J. Insur. Res., 14(2): Pages: 109-118, Spring 2025, Serial #52

Journal of Insurance Research
(IJIR)

Homepage: https://ijir.irc.ac.ir/?lang=en

Iranian Insurance Research Center (IIRC)

ORIGINAL RESEARCH PAPER

Detecting car insurance fraud using improved clustering with genetic algorithm

B. Yousefimehr, M. Ghatee*, S. Moradi, Y. Tafakor, S. Tavakoli

Department of Computer Science, Faculty of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Ploytechnic), Tehran, Iran.

ARTICLE INFO

Article History:
Received 04 August 2024
Revised 11 November 2024
Accepted 07 December 2024

Keywords:
Artificial intelligence
Car insurance
Clustering
Fraud detection
Genetic algorithm

ABSTRACT

BACKGROUND AND OBJECTIVES: Clustering is one of the basic techniques in data mining and machine learning, which is used to divide a set of data into homogeneous subsets. There are different methods for clustering, each of which has its own strengths and weaknesses. One of the main challenges in clustering is finding the optimal number of clusters and optimal allocation of data to these clusters. Genetic algorithm, as an optimization method based on natural evolution, has a high ability to solve complex problems and search for large solution spaces and can be used as an effective tool in clustering. The purpose of this article is to investigate the efficiency and accuracy of genetic algorithm in data classification and compare it with traditional clustering methods for classification. In order to evaluate the performance of this algorithm, several insurance data sets are used and the obtained results are analyzed with different criteria such as accuracy. Also, different parameters of the genetic algorithm are examined and their effects on the final performance of the algorithm are studied in order to determine the most optimal settings for data classification.

METHODS: In this research, to form chromosomes, at first, the number of clusters was determined. Considering that each cluster center had as many features as the number of features in the data set, the length of each chromosome was determined by multiplying the number of clusters by the number of features. New and diverse methods were used for Crossover, Mutation and Survival processes. Also, the evaluation criterion similar to the K-means algorithm was chosen to optimize the clustering performance. This innovative approach led to improving the accuracy and efficiency of the classification process.

FINDINGS: By applying the method described in this article to three insurance data sets for fraud detection, we have interesting results with 12% improvement in F1 and 10% increase in accuracy in the first data set, 1% improvement in F1 and 1% improvement in accuracy in the first data set. Second and finally, 1% improvement in F1 and 2% improvement in the accuracy of the third data set compared to the K-means method and other methods have been achieved. Due to the 2-mode data in this data set, the problem is solved for two clusters using the algorithm and the best label for each cluster is selected according to the real labels of the data and the result is presented as the results of classification problems. Additionally, significant improvements in metrics such as ARI and other clustering evaluation criteria have been achieved, and remarkable progress has been made compared to the standard genetic algorithm.

CONCLUSION: Genetic Algorithm is able to solve complex problems without definite solution and can perform better in data clustering than traditional methods such as K-means. By combining probabilities and randomness, this approach provides the possibility to examine more points as cluster centers and improve clustering performance. The results show that this method works better than the famous methods in some cases and provides a suitable structure for data clustering.

*Corresponding Author:
Email: *ghatee@aut.ac.ir*
Phone: +9821 64542531
ORCID: 0000-0002-9558-8286

DOI: 10.22056/ijir.2025.02.02

نشریه علمی پژوهشنامه بیمه

سایت نشریه: https://ijir.irc.ac.ir/?lang=fa

مقاله علمی

تشخیص تقلب در بیمهٔ خودرو با استفاده از خوشه‌بندی بهبودیافته با الگوریتم ژنتیک

بهنام یوسفی‌مهر، مهدی قطعی*، سینا مرادی، یاسمین تفکر، ساجد توکلی

گروه علوم کامپیوتر، دانشکده ریاضی و علوم کامپیوتر، دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)، تهران، ایران.

اطلاعات مقاله

تاریخ های مقاله:

تاریخ دریافت: ۱۴ مرداد ۱۴۰۳

تاریخ داوری: ۲۱ آبان ۱۴۰۳

تاریخ پذیرش: ۱۷ آذر ۱۴۰۳

کلمات کلیدی:

الگوریتم ژنتیک

بیمهٔ خودرو

تشخیص تقلب

خوشه‌بندی

هوش مصنوعی

***نویسنده مسئول:**

ایمیل: *ghatee@aut.ac.ir*

تلفن: ۶۴۵۴۲۵۳۱ ۹۸۲۱+

ORCID: 0000-0002-9558-8286

DOI: 10.22056/ijir.2025.02.02

چکیده:

پیشینه و اهداف: خوشه‌بندی یکی از روش‌های اساسی در داده‌کاوی و یادگیری ماشین است که برای تقسیم مجموعه‌ای از داده‌ها به زیرمجموعه‌های همگن به کار می‌رود. روش‌های مختلفی برای انجام خوشه‌بندی وجود دارد که هریک نقاط قوت و ضعف خاص خود را دارند. یکی از چالش‌های اصلی در خوشه‌بندی، یافتن تعداد خوشه‌های بهینه و تخصیص بهینهٔ داده‌ها به این خوشه‌هاست. الگوریتم ژنتیک، به‌عنوان روش بهینه‌سازی مبتنی بر تکامل طبیعی، توانایی بالایی در حل مسائل پیچیده و جست‌وجوی فضای جواب‌های بزرگ دارد و می‌تواند به‌عنوان یک ابزار مؤثر در خوشه‌بندی به کار رود. هدف این مقاله، بررسی کارایی و دقت الگوریتم ژنتیک در کلاس‌بندی داده‌ها و مقایