University of Qom

Engineering Management and Soft Computing

JEMSC.QOM
Online ISSN: 2538-2675

Discovering the Rules of Data Mining Classification using Distributed Memetic Algorithm

Mohammadreza DehghaniMahmoudabadi[1] ✉ **and Mohammadreza Mollahoseini Ardakan**[2]

1. Corresponding author, Ph.D. Student in computer engineering software, Faculty of Computer Engineering, University of Azad Islamic Maybod, Maybod, Iran. Email: m.r.dehghani.m.a@gmail.com
2. Assistant Professor, Department of Computer Engineering, Maybod Branch, Islamic Azad University, Maybod, Iran. Email: Mollahoseiny@yahoo.com

Article Info

Article type:
Research Article

Article history:
Received 2023 September 28
Received in revised form 2023 November 6
Accepted 2023 November 21
Published online 2024 March 15

Keywords:
Classification rules,
Distributed memetic algorithm,
parallel memetic algorithm.

ABSTRACT

In distributed population systems, cohesive structures prevail, playing a crucial role in the evolution of species across different sites and fostering diversity. These structures employ local selection and reproduction methods to enhance the evolution process. Alterations in migration rules on specific areas, coupled with the execution of search operations, have led to a significant improvement in discovering classification rules. Ultimately, information sharing is employed to mitigate the complexity of the identified law set. This study evaluates the effectiveness of the Distributed Memetic Algorithm in discovering classification rules in data mining. The algorithm is analyzed based on results obtained from five datasets collected from UCI and KEEL repositories. The findings indicate that the Distributed Memetic Algorithm outperforms the traditional Memetic Algorithm in precision for predicting and discovering classification rules in data mining. This research underscores the observable impact of migration operations and search execution in learning classification rules in data mining.

Cite this article: DehghaniMahmoudabadi, M.R. & Mollahoseini Ardakan, M.R. (2023). Discovering the Rules of Data Mining Classification using Distributed Memetic Algorithm. *Enginiiring Management and Soft Computing, 9 (2)*. 50-62. **DOI**: https://doi.org/

© The Author(s)
DOI: https://doi.org/

Publisher: University of Qom

JEMSC.QOM
شاپا الکترونیکی: ۲۶۷۵-۲۵۳۸

مدیریت مهندسی و رایانش نرم

دانشگاه قم

کشف قوانین دسته‌بند داده‌کاوی با استفاده از الگوریتم ممتیک توزیع‌شده

محمدرضا دهقانی محمودآبادی[1]✉ و محمدرضا ملاحسینی اردکانی[2]

۱. نویسنده مسئول، دانشجوی دکتری، مهندسی کامپیوتر نرم افزار، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد میبد، میبد، ایران. رایانامه: m.r.dehghani.m.a@gmail.com

۲. استادیار، مهندسی کامپیوتر نرم افزار، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی واحد میبد، میبد، ایران. رایانامه: Mollahoseiny@yahoo.com

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۲/۰۷/۰۶

تاریخ بازنگری: ۱۴۰۲/۰۸/۱۵

تاریخ پذیرش: ۱۴۰۲/۰۸/۳۰

تاریخ انتشار: ۱۴۰۲/۱۲/۲۵

کلیدواژه‌ها:
الگوریتم ممتیک توزیع‌شده،
الگوریتم ممتیک موازی،
قوانین دسته‌بندی.

چکیده

در سیستم‌های توزیع‌شده جمعیت، فرم‌های به‌هم‌پیوسته حاکم هستند. این فرم‌ها برای تکامل گونه‌ها و افزایش تنوع آنها به‌وسیله روش انتخاب و تولیدمثل محلی مورد استفاده قرار می‌گیرند. در برخی از سایت‌ها، تغییرات و نیز اجرای جستجو در قوانین مهاجرت منجر به بهبود قابل‌توجه در کشف قوانین دسته‌بند شده‌است. در نهایت، اشتراک اطلاعات جهت کاهش پیچیدگی مجموعه مشخص کشف‌شده بکار گرفته می‌شود. در این مطالعه اثربخشی الگوریتم ممتیک توزیع‌شده در کشف قوانین دسته‌بند داده‌کاوی با تحلیل نتایج این الگوریتم در پنج مجموعه داده از مخازن UCI و KEEL مورد ارزیابی قرار گرفته‌است. نتایج حاکی از این است که الگوریتم ممتیک توزیع‌شده با دقت پیش‌بینی بالاتری نسبت به الگوریتم ممتیک سنتی در کشف قوانین دسته‌بند داده‌کاوی عمل می‌کند. این مطالعه نشان می‌دهد که تاثیر عملیات مهاجرت و اجرای جستجو، در کشف قوانین دسته‌بند داده‌کاوی قابل مشاهده است.

استناد: دهقانی محمودآبادی، محمدرضا و ملاحسینی اردکانی، محمدرضا. (۱۴۰۲). «کشف قوانین دسته‌بند داده‌کاوی با استفاده از الگوریتم ممتیک توزیع‌شده». *مدیریت مهندسی و رایانش نرم، دوره ۹ (۲)*. صص: ۶۲-۵۰. https://doi.org/

ناشر: دانشگاه قم

© نویسندگان.

CC BY NC

۱) مقدمه

در دهه‌های اخیر رشد چشمگیر حجم داده‌ها باعث ارتقای فوری و موثر الگوریتم‌های داده‌کاوی به‌منظور استخراج دانش مفید و کاربردی شده‌است. دسته‌بندی، یکی از تکنیک‌های حیاتی در زمینه داده‌کاوی است و توانایی ساخت الگوهایی دارد که کلاس‌های داده را توصیف و تفکیک می‌کند حتی برای مواردیکه برچسب کلاس آنها ناشناخته است. الگوریتم‌های ممتیک نمونه‌هایی از الگوریتم‌های قوی و سازگار جستجو هستند که در فضای راه‌حل‌ها به دنبال جستجوی جهانی می‌پردازند. با توجه به حجم بسیار زیاد فضای احتمالی کاندیداها برای مجموعه داده‌های متوسط، الگوریتم‌های ممتیک گزینه محبوبی برای کشف قوانین در دسته‌بندی با استفاده از قوانین نمادین "اگر-آنگاه" هستند (Gutiérrez et al, ۲۰۱۷).

