیت منجر به ایجاد شبکه ایمنی می‌شود که در آن، چه مولکول-های آنتی‌بادی به صورت آزاد و چه به عنوان مولکول گیرنده B-cell[۱۵] به یکدیگر متصل می‌شوند. پس از اینکه یک آنتی‌بادی یک اپیتوپ یا یک ایدیوتوپ را شناسایی کرد، می‌تواند به صورت مثبت یا منفی به این سیگنال تشخیص پاسخ دهد. یک پاسخ مثبت منجر به تحریک سلول[۱۶]، تکثیر سلول[۱۷] و ترشح آنتی‌بادی[۱۸] می شود. درحالیکه یک پاسخ منفی منجر به تحمل[۱۹] یا بازدارندگی[۲۰] در سلول می‌شود[۱۳].

۴. رهیافت پیشنهادی

رهیافت پیشنهادی برای خوشه بندی داده ها بصورت پویا شامل دو مرحله می باشد. در مرحله اول کاهش حجم داده ها مبتنی بر شبکه ایمنی مصنوعی انجام می گیرد و در مرحله دوم یک خوشه بندی پویا مبتنی بر اتوماتای یادگیر ارائه شده است.

برای ارزیابی نتیجه خوشه بندی از شاخص دون، طبق رابطه (۱) استفاده شده است.

$$D = \min_{i=1\ldots n_c}\left\{\min_{j=i+1\ldots n_c}\left(\frac{d(c_i,c_j)}{\max_{k=1\ldots n_c}(diam(c_k))}\right)\right\} \qquad (۱)$$

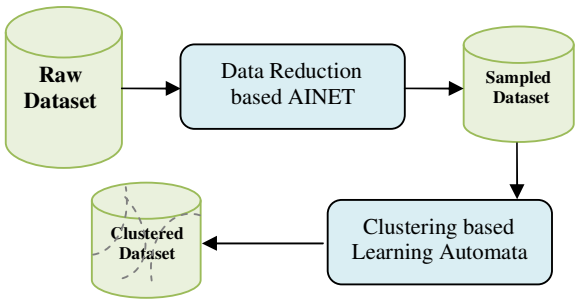
که C_i مرکز خوشه i ام، n_c تعداد خوشه ها می باشند. همچنین d(x,y) فاصله اقلیدسی بین دو شی ء داده ای و diam(c_i) قطر خوشه j ام می باشند که به ترتیب با استفاده از روابط (۲) و (۳) محاسبه می‌شوند.

$$d(c_i,c_j) = \min_{x\in c_i, y\in c_j}\{d(x,y)\} \qquad (۲)$$

$$diam(c_i) = \max_{x,y\in c_i}\{d(x,y)\} \qquad (۳)$$

در رابطه (۱)، هر چه مقدار D بزرگتر باشد خوشه بندی بهتری انجام شده است.

در شکل (۲) بلوک دیاگرام رهیافت پیشنهادی برای خوشه بندی داده ها نمایش داده شده است.



شکل (۲): بلوک دیاگرام رهیافت پیشنهادی برای خوشه بندی

در ادامه به تفصیل، هر یک از این مراحل تشریح شده است.

۴-۱- کاهش حجم داده ها مبتنی بر شبکه ایمنی مصنوعی

یکی از روشها برای افزایش کارایی آنالیز داده ها، کاهش حجم داده ها و نمونه برداری می باشد. این کار باید بصورتی انجام شود کـه از دسـت دادن دقت داده ها به حداقل خود برسد. الگوریتم های مختلفی برای این کار وجود دارد. در این مقاله از شبکه ایمنی مصـنوعی بـرای ایـن منظـور استفاده شده است. برای این کار نمونه های موجود در مجموعه داده نقش آنتی ژن ها را بازی می کنند و با تولید یـک مجموعـه تصـادفی از آنتـی بادی ها کار شروع می شود. سپس در یک فرایند تکاملی آنتی بادی هایی که بیشترین شباهت را با آنتی ژن ها داشته باشند تولیـد مثـل و جهـش پیدا می کنند و بهترین ها در هر تکرار در حافظه نگهداری می شوند. این کار تارسیدن به شرط پایان الگوریتم (تعداد تکرار از قبـل تعیـین شـده) ادامه می یابد. آنتی بادی ها در نهایت نمونه هایی هستند که برای خوشه بندی انتخاب شده اند. این الگوریتم بصورت زیر می باشد.

