

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

بابک نصیری^۱؛ محمد رضا 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 سطحی با کم کردن نمونه ها در مجموعه داده صورت می گیرد. شبکه ایمنی مصنوعی، نمونه هایی که می باشند 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، K-MEANS و EM 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، خوش بندی، پیش بینی، کاوش قوانین پیوندی و تشخیص آنومالی نامید. خوش بندی یعنی، تقسیم داده ها به گروه هایی از اشیا مشابه. هدف از خوش 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. خوش 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 خوش بندی داده ها توسعه داده شده اند که می توان آنها را به دسته های زیر تقسیم نمود: خوش بندی جزء بندی^۱، خوش بندی سلسله مراتبی^۲، خوش بندی مبتنی بر تراکم^۳، خوش بندی مبتنی بر توری^۴ و ... [۱]. استفاده از روش های هوشمند برای خوش بندی داده ها نیز در سالهای اخیر بسیار مورد توجه قرار گرفته است که از بین آنها می توان به خوش بندی مبتنی بر شبکه عصبی مصنوعی، الگوریتم ژنتیک [۲]، الگوریتم مورچگان [۳]، جستجوی Tabu [۴] و بعضی الگوریتم های ترکیبی نظر SOM و K-means [۵]، SOM و GA [۶] اشاره نمود. همچنین اخیراً روش هایی مبتنی بر اتماتای سلولی [۷]، اتماتای یادگیر [۸] و اتماتای یادگیر سلولی [۹] برای خوش بندی داده ها ارائه شده است.

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

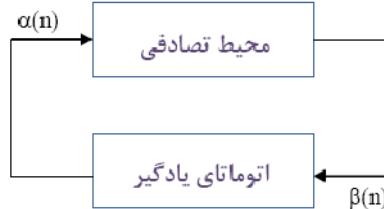
اکثر الگوریتم های ارائه شده خوش بندی را بصورت ایستا انجام می دهند و زمانیکه با حجم انبوهی از داده ها رو به رو باشیم دچار مشکل می شوند. چنانچه بخواهیم از این الگوریتم ها برای خوش بندی پویا بهره ببریم بسیار زمانگیر خواهد بود. در این مقاله یک رهیافت دو مرحله ای برای خوش بندی داده ها بصورت پویا ارائه شده است. ابتدا از یک روش نمونه گیری مبتنی بر سیستم ایمنی مصنوعی (شکله 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 خوش بندی انتخاب شده اند.

در مرحله بعد پس از نمونه گیری داده ها و کاهش حجم آن، نوبت به خوش بندی داده ها بصورت پویا با استفاده از اتماتای یادگیر می رسد. برای این کار به هر نمونه یک اتماتای 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، اتماتای 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 تشخیص دارد.

۲. اتماتای یادگیر

اتماتای یادگیر یک مدل انتزاعی است که بطور تصادفی یک عمل از مجموعه متنهای اعمال خود را انتخاب کرده و بر محیط اعمال می کند. محیط، عمل انتخاب شده توسط اتماتای یادگیر را ارزیابی کرده و نتیجه ارزیابی خود را توسط یک سیگنال تقویتی به اتماتای یادگیر اطلاع می دهد. سپس اتماتای 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 اتماتا یادگیرد تا از بین اعمال خود بهترین عمل را انتخاب کند. بهترین عمل، عملی است که احتمال دریافت پاداش از محیط را به حداقل برساند. کارکرد اتماتای یادگیر در تعامل با محیط، درشکل (۱) مشاهده می شود.



شکل (۱): ارتباط بین اتوماتیک یادگیر با محیط

محیط را می‌توان توسط سه تابی $E = \{\alpha, \beta, C\}$ نشان داد که در آن $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_r\}$ مجموعه ورودیها، $\beta = \{\beta_1, \beta_2, \dots, \beta_m\}$ مجموعه احتمالات جریمه می‌باشد. هرگاه β مجموعه دو عضوی باشد محیط از نوع P می‌باشد. در چنین محیطی $\beta_1 = \beta_2 = \dots = \beta_r$ بعنوان پاسخ نامطلوب یا شکست است. β بعنوان پاسخ مطلوب یا موفقیت در نظر گرفته می‌شوند. در محیط از نوع Q مجموعه β دارای تعداد متناهی عضوی 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 می‌باشد. در یک محیط ایست مقادیر C_i ها ثابت هستند، حال آنکه در یک محیط غیر ایست این مقادیر در طی زمان تغییر می‌کنند. بر اساس اینکه تابع بروزرسانی وضعیت اتوماتیک یادگیر (که با اطلاع از عمل انتخاب شده و سیگنال تقویتی β) وضعیت بعدی اتوماتیک یادگیر را محاسبه کند) ثابت یا متغیر باشد، اتوماتیک یادگیر به دو دسته اتوماتیک یادگیر با ساختار ثابت و اتوماتیک یادگیر با ساختار متغیر تقسیم می‌گردد.