گرچه محققان در حوزه داده‌کاوی به موفقیت‌هایی در کشف قوانین دقیق و جالب دست یافته‌اند اما دستیابی به تعادل مناسب بین اکتشاف و بهره‌برداری همچنان چالش‌هایی را به همراه دارد. بنابراین تنظیم تعادل بین اکتشاف و بهره‌برداری و بهینه‌سازی پارامترهای الگوریتم‌های ممتیک به‌عنوان یک راه‌حل مطلوب و ضروری مطرح شده‌است. این تعادل با ترکیب متنوع روش‌های انتخاب، مقیاس‌بندی و مدل‌های جستجو، تحقق می‌یابد. در نهایت، الگوریتم‌های ممتیک موازی یا توزیع‌شده به‌عنوان ابزارهای اساسی برای حل مسائل همگرایی زودرس در داده‌کاوی بکار گرفته می‌شوند (Zhang & Xing, ۲۰۱۸).

راه‌حل اصلی برای مواجهه با چالش همگرایی زودرس در جستجو، تنظیم تعادل میان اکتشاف، بهره‌برداری و بهینه‌سازی پارامترهای الگوریتم ممتیک است. محققان در این زمینه برای دستیابی به تعادل مطلوب میان اکتشاف و بهره‌برداری، از روش‌های متعددی ازجمله انتخاب، مقیاس‌بندی و مدل‌های متنوع استفاده کرده‌اند. اگرچه این روش‌ها نمی‌توانند مسئله همگرایی زودرس را به‌طور کامل حل کنند ولیکن الگوریتم‌های ممتیک موازی یا توزیع‌شده در این زمینه بسیار بکار گرفته می‌شوند (Wu et al, ۲۰۱۸).

الگوریتم ممتیک توزیع‌شده به‌عنوان یک مدل جزیره یا ممتیک دانه‌درشت شناخته می‌شود. این الگوریتم، جمعیت را به زیرجمعیت‌هایی تقسیم می‌کند و به دلیل نیمه‌انزوایی از زیرجمعیت‌ها، در طول تکامل تنوع را حفظ می‌کند و جستجو را انجام می‌دهد. افراد زیرجمعیت به‌طور مستقل و همزمان با جستجو تکامل می‌یابند و به‌طور دوره‌ای اپراتورها مهاجرت برخی از افراد با عملکرد بهتر را در زیرجمعیت‌ها تبادل می‌کنند تا در مسیرهای جستجوی امیدوارکننده قرار بگیرند (Townsend & De Corcuera, ۱۹۹۳).

در این تحقیق، الگوریتم ممتیک توزیع‌شده جهت کشف قوانین دسته‌بندی، بکار گرفته شده‌است. این الگوریتم توزیع‌شده در قالب یک راه‌حل پیشنهادی ارائه شده‌است که به کنترل ترافیک داخل سایت‌ها می‌پردازد. با تعیین یک بازه اپراتور مهاجرت بین سایت‌ها، الگوریتم به‌منظور رعایت قوانین با پوشش کمتر از مهاجرت استفاده می‌شود. نتایج تحقیق نشان می‌دهد که الگوریتم ممتیک توزیع‌شده قادر به تولید راه‌حل‌هایی با کارآیی و استفاده زمان بهینه‌تری است. از این رو، تکامل پارامترهای الگوریتم ممتیک توزیع‌شده منجر به بهبود زمانی می‌شود که این امر به‌عنوان یک مزیت مهم این الگوریتم در مقایسه با روش‌های دیگر برجسته می‌شود. این نتایج با استفاده از مقایسه با الگوریتم ممتیک سنتی در چند

مجموعه داده مختلف از مخازن یادگیری ماشین به‌دست‌آمده و به‌عنوان یک پیشنهاد موثر در زمینه کنترل ترافیک و بهینه‌سازی الگوریتم‌های ممتیک توزیع‌شده مطرح شده‌اند.

در این تحقیق، از الگوریتم ممتیک توزیع‌شده به‌صورت موازی برای کشف قوانین دسته‌بندی و به‌دست‌آوردن قوانین دقیق و قابل‌فهم استفاده شده‌است. سپس رویکرد پیشنهادی در یک پردازنده شبیه‌سازی ارائه شده و اثربخشی الگوریتم پیشنهادی با مقایسه نتایج الگوریتم ممتیک سنتی در ۵ مجموعه داده از مخزن یادگیری ماشین UCI5 و مخزن KEEL4 ارزیابی شده‌است. نتایج حاکی از این است که الگوریتم ممتیک توزیع‌شده در کل مجموعه داده‌ها دارای پیش‌بینی با دقت بالاتری است. عملکرد این الگوریتم توزیع‌شده با توجه به نرخ مهاجرت، فراوانی مهاجرت و نوع قوانین، مورد تجزیه‌وتحلیل قرار گرفته‌است که این تحلیل به درک بهتر از عملکرد الگوریتم در شرایط مختلف کمک می‌کند. این نتایج به‌عنوان یک پیشرفت مهم در حوزه الگوریتم‌های ممتیک توزیع‌شده و کاربرد آنها در مسائل دسته‌بندی مورد تأیید بین‌المللی قرار می‌گیرد.

۲) پیشینه پژوهش

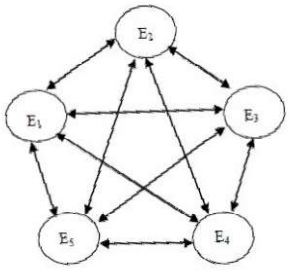
الگوریتم‌های داده‌کاوی به دلیل روبه‌روشدن با مجموعه داده‌های حجیم و چالش‌های ابعاد، نیاز به مقیاس‌پذیری دارند. فضای جستجوی قوانین در کشف الگوها و قوانین دسته‌بندی با افزایش تعداد ویژگی‌ها گسترش می‌یابد. به همین دلیل محققان از الگوریتم‌های ممتیک برای حل این چالش‌ها و بهبود کارآیی در مقابل داده‌های حجیم استفاده کرده‌اند (Buayen & Werapun, 2018).