الگوریتم ۱: الگوریتم کاهش حجم داده ها مبتنی بر شبکه ایمنی مصنوعی

Algorithm ۱. Data Reduction based AINET

Input : *dataset **X**, immune suppression threshold **ts***

Output : *Reduced Dataset **R***

Step۱: Initialize network cells(antibodies) set ***Abs*** and let **M** ={ } ;

Step۲: For each object(**Ag**) in dataset **X** do:

۲.۱: Determine the affinity with each cell in ***Abs*** according to a distance metric;

۲.۲: Select the c highest affinity cells from ***Abs***;

۲.۳: Clone selected ***Abs*** proportional to its affinity and add into a temporary antibodies set, *tempM*;

۲.۴: Apply clone mutation to the *tmpM*;

۲.۵: apply immune suppression to *tmpM*;

۲.۶: add Ag to M in tail position;

۲.۷: add *tmpM* into M in tail position;

۲.۸: apply immune suppression to M;

Step۳:Test the stopping criterion, if it not stops, then let Abs = M ,M ={ }, and go to Step۲;

Otherwise let R = M and stop.

الگوریتم بالا در مقایسه با الگوریتم AINET کلاسیک دارای تفاوت هایی می باشد. از آن جمله می توان به اضافه شدن مرحله ۲.۶ برای اینکه هر آنتی ژن شانس ورود به شبکه ایمنی را در طول تکرارها داشته باشد و از احتمال از دست رفتن اطلاعات در فضاهای خلوت اجتناب شود، اشـاره نمود.

۴-۲- خوشه بندی داده ها بصورت پویا مبتنی بر اتوماتای یادگیر

پس از انجام نمونه گیری توسط Algorithm ۱، مجموعه داده ای از نمونه ها بدست می آید. حال به هر یک از نمونه های این مجموعـه یـک اتوماتای یادگیر اختصاص داده می شود و همسایگان هر نمونه با توجه به شعاع همسایگی rad (یکی از پارامترهای ورودی مسـئله) مشـخص شـده،

بعنوان عملهای آن نمونه انتخاب می شوند. انتخاب یک عمل توسط یک اتوماتای یادگیر به مفهوم انتخاب همخوشه ای برای آن اتوماتا می باشد. الگوریتم خوشه بندی با استفاده از اتوماتای یادگیر بصورت زیر می باشد.

الگوریتم ۲: الگوریتم خوشه بندی مبتنی بر آتوماتای یادگیر

Algorithm ۲. Clustering based Learning Automata

Input : *Reduced Dataset **R** , neighborhood radius **rad***

Output: *Cluster Number **Cnum**, Clustered Dataset **C***

Step۱: find_neighborhoods_of_each_data();

Step۲: initialize probability of actioni from $\mathbf{LA_j}$ as $\mathbf{d_{ij}}$ /$\mathbf{d_{ij}}$

Step۳: Do While **fitness** not changed for a period of time

۳.۱: Add data from Dataset **R** to **Queue** randomly.

۳.۲: Do While **Queue** is not EMPTY.

۳.۲.۱: delete an item from **Queue** as **Selected**.

۳.۲.۲: select an action for $\mathbf{LA_{Selected}}$ based probability as Action_j.

۳.۲.۳: disable Action_j for $\mathbf{LA_{Selected}}$ and also disable Action Selected for LA**j**.

۳.۲.۴: add **i** , **j** to **Path** and add **j** to **Queue**.

Step۴: [Cnum,C]=find_weakly_Connected_Graph(**Path**).

Step۵: fitness = calculate_Dunn_Index(**Cnum** , **C**).

Step۶: if it is the best fitness up to now.

۶.۱: reward Action_j of LA_i for each LA, Action in Path.

Step۷: else

۷.۱: penalize Action_j of LA_i for each LA, Action in Path.

۵– ارزیابی رهیافت پیشنهادی

در این بخش به بررسی و ارزیابی رهیافت پیشنهادی در مقایسه با سایر الگوریتم ها می پردازیم. برای اینکار از سه مجموعه داده که یکی از آنها بصورت آزمایشی و بقیه مجموعه داده های واقعی می باشند استفاده شده است. مجموعه داده اول (Sample) شامل ۵۰۰۰ داده دو بعدی که توسط توزیع نرمال $N(\mu,\sigma^2)$ تولید شده و متعلق به پنج خوشه (هر خوشه ۱۰۰۰ داده) با مشخصات زیر می باشد.

$$\{N(۰.۲,۰.۱^2),\quad \{N(۰.۲,۰.۱^2),\quad \{N(۰.۵,۰.۱^2),$$
$$N(۰.۲,۰.۱^2)\}\quad N(۰.۸,۰.۱^2)\}\quad N(۰.۵,۰.۱^2)\}$$

$$\{N(0.8,0.1^2), \quad \{N(0.8,0.1^2),$$
$$N(0.2,0.1^2)\} \quad N(0.8,0.1^2)\}$$