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

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

که a پارامتر پاداش و b پارامتر جریمه می‌باشد. اگر a و b با هم برابر باشند، الگوریتم L_{R-P} ^۵، اگر b از a خیلی کوچکتر باشد، الگوریتم L_{R-E} ^۶ و اگر b صفر باشد، الگوریتم L_{R-I} ^۷ نام دارد.^[۱۰]

۳. سیستم ایمنی مصنوعی

سیستم ایمنی مصنوعی الهام گرفته از سیستم ایمنی طبیعی (سیستم ایمنی موجود در بدن موجودات زنده) می‌باشد. سیستم ایمنی طبیعی وظیفه محافظت از بدن را در برابر موجودات خارجی نظیر باکتری‌ها، ویروس‌ها و ... بر عهده دارد. برای این منظور می‌باشد ابتدا سلول‌های خودی از غیر خودی تشخیص^۸ داده شود، سپس در برابر سلول‌های غیر خودی از خود واکنش نشان دهد. این سیستم همچنین دارای حافظه بوده و آنتی بادی هایی را که قبلاً برای از بین بردن آنتی ژن‌ها بکار رفته اند را نگهداری می‌کند تا در صورت حمله مجدد از طرف این آنتی ژن‌ها سریعاً به آنها پاسخ گوید. الگوریتم‌های موجود در سیستم ایمنی مصنوعی را می‌توان بطور کلی به چهار دسته مختلف تقسیم نمود. این الگوریتم‌ها عبارتند از: انتخاب منفی^۹، انتخاب تولیدمثل^{۱۰}، تئوری شبکه‌های ایمنی^{۱۱} و تئوری خطر^{۱۲} 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 مجموعه‌ای از ایدیوتوپ‌ها شناسایی می‌شوند، تعریف می‌شود. بنابراین هر یک از عناصر شبکه قابلیت شناسایی و شناخته‌شدن را به صورت هم‌مان دارا می‌باشد. این خاصیت منجر به ایجاد شبکه ایمنی می‌شود که در آن، چه مولکول‌های آنتی‌بادی به صورت آزاد و چه به عنوان مولکول گیرنده^{۱۵} B-cell 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} می‌شود. در حالیکه یک پاسخ منفی منجر به تحمل^{۱۹} یا بازدارندگی^{۲۰} در سلول می‌شود[۱۳].

۴. رهیافت پیشنهادی

رهیافت پیشنهادی برای خوش بندی داده‌ها بصورت پویا شامل دو مرحله است. در مرحله اول کاهش حجم داده‌ها مبتنی بر شبکه ایمنی مصنوعی انجام می‌گیرد و در مرحله دوم یک خوش بندی پویا مبتنی بر اتماتای یادگیر ارائه شده است.

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

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

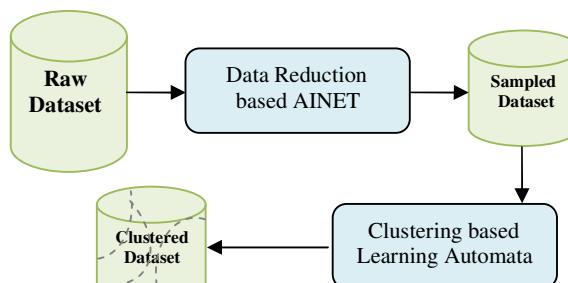
که C_i مرکز خوش i ام، n_c تعداد خوش‌ها می‌باشد. همچنین $d(x, y)$ فاصله اقلیدسی بین دو شیء داده i و j و $\text{diam}(c_i)$ قطر خوش i می‌باشد که به ترتیب با استفاده از روابط (۲) و (۳) محاسبه می‌شوند.

$$d(c_i, c_j) = \min_{x \in c_i, y \in c_j} \{d(x, y)\} \quad (2)$$

$$\text{diam}(c_i) = \max_{x, y \in c_i} \{d(x, y)\} \quad (3)$$

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

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



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

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

۱-۴ - 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 که از دست دادن دقت داده ها به حداقل خود برسد. الگوریتم های مختلفی برای این کار وجود دارد. در این مقاله از شبکه ایمنی مصنوعی برای این منظور استفاده شده است. برای این کار نمونه های موجود در مجموعه داده نقش آنتی ژن ها 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 تولید مثل و جهش پیدا می کنند و بهترین ها در هر تکرار در حافظه نگهداری می شوند. این کار تاریخی به شرط پایان الگوریتم (تعداد تکرار از قبل تعیین شده) 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, $tmpM$;

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 شناس ورود به شبکه ایمنی را در طول تکرارها داشته باشد و از احتمال از دست رفتن اطلاعات در فضاهای خلوت اجتناب شود، اشاره نمود.