با چالش‌های وجودی در زمینه داده‌کاوی مانند مسائل همگرایی زودرس، مدت‌زمان طولانی و جستجوی محلی که با الگوریتم‌های ممتیک ساده همراه است، نسخه‌های موازی از الگوریتم‌های ممتیکی تدوین شده‌اند. این الگوریتم‌های ممتیک موازی به‌خصوص برای حل مسائل داده‌کاوی طراحی شده‌اند اما تعداد محدودی از مقالات تحقیقاتی در این زمینه مورد بررسی قرار گرفته‌است. انجام عملیات محاسباتی برای حل مسائل داده‌کاوی امری زمان‌بر است بنابراین یک ساختار، برای توزیع بهینه محاسبات در میان چندین پردازنده ارائه شده‌است. اثبات کارآیی این ساختار با استفاده از آزمون‌های استاندارد و حل مسائل دسته‌بندی موردی در داده‌کاوی انجام شده و نشان می‌دهد که الگوریتم‌های ممتیک موازی علاوه بر افزایش بهره‌وری و بهینه‌سازی زمانی، کارآیی را نیز بهبود می‌بخشند (Gowanlock & Karsin, 2019).

اگر الگوریتم ممتیک موازی، روی یک پردازنده واحد اجرا شود، قادر به ارائه راه‌حل‌های بهینه است. الگوریتم‌های ممتیک توزیع‌شده، یک کلاس مؤثر از الگوریتم‌های بهینه‌سازی در حوزه علم داده‌کاوی به‌شمار می‌روند. مدل مهاجرتی خودسازگار الگوریتم ممتیک، به‌منظور کشف قوانین دسته‌بند داده‌کاوی، پیشنهاد شده‌است. در این مدل، پارامترهایی مانند اندازه جمعیت، تعداد نقاط متقاطع، جستجوی محلی و سرعت جهش برای هر نسل متوالی تطبیق‌پذیر است. همچنین در مدل‌های جزیره‌ای از پارامترهایی با اندازه متغیر استفاده می‌شود. این الگوریتم ممتیک توزیع‌شده با اجرای مدیریت دینامیک مهاجرت افراد بین جمعیت حتی در شرایط استثناها و سلسله‌مراتب، به کشف قوانین مفید و قابل‌فهم می‌پردازد (Mohammadreza, n.d.).

۳) روش‌شناسی پژوهش

الگوریتم‌های ممتیک توزیع‌شده به‌تنهایی نسخه‌های موازی الگوریتم‌های ممتیک متوالی نیستند بلکه هدف آنها ارتقای عملکرد نسبت به جمع رفتار الگوریتم‌های سنتی است. رویکرد پیشنهادی در الگوریتم ممتیک توزیع‌شده باعث یکپارچگی جستجوی محلی در سطح سایت‌های محلی با جستجوی عمومی می‌شود. زیرمجموعه‌های جدا شده به‌عنوان سایت‌ها در چندین نسل اولیه به‌صورت محلی تغییر می‌کنند و سپس مهاجرت قوانین بین این زیرجمعیت‌ها اتفاق می‌افتد. فراوانی و میزان مهاجرت ازجمله عوامل مهمی هستند که بر عملکرد هر الگوریتم ممتیک توزیع‌شده تاثیرگذارند.



شکل ۱. اهداف الگوریتم ممتیک توزیع‌شده

الگوریتم ممتیک توزیع‌شده پیشنهادی با تمرکز بر پنج دید مهم E1 تا E5، یک رویکرد نوین را به ارتقای عملکرد الگوریتم‌های ممتیک ارائه می‌دهد. این دیدگاه‌ها از جفت‌های سایت‌ها برای تبادل قوانین بهره می‌برند و با مبادله قواعد، دقت مجموعه‌ها را بهبود می‌بخشند. از آنجاییکه دقت متوسط این مجموعه‌ها نقش حیاتی در بهبود کارآیی الگوریتم دارد، الگوریتم ممتیک توزیع‌شده پیشنهادی این امکان را ایجاد می‌کند که با افزایش دقت، بهترین نتایج را در جستجوی قوانین بین سایت‌ها به‌دست آورد.

۳-۱) جمعیت اولیه[۴۵]

تازه‌ترین تحقیقات در زمینه تولید قوانین دسته‌بندی از مجموعه داده Iris با استفاده از جمعیت اولیه‌ای از زیرگروه‌ها آغاز می‌شود. این جمعیت از قوانین تصادفی شروع به تولید می‌کند و سپس با استفاده از رمزگذاری عدد صحیح با طول ثابت، افراد در مدل نمایش داده می‌شوند. در این روند، گزاره‌هایی به شکل "اگر P، سپس D" بیان می‌شوند که P یک فرضیه پیشین است و یک جفت ویژگی ارزش را نمایان می‌کند. تصمیم‌گیری براساس نتیجه D انجام می‌شود و این یک اصطلاح واحد است که شامل ارزش برای ویژگی هدف می‌شود. مجموعه داده Iris از مخزن UCI به‌عنوان یک مثال اولیه مورد استفاده قرار گرفته و عملگرها و میزان ارزش‌گذاری مورد بحث و توضیح قرار می‌گیرند. این مجموعه داده شامل داده‌های مداوم است که با استفاده از ابزار Weka گسسته می‌شوند و تمام صفات پس از تفسیر دارای سه مقدار هستند. جدول مشخصه‌های مجموعه داده Iris نیز برای بهترین درک و دریافت اطلاعات بیشتر ارائه شده‌است (Omara & Arafa, 2010).