مجموعه داده دوم IRIS بوده که شامل ۱۵۰ داده چهار بعدی متعلق به سه خوشه ۵۰ داده ای می باشد. در این مجموعه داده خوشه اول بصورت خطی از دو خوشه دیگر جدا شده و دو خوشه دیگر بصورت غیر خطی از یکدیگر جدا می شوند. در نهایت مجموعه داده سوم، مجموعه داده Wine می باشد که شامل ۱۷۸ داده سیزده بعدی می باشد. این مجموعه داده شامل سه خوشه ۵۹، ۷۱ و ۴۸ داده ای بوده که خوشه دوم با خوشه اول و سوم داده های هم مرز دارد. بر روی مجموعه داده های ذکر شده در بالا، چهار الگوریتم مختلف اعمال شده است که عبارتند از روش ترکیبی شبکه ایمنی مصنوعی و اتوماتای یادگیر(AIN-LA)، روش ترکیبی شبکه ایمنی مصنوعی و اتوماتای یادگیر تطبیق پذیر(Adaptive AIN-LA)، DBScan و EM.

در روش ترکیبی شبکه ایمنی مصنوعی و اتوماتای یادگیر تطبیق پذیر، پارامترهای پاداش و جریمه در طول اجرا تنظیم می شوند. پارامترهای پاداش و جریمه نقش قدم ها را برای رسیدن به جواب بازی می کنند. به همین خاطر ابتدا قدم ها را بزرگ برداشته و سپس آن را در طول اجرا کوچک می کنیم.

نتایج حاصل از اعمال الگوریتم۱ بر روی مجموعه داده ها بمنظور کاهش حجم داده بصورت جدول (۱) می باشد. نتایج بدست آمده، حاصل از ۲۰ بار اجرای الگوریتم و میانگین گیری می باشد.

جدول (۱) : نتایج حاصل از اجرای الگوریتم ۱ بر روی مجموعه داده ها

مجموعه داده	تعداد نمونه ها	تعداد نمونه ها پس از اجرای Algorithm۱	آستانه بازدارندگی (ts)
Sample	۵۰۰۰	۲۳	۰.۱۴
IRIS	۱۵۰	۱۵	۰.۲
Wine	۱۷۸	۱۸	۰.۱

همانطور که در جدول بالا مشاهده می شود، حجم داده ها پس از اعمال الگوریتم ۱، فوق العاده کاهش می یابد. t_s آستانه بازدارندگی می باشد که چنانچه میزان شباهت بین دو داده کمتر از آن باشد، یکی از آن دو داده بعنوان نماینده باقی مانده و دیگری حذف می شود.

همچنین نتایج حاصل از اعمال الگوریتم ۲ بر روی مجموعه داده های ذکر شده در مقایسه با سایر روشها بصورت جدول (۲) می باشد.

جدول(۲): نتایج حاصل از مقایسه اجرای الگوریتمها.

	AIN-LA / Adaptive AIN-LA			DBScan		EM	
	تعداد خوشه	میزان درستی(٪)	تعداد دفعات تکرار	تعداد خوشه	میزان درستی(٪)	تعداد خوشه	میزان درستی(٪)
Sample	۵	۱۰۰	۷۵۲	۵	۹۵	۵	۱۰۰
	۵	۱۰۰	۶۴۳				
IRIS	۵	۸۶.۳	۳۴۰	۳	۸۷.۳	۵	۷۰
	۳	۸۴.۶	۲۳۲				
Wine	۳	۸۴.۲	۳۹۱	۲	۶۴	۳	۸۱.۴

				۲۳۶	۸۴.۲	۳	

همانطور که مشاهده می شود، در اکثر موارد رهیافت پیشنهادی در مقایسه با رهیافت های دیگر بهتر جواب داده است و همچنین استفاده از اتوماتای یادگیر تطبیق پذیر باعث تسریع در همگرایی و پاسخ دهی سریعتر شده است.

۶- نتیجه گیری

در این مقاله یک روش جدید برای خوشه بندی اطلاعات مبتنی بر شبکه ایمنی مصنوعی و اتوماتای یادگیر تطبیق پذیر مورد معرفی قرار گرفت که بخوبی از قابلیت های شبکه ایمنی مصنوعی برای کاهش حجم داده ها و از اتوماتای یادگیر برای جستجو در فضاهای بزرگ بهره گرفته شد. بمنظور ارزیابی، رهیافت خوشه بندی ارائه شده بر روی تعدادی مجموعه داده استاندارد آزمایش شد و نتایج حاصله (اتوماتای یادگیر تطبیق پذیر و معمولی) با نتایج حاصل از الگوریتم های K-Means و مورد مقایسه قرار گرفت که نتایج نشان دهنده کارایی این الگوریتم در مقابل روشهای دیگر بود.