۲-۴ - 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 مجموعه یک اتوماتای یادگیر اختصاص داده می شود و همسایگان هر نمونه با توجه به شعاع همسایگی rad (یکی از پارامترهای ورودی مسئله) مشخص شده.

بعنوان عملهای آن نمونه انتخاب می شوند. انتخاب یک عمل توسط یک اتوماتی یادگیر به مفهوم انتخاب همو胥ه ای برای آن اتوماتا می باشد.
الگوریتم خوش بندی با استفاده از اتوماتی یادگیر بصورت زیر می باشد.

الگوریتم ۲: الگوریتم خوش بندی مبتنی بر آتوماتی یادگیر

Algorithm ۲. Clustering based Learning Automata

Input : Reduced Dataset \mathbf{R} , neighborhood radius rad

Output: Cluster Number C_{num} , Clustered Dataset \mathbf{C}

Step۱: find_neighborhoods_of_each_data();

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

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

۳.۱: Add data from Dataset \mathbf{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 $LA_{Selected}$ based probability as $Action_j$.

۳.۲.۳: disable Action $_j$ for $LA_{Selected}$ and also disable Action Selected for LA_j .

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

Step۴: $[C_{num}, C] = \text{find_weakly_Connected_Graph}(\mathbf{Path})$.

Step۵: $\text{fitness} = \text{calculate_Dunn_Index}(C_{num}, C)$.

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 توزیع نرمال (μ, σ^2) تولید شده و متعلق به پنج خوش (هر خوش ۱۰۰۰ داده) با مشخصات زیر می باشد.

$$\begin{array}{lll} \{N(0.2, 0.1^2), & \{N(0.2, 0.1^2), & \{N(0.5, 0.1^2), \\ N(0.2, 0.1^2)\} & N(0.8, 0.1^2)\} & N(0.5, 0.1^2)\} \end{array}$$

$$\begin{aligned} & \{N(0.8, 0.1^2), \quad \{N(0.8, 0.1^2), \\ & N(0.2, 0.1^2)\} \quad N(0.8, 0.1^2)\} \end{aligned}$$

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 اتوماتیک یادگیر (AIN-LA)، روش ترکیبی شبکه ایمنی مصنوعی و اتوماتیک یادگیر تطبیق پذیر (Adaptive AIN-LA)، EM و DBScan.

در روش ترکیبی شبکه ایمنی مصنوعی و اتوماتیک 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 الگوریتم ۱ 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

آستانه بازدارندگی (ts)	تعداد نمونه ها پس از اجرای Algorithm ۱	تعداد نمونه ها	مجموعه داده
۰.۱۴	۲۳	۵۰۰۰	Sample
۰.۲	۱۵	۱۵۰	IRIS
۰.۱	۱۸	۱۷۸	Wine

همانطور که در جدول بالا مشاهده می شود، حجم داده ها پس از اعمال الگوریتم ۱، فوق العاده کاهش می یابد. آستانه بازدارندگی می باشد که چنانچه میزان شیاهت 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.

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

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 تسريع 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 حجمی و عدم نیاز به مشخص کردن تعداد خوشه ها قبل از اجرای الگوریتم اشاره نمود. همچنین از مزایای 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 ۳۰-th 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, Prentice-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. ۲۲, No. ۶, PP. ۷۱۱-۷۲۲, ۲۰۰۲.

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

زنگنه

- ✓ Partitioning
- ✗ Hierarchical
- ✗ Density-based
- ✗ Grid-Based
- ✗ Linear Reward-Penalty
- ✗ Linear Reward Epsilon Penalty
- ✓ Linear Reward Inaction
- ✗ Self/Non-self discrimination
- ✗ Negative Selection
- ✓ Clonal Selection
- ✓ Immune Network Theory
- ✓ Danger Theory
- ✓ Paratope
- ✓ Idiotope
- ✓ B-cell receptor
- ✓ Cell activation
- ✓ Cell proliferation
- ✓ Antibody secretion
- ✓ Tolerance
- ✗ Suppression