[45] Papulation

جدول ۱. توضیحات از دیتاست Iris

جایگاه	صفات	ارزش	مشخصات
۰	Petal length	High, Medium, Small	‘1’,’2’,’3’
۱	Petal width	Large, Medium, Small	‘1’,’2’,’3’
۲	Sepal length	High, Medium, Small	‘1’,’2’,’3’
۳	Sepal width	Large, Medium, Small	‘1’,’2’,’3’
٤	Iris	Setosa, Versicolor, Virginica	‘1’,’2’,’3’

با توجه به رمزگذاری مقادیر کلاس‌ها "Setosa"، "versicolor"و "virginica" به‌ترتیب به‌عنوان "۱"، "۲" و "۳"، مجموعه قوانین نمونه در شکل ۲ به‌صورت دو بخش رمزگذاری‌شده (الف) و رمزگشایی‌شده (ب) نمایش داده شده‌است. اینجا زیرجمعیت‌های محلی با استفاده از بازه زمانی معین شروع به تکامل می‌کنند و پس از آن مهاجرت بین جمعیت‌ها رخ می‌دهد(Mishra & Singh, 2015) . این روند به معیارهای مشخصی ادامه می‌یابد و در نهایت، از اپراتور زیرمجموعه برای کاهش پیچیدگی قوانین مشخص و جمع‌آوری از تمام سایت‌ها بهره می‌برد.

ب

```
IF Sepal_length=small AND Petal_length=small THEN
Iris=setosa
IF Petal_length=medium THEN Iris=versicolor
IF Sepal_width=medium AND Petal_width=small THEN
Iris=setosa
IF Petal_width=medium THEN Iris=versicolor
```

الف

۱	۰	۳	۰	۳
۲	۰	۲	۰	۰
۱	۳	۰	۲	۰
۲	۲	۰	۰	۰

شکل ۲. نماینده: (الف) ماتریس؛ (ب) قوانین

۳-۲) تکامل محلی

در این شبیه‌سازی، الگوریتم ممتیکی ازدحام‌سازی برای تکامل زیرمجموعه‌های E1 ، E2، E3، E4 و E5 به‌صورت محلی بکار گرفته شده‌است. تمام زیرمجموعه‌ها به‌صورت پردازش موازی در شبیه‌سازی جداگانه تکامل یافته و آنها جداگانه ارزیابی شده‌اند. در این فرآیند، میزان ارزش‌گذاری مورداستفاده، عملگرهای انتخابی و نرخ ترکیبی با جزئیات بیشتر در مقاله ارائه شده توسط (Vonásek et al., ۲۰۱۸) توضیح داده شده‌است.

۳-۲-۱) عملکرد ارزش‌گذاری

قبل از تعریف عملکرد ارزش‌گذاری، ابتدا مفاهیم اساسی مرتبط با اقدامات دقت در دسته‌بندی قوانین (کروموزوم‌ها) در جمعیت باید یادآوری شوند. هنگامی که از یک قاعده برای دسته‌بندی یک کلاس پیش‌بینی‌شده استفاده می‌شود، این قانون نتایج متفاوتی را در مقایسه با کلاس واقعی به‌وجود می‌آورد که عبارتنداز:

- مثبت واقعی: کلاس مثبت که به‌طور صحیح توسط قانون پیش‌بینی می‌شود، P ^ D؛(TP[46])
- مثبت نادرست: کلاس منفی که به‌اشتباه توسط قانون پیش‌بینی شده‌است. P ^ ~ D؛(FP[47])

[46] True Positive
[47] False Positive

- منفی واقعی: کلاس منفی که به‌طور صحیح توسط قانون پیش‌بینی شده‌است ~ D ^ ~ P؛(TN[48])
- منفی نادرست: کلاس منفی که به‌طور اشتباه با قانون پیش‌بینی شده‌است ~ D ^ P؛(FN[49])

در اینجا، فرضیه P و D بخشی از تصمیم یک قاعده هستند. عملکرد ارزش‌گذاری دو اقدام حساسیت و ویژگی به شرح ذیل تعریف شده‌است:

رابطه ۱) $$Sensitivity = \frac{|TP|}{|TP+FN|}$$

رابطه ۲) $$totaltotalSpecificity = total\frac{|TN|}{|TN+FP|}$$

سرانجام عملکرد ارزش‌گذاری استفاده‌شده توسط دسته‌بندی با توجه به حساسیت و ویژگی به‌عنوان امتیاز قانون شناخته می‌شود یعنی:

رابطه ۳) $$Rule\ Score = Sensitivity * Specificity$$

هدف اصلی از ایجاد مکانیسم تکاملی، حداکثر کردن حساسیت و ویژگی‌ها در یک زمان مشخص است. با این حال، اگر یک قانون به حالت مراقبت‌های خاصی محدود شود، ارزش‌گذاری آن به صفر کاهش پیدا می‌کند (Martín et al., ۲۰۱۸).

۳-۲-۲) عملگرهای ممتیک

در این تحقیق از انتخاب چرخ رولت، برای انتخاب افراد جهت تولید نسل بعدی استفاده شده‌است تا با تولید فرزندان جدید و بهتر، در فرآیند تکاملی آینده شرکت کنند. از اپراتور متقاطع یک نقطه‌ای، برای ترکیب افراد تولیدکننده ممتیک جدید استفاده شده‌است. سایت‌های تصادفی انتخاب شده و مقادیر مربوط به آنها جهت ایجاد افراد جدید، تبادل می‌شود. هنگامیکه دو فرد از دسته‌های مختلف با یکدیگر ترکیب می‌شوند، بیشتر آنها بدترین فرزندان را تولید می‌کنند زیرا تعلق به گونه‌های مختلف در جمعیت وجود دارد. بنابراین یک متقاطع به افراد متعلق به همان دسته محدود می‌شود (Yuming et al., ۱۹۹۵).

در این تحقیق، از اپراتور جهش برای حفظ تنوع در جمعیت استفاده می‌شود. با هدف بهبود تکامل، از جهش سازشی برای تاخیر در از دست دادن افراد بهتر تا آخرین مراحل مسابقات استفاده شده‌است. در این راستا، ابتدا احتمال جهش را به ۰.۱ تنظیم می‌کنیم و سپس با توجه به دقت نسل قبلی، احتمال تغییرات در جمعیت تعیین می‌شود.