از مزایای این روش می توان به کارایی این روش بر روی مجموعه داده های حجیم و عدم نیاز به مشخص کردن تعداد خوشه ها قبل از اجرای الگوریتم اشاره نمود. همچنین از معایب آن می توان به مشکل ذاتی اتوماتای یادگیر که همانا سرعت پائین در همگرایی می باشد اشاره کرد. که در این مقاله با استفاده از اتوماتای یادگیر تطبیق پذیر تا حدی این مشکل برطرف شده است.

مراجع

[۱] Rui XU, Wunsch, D. “Survey of Clustering Algorithms”, IEEE Trans. Neural Networks, Vol. ۱۶, pp ۶۴۵-۶۷۸, ۲۰۰۵.

[۲] Freitas, A.A., “A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery,” Advances in Evolutionary Computation, Springer-Verlag, ۲۰۰۲.

[۳] Labroche, N., Monmarché, N. and Venturini, G.,”A new clustering algorithm based on the chemical recognition system of ants,” ۱۵th European Conference on Artificial Intelligence, ۲۰۰۲.

[۴] Wang, L. and Jiao, L., “A novel genetic algorithm based on immunity,” ۲۰۰۰ IEEE International Symposium on Circuits and Systems, pp. ۳۸۵-۳۸۸, ۲۰۰۰.

[۵] Kuo, R. J. and Chung, W. J., “Integration of Self-Organizing Map and Genetic K-Means Algorithm for Data Mining”, Proceedings of ۳۰th International Conference of Computer and Industrial Engineering, Tinos Island, Aggean, Sea, Greece, Jun. ۲۹ – Jul. ۲, ۲۰۰۲a.

[۶] Kuo, R. J., Chang, K. and Chien, S.Y., “Integration of Self-Organizing Feature Map and Genetic Algorithm Based Clustering Method for Market Segmentation,” Journal of Organizational Computing and Electronic Commerce, ۲۰۰۲b.

[۷] Morshedlou, H. and Meybodi, M. R., "A Cellular Automata based Data Clustering Method", Proceedings of the First Iranian Data Mining Conference, Amirkabir University of Technology, Tehran, Iran, Nov. ۱۹-۲۰, ۲۰۰۷.

[۸] Farajzadeh, N. and Meybodi, M. R., "Learning Automata-based Clustering Algorithm for Sensor Networks",Proceedings of ۱۲th Annual CSI Computer Conference of Iran, Shahid Beheshti University, Tehran, Iran, pp. ۷۸۰-۷۸۷, Feb. ۲۰-۲۲, ۲۰۰۷.

[۹] Hossieni Sedehi, M. and Meybodi, M. R., "A Data Clustering Algorithm based on Cellular Learning Automata", Proceedings of the First Iranian Data Mining Conference, Amirkabir University of Technology, Tehran, Iran, Nov. ۱۹-۲۰, ۲۰۰۷.

[۱۰] Narendra, K. S., and Thathachar, M. A. L., Learning Automata: An Introduction, Printice-Hall Inc, ۱۹۸۹.

[۱۱] Thathachar, M. A. L., Sastry, P. S., “Varieties of Learning Automata: An Overview”, IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. ۳۲, No. ۶, PP. ۷۱۱-۷۲۲, ۲۰۰۲.

[۱۲] de Castro, L. N., Timmis, J.: Artificial Immune Systems: A New Computational Intelligence Approach. Springer-Verlag (۲۰۰۲).
[۱۳] Xian Shen, X. Z. Gao, Rongfang Bie, and Xin Jin, “Artificial Immune Networks: Models and Applications”, IEEE, ۲۰۰۶.

زیرنویس

۱ Partitioning
۲ Hierarchical
۳ Density-based
۴ Grid-Based
۵ Linear Reward-Penalty
۶ Linear Reward Epsilon Penalty
۷ Linear Reward Inaction
۸ Self/Non-self discrimination
۹ Negative Selection
۱۰ Clonal Selection
۱۱ Immune Network Theory
۱۲ Danger Theory
۱۳ Paratope
۱۴ Idiotope
۱۵ B-cell receptor
۱۶ Cell activation
۱۷ Cell proliferation
۱۸ Antibody secretion
۱۹ Tolerance
۲۰ Suppression