رابطه ٤) $$\text{جهش ضریب} = \begin{cases} 0.1 & \text{اگر دقت ژن قبلی} > \text{دقت} \\ \text{دقت ژن قبلی} - \text{دقت} \Big/ \text{دقت} & \text{در غیراینصورت} \end{cases}$$

اپراتور جهش‌یافته، با اعمال آن قادر است وضعیت مراقبت ویژه را به شکلی تخصصی با ارزش صفر جهش‌یافته تعیین کند یا مقادیر ویژگی منتقل‌شده به وضعیت را با افزودن یا حذف بندهای شرطی در بخش‌های قبلی تعمیم دهد. این امکان به این تحقیق اجازه می‌دهد تا با استفاده از اپراتور جهش‌یافته، تغییرات دقیق و حیاتی در قوانین کاندیداها را ایجاد و ارزیابی نماید.

[48] True Negative
[49] False Negative

۳-۲-۳) ازدحام‌سازی

استفاده از تکنیک ازدحام‌سازی جهت حفظ تنوع در زیرمجموعه‌ها و جلوگیری از همگرایی به یک قانون، این الگوریتم ممتیک را در بهینه‌سازی کارآیی بهبود می‌بخشد. در این رویکرد نوین، با حفظ همپوشانی در جمعیت نسل جدید، از تکرار چندباره الگوریتم جلوگیری شده و به جای آن، فرزندان با استفاده از عملگرهای ممتیک از بین بدترین افراد اما با تنوع مشابه ایجاد می‌شوند. این رویکرد نه‌تنها به بهبود کارآیی الگوریتم کمک می‌کند بلکه از حفظ گونه‌های مختلف در جمعیت نیز پشتیبانی می‌کند .(Veloso et al.,2018) برای اندازه‌گیری شباهت بین افراد جامعه و فرزندان، از مکانیسم زیر استفاده شده‌است:

رابطه ۵) $$\text{ضریب اشتراک} = |N_1 \cap N_2|$$

رابطه ٦) $$\text{ضریب اجتماع} = |N_1 \cup N_2|$$

رابطه ۷) $$\text{ضریب شباهت} = \frac{\text{ضریب اشتراک}}{\text{ضریب اجتماع}}$$

در اینجا با ارائه مفهومی از شباهت بین قوانین فرزندان و بدترین قوانین، الگوریتم ممتیک از معیارهای اشتراک‌پذیری و اجتماع‌پذیری برای ارزیابی کیفیت آنها استفاده می‌کند. مقدار شباهت که در بازه ۰ تا ۱ قرار دارد، نشان‌دهنده اندازه اشتراک و اجتماع میان دو قانون است. این اطلاعات مهم در فرآیند جایگزینی قانون فرزندان با بدترین قاعده از جمعیت الگوریتم ممتیک به‌منظور حداکثرکردن تشابه با تعداد بیشتری از بدترین قوانین موجود است. این رویکرد بهبودی موثر در بهینه‌سازی جمعیت الگوریتم فراهم می‌کند و نتایج بهتری را در فرآیند تکاملی فراهم می‌آورد.

۳-۳) تکامل بین سایت‌ها یا عمومی

فرآیند تکامل بین سایت‌ها یا عناصر ممتیکی در زیرمجموعه‌ها با هدف برقراری تعادل میان اکتشاف و بهره‌برداری انجام می‌شود. ارزیابی دقیقی از کیفیت زیرمجموعه‌ها به‌عنوان معیار کلی درنظر گرفته‌می‌شود تا امکان تصمیم‌گیری بهتر درمورد تبادل و مهاجرت قوانین در میان نماها ایجاد شود. این رویکرد موثر به بهبود توازن بین اکتشاف و بهره‌برداری در فرآیند تکامل کمک کرده و اطمینان از کیفیت بالاتر تصمیم‌گیری‌ها را فراهم می‌آورد (.Cano-Cano et al ۲۰۱۹).

۳-۳-۱) دقت[50]

برای ارزیابی دقت یک جمعیت لازم است از زیرجمعیت‌ها (که قوانین در یک سایت را شامل می‌شوند) در فرآیند تکامل داخلی استفاده شود. این انتخاب دقیق عملکرد ارزش‌گذاری را به یک ابزار حیاتی تبدیل می‌کند زیرا باعث می‌شود جستجو به سوی بهبود مجموعه قوانین به‌صورت جهانی هدایت شود. عملکرد ارزش‌گذاری برای تکامل داخلی سایت به شرح زیر استفاده می‌شود (.Lebib et al ۲۰۱۷).

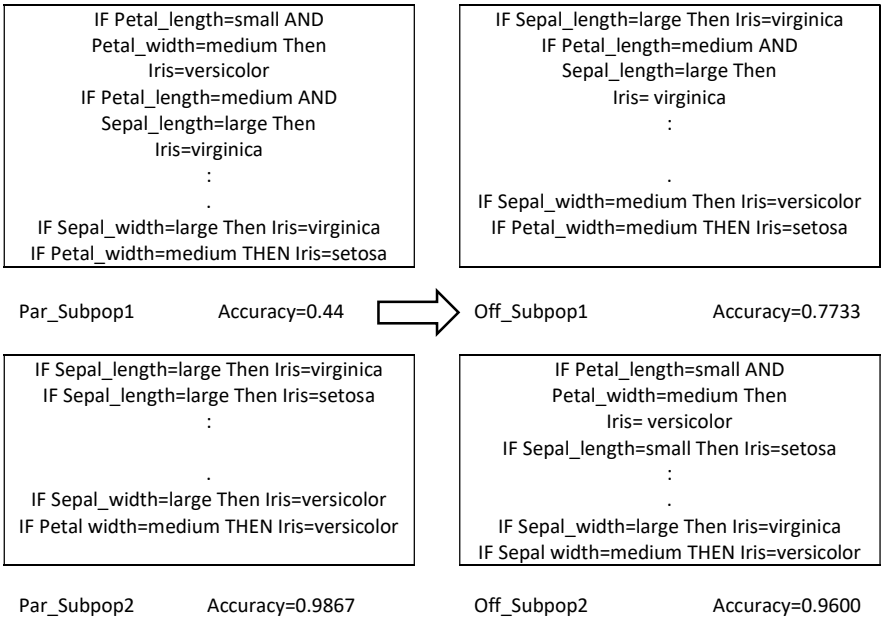
رابطه ۸) $$\text{دقت} = \frac{|TP+TN|}{N}$$

[50] Accuracy

که در آن N تعداد کل موارد موجود در مجموعه داده است.

۳-۳-۲) اپراتور مهاجرت

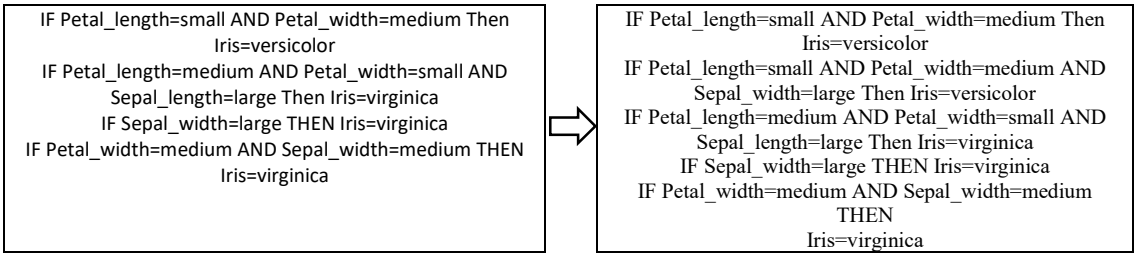
جهت بهبود کارآیی الگوریتم ممتیک توزیع‌شده، چندین جفت از سایت‌ها به‌صورت تصادفی انتخاب شده و ارزش‌گذاری آنها محاسبه شده‌است. این انتخاب جفت‌ها براساس تعداد جزایر یا به‌صورت کمتر از آن انجام می‌شود. سایتی که دوبار در یک جفت انتخاب می‌شود، دورهای تبادل را طی می‌کند. در شکل ۱ این انتخاب چندین جفت را نشان می‌دهد. سپس در شکل ۳ تعداد محدودی از آنها به‌طور تصادفی انتخاب شده و فرآیند تبادل افراد بین سایت‌های انتخاب‌شده را نشان می‌دهد. پس از تبادل بهترین افراد بین جفت‌های انتخاب‌شده، میانگین آمادگی فرزندان جدید محاسبه و با میانگین ارزش‌گذاری اصلی مقایسه می‌شود. اگر میانگین زیرمجموعه والدین کمتر از میانگین فرزندان باشد، زیرمجموعه والدین توسط زیرمجموعه‌های فرزندان جایگزین می‌شوند. در غیر این صورت، هیچ تغییری در جمعیت اتفاق نمی‌افتد. به‌عنوان مثال در شکل زیر، زیرمجموعه‌های والدین (Rule1 subpop1) و(Rule1 subpop2) با یکدیگر ردوبدل شده و ارزش دقت آنها با استفاده از تابع ارزیابی به‌دست آمده‌است. ارزش دقت زیرمجموعه والدین ۰.۷۲ و ارزش دقت سایت‌های فرزندان ۰.۸۶۵ می‌باشد بنابراین ارزش سایت‌های والدین کمتر از ارزش سایت‌های فرزندان است و تبادل قوانین بین زیرمجموعه‌ها انجام می‌شود.



شکل ۳. اپراتور مهاجرت

۳-۴) اپراتور اشتراک

اپراتور اشتراک به‌عنوان یک مرحله پس از پایان فرآیند تکامل بکار می‌رود. این مرحله از تکامل قوانین، با جمع‌آوری قوانین از تمام سایت‌ها شروع می‌شود. سپس از اپراتور اشتراک استفاده می‌شود تا قوانین را از زیرمجموعه‌هایی که نمونه‌های داده تحت پوشش هر قانون دیگری در جمعیت را شامل می‌شود، حذف کند. این عملگر با حذف تکرارها و افزایش یکپارچگی از مجموعه کلی قوانین، پیچیدگی را بهبود می‌بخشد. شکل ۴ نمایانگر عملکرد این اپراتور اشتراک است.



شکل ٤. عملگر اشتراک

نمونه‌های داده تحت پوشش قانون

‘IF Petal_length=small AND Petal_width=medium AND Sepal_width=large Then Iris=versicolor’

زیرمجموعه نمونه‌های داده تحت پوشش این قانون است:

‘IF Sepal_width=large Then Iris=versicolor’.

بنابراین قاعده

‘IF Petal_l length = small AND Petal_width = medium AND '

از جمعیت حذف می‌شود زیرا طبق قانون کلی فرض می‌شود

‘IF Petal_length=small AND Petal_width=medium AND Sepal_width=large Then Iris=versicolor’ Sepal_width=large Then Iris=versicolor’.

۴) یافته‌های پژوهش

رویکرد الگوریتم ممتیک توزیع‌شده، جهت ارزیابی عملکرد خود از پنج مجموعه داده که از KEEL و مخزن یادگیری ماشین UCI گرفته شده‌اند، استفاده می‌کند. این مجموعه داده‌ها در جدول ۲ خلاصه شده‌اند. برای اعتبارسنجی، از روش متقاطع دو برابر برای مجموعه داده‌های با کمتر از ۷۰۰ هزار نمونه و اعتبارسنجی متقاطع پنج برابر برای مجموعه دادهای با بیش از ۷۰۰ هزار نمونه استفاده می‌شود. تنظیم دقیق پارامترها یک جنبه مهم در این رویکرد است و پس از آزمایش‌های انجام‌شده، مقادیر بهینه پارامترها در جدول ۳ گزارش شده‌اند.

جدول ۲. مجموعه داده‌های مورد استفاده در آزمایش‌ها

مجموعه داده‌ها	موارد	ویژگی‌های	کلاس‌ها	روش آزمون
Mushroom	۵۶۴۴	۲۳	۲	۱۰-CV
Vote	۴۳۵	۱۷	۲	۲-CV
Iris	۱۵۰	۵	۳	۲-CV
Zoo	۱۰۱	۱۷	۷	۲-CV
Nursery	۱۲۹۶۰	۹	۵	۱۰-CV

جدول ۳ . پارامترهای بهینه‌سازی شده الگوریتم ممتیک توزیع‌شده

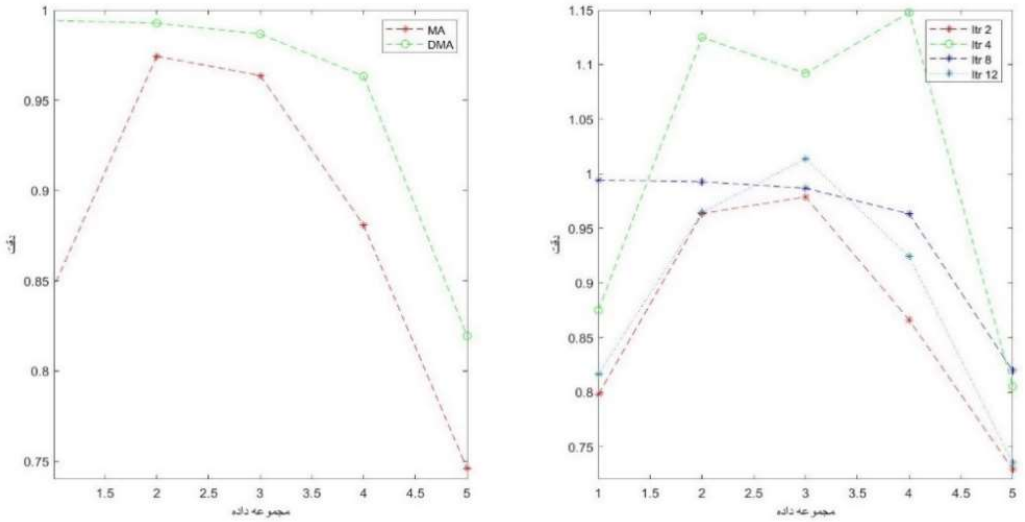
پارامتر	ارزش
اندازه جمعیت	٤۰۰
تعداد سایت‌ها	۵
سایز اولیه	۸۰
روش انتخاب	رولت ویل

تعداد ماکزیمم تولید	۱۵۰
احتمال جهش	۰.۱
احتمال متقاطع	۰.٦
فاکتور ازدحام	٤
فرکانس مهاجرت	۵
تعداد مهاجرت	۸
روش جستجو	دودویی

جدول ٤ . مقایسه الگوریتم ممتیک توزیع‌شده با الگوریتم ازدحام ممتیک

مجموعه داده	الگوریتم ممتیک	الگوریتم ممتیک توزیع‌شده
Mushroom	۰.۰۸۷۰±۰.۷۵۹۲	۰.۰۱۲٦±۰.۹۸۱٦
Vote	۰.۰۲۳٦±۰.۹۵۰۸	۰.۰۰٦٤±۰.۹۸٦۳
Iris	۰.۰۱۸۳±۰.۹٤۵۳	۰.۰±۰.۹۸٦۷
Zoo	۰.۰۳۵۵±۰.۸٤۵۳	۰.۰۱۳۰±۰.۹۵۰۳
Nursery	۰.۰٦۹۸±۰.٦۷٦۲	۰.۰۵۱۳±۰.۷٦۸۲

در مقایسه میان دقت پیش‌بینی‌شده، الگوریتم ممتیک ساده و الگوریتم ممتیک توزیع‌شده در جدول ۴، عملکرد بهتر الگوریتم ممتیک توزیع‌شده در تمام پنج مجموعه داده مشاهده می‌شود. این بهبود عملکرد می‌تواند به این دلیل باشد که الگوریتم ممتیک توزیع‌شده ابتدا بهینه‌سازی لیست بهترین قوانین را انجام داده و سپس با مهاجرت بین سایت‌ها، مجموعه قوانین را بهینه‌تر می‌کند. نمودارهای شکل ۵ نیز دقت بیش از ۵ بار اجرای هر الگوریتم را برای پنج مجموعه داده نشان می‌دهد. الگوریتم ممتیک توزیع‌شده با ادغام صفت‌ها و همچنین فعل و انفعالات قاعده، عملکرد بهتری از الگوریتم ممتیک ساده کسب می‌کند. این نشان‌دهنده توانایی الگوریتم ممتیک توزیع‌شده در تولید و بهینه‌سازی لیست بهتری از قوانین است.



شکل ۵. دقت در ۵ اجرا بر روی مجموعه داده‌ها

جدول ۵. دقت الگوریتم ممتاز توزیع‌شده با تعداد متفاوت مهاجرت و مبادله در مجموعه داده‌های مختلف

مجموعه داده	تعداد مهاجرت			
	۲	۴	۸	۱۲
Mushroom	۰.۰۱۷۰±۰.۷۸۱۰	۰.۰۱۳۰±۰.۸۶۲۰	۰.۰۱۲۶±۰.۹۱۶	۰.۰۱۵۴±۰.۸۰۱۲
Vote	۰.۰۳۱۵±۰.۹۳۲۰	۰.۱۴۳۰±۰.۹۸۲۰	۰.۰۰۶۴±۰.۹۸۶۳	۰.۰۴۳۵±۰.۹۲۱۰
Iris	۰.۰۱۷۵±۰.۹۶۱۳	۰.۱۱۰۰±۰.۹۸۲۰	۰.۰±۰.۹۸۶۷	۰.۰۴۷۱±۰.۹۶۶۷
Zoo	۰.۰۴۳۰±۰.۸۲۳۱	۰.۲۷۶۰±۰.۸۷۱۶	۰.۰۱۳۰±۰.۹۵۰۳	۰.۰۸۲۴±۰.۸۴۱۸
Nursery	۰.۰۲۸۱±۰.۷۰۱۲	۰.۰۳۷۱±۰.۷۶۸۰	۰.۰۵۱۳±۰.۷۶۸۲	۰.۰۳۸۲±۰.۶۹۷۱

در جدول ۵ تعداد مهاجرت و مبادله در مجموعه‌های مختلف به‌دقت نمایش داده شده‌است. آزمایشات نشان می‌دهد که نرخ مهاجرت بسیار کم است و الگوریتم ممتیک توزیع‌شده مانند الگوریتم ممتیک عمل می‌کند. افزایش نرخ مهاجرت فراتر از حد مجاز، منجر به کاهش عملکرد الگوریتم ممتیک توزیع‌شده می‌شود. این اتفاق به دلیل کاهش تنوع در نرخ مهاجرت است. برای حفظ عملکرد مثبت، الگوریتم ممتیک توزیع‌شده نیاز به یافتن نرخ مهاجرت بهینه دارد. نرخ مهاجرت بهینه برای الگوریتم ممتیک توزیع‌شده پیشنهادی در تمام مجموعه‌های داده ۸ است.

۵) نتیجه‌گیری و پیشنهادها

الگوریتم ممتیک توزیع‌شده، یک رویکرد پیشنهادی برای کشف قوانین دسته‌بندی دقیق است. این رویکرد الگوریتمی مبتنی بر تقسیم جمعیت به زیرجمعیت‌هاست که به‌طور همزمان تکامل می‌یابند و از مهاجرت افراد برای تبادل اطلاعات استفاده می‌کنند. الگوریتم در پنج مجموعه داده مختلف اعتبارسنجی شده‌است و نشان داده‌است که با دقت بالایی می‌تواند مجموعه‌های قوانین را کشف کند. این رویکرد بهینه‌سازی لیست بهترین قوانین بین سایت‌ها را ارتقا می‌دهد و از مهاجرت برای بهینه‌سازی این لیست بهره می‌برد. هرچند عملکرد الگوریتم به میزان مهاجرت وابسته است اما با تنظیم نرخ مهاجرت، بهترین نتایج را ارائه می‌دهد. این الگوریتم به‌صورت همگن یعنی با ممتیک واحد و با تنظیم پارامترها برای تکامل درونی، عمل می‌کند. نتایج ناهمگن و سلسله‌مراتبی نشان‌دهنده کارآیی موفق در بهینه‌سازی کارکردهای عددی است و این امکان را فراهم می‌کند که این الگوریتم در زمینه استخراج قوانین نیز موثر باشد.

منابع

Buayen, P., & Werapun, J. (2018). Parallel time–space reduction by unbiased filtering for solving the 0/1-Knapsack problem. *Journal of Parallel and Distributed Computing*, *122*, 195–208. https://doi.org/10.1016/j.jpdc.2018.08.003

Cano-Cano, J., Andújar, F. J., Alfaro, F. J., & Sánchez, J. L. (2019). Speeding up exascale interconnection network simulations with the VEF3 trace framework. *Journal of Parallel and Distributed Computing*, *133*, 124–135. https://doi.org/10.1016/j.jpdc.2019.06.013

Gowanlock, M., & Karsin, B. (2019). Accelerating the similarity self-join using the GPU. *Journal of Parallel and Distributed Computing*, *133*, 107–123. https://doi.org/10.1016/j.jpdc.2019.06.005

Gutiérrez, M., Gregorio-Godoy, P., Pérez Del Pulgar, G., Munoz, L. E., Sáez, S., & Rodríguez-Patón, A. (2017). A New Improved and Extended Version of the Multicell Bacterial Simulator gro. *ACS Synthetic Biology*, *6*(8), 1496–1508. https://doi.org/10.1021/acssynbio.7b00003

Lebib, F. Z., Mellah, H., & Drias, H. (2017). Enhancing information source selection using a genetic algorithm and social tagging. *International Journal of Information Management*, *37*(6), 741–749. https://doi.org/10.1016/j.ijinfomgt.2017.07.011

Martín, A., Lara-Cabrera, R., Fuentes-Hurtado, F., Naranjo, V., & Camacho, D. (2018). EvoDeep: A new evolutionary approach for automatic Deep Neural Networks parametrisation. *Journal of Parallel and Distributed Computing*, *117*, 180–191. https://doi.org/10.1016/j.jpdc.2017.09.006

Mishra, V., & Singh, V. (2015). Generating Optimal Query Plans for Distributed Query Processing using Teacher-Learner Based Optimization. *Procedia Computer Science*, *54*, 281–290. https://doi.org/10.1016/j.procs.2015.06.033

Mohammadreza. (n.d.). A novel aggregation-based distributed search model for optimal design using the memetic algorithm.

Omara, F. A., & Arafa, M. M. (2010). Genetic algorithms for task scheduling problem. *Journal of Parallel and Distributed Computing*, *70*(1), 13–22. https://doi.org/10.1016/j.jpdc.2009.09.009

Townsend, J. G., & De Corcuera, J. B. (1993). Feminists in the rainforest in Mexico. *Geoforum*, *24*(1), 45–54. https://doi.org/10.1016/0016-7185(93)90013-8

Veloso, B., Leal, F., González-Vélez, H., Malheiro, B., & Burguillo, J. C. (2018). Scalable data analytics using crowdsourced repositories and streams. *Journal of Parallel and Distributed Computing*, *122*, 1–10. https://doi.org/10.1016/j.jpdc.2018.06.013

Vonásek, V., Vick, A., & Krüger, J. (2018). Distributed motion planning for industrial random bin picking. *Procedia CIRP*, *76*, 121–126. https://doi.org/10.1016/j.procir.2018.01.039

–Wu, L., Yan, C., Jian, M., Liu, S., Dong, W., & Chen, C. W. (2018). A fast hybrid retargeting scheme with seam context and content aware strip partition. *Neurocomputing*, *286*, 198–213. https://doi.org/10.1016/j.neucom.2018.01.058

Yuming, C., Jing, D., & Fanlun, X. (1995). Genetic Algorithms for Irrigation Optimization. *IFAC Proceedings Volumes*, *28*(4), 145–150. https://doi.org/10.1016/s1474-6670(17)45555-8

–Zhang, G., & Xing, K. (2018). Memetic social spider optimization algorithm for scheduling two-stage assembly flowshop in a distributed environment. *Computers and Industrial Engineering*, *125*, 423–433. https://doi.org/10.1016/j.cie.2018.09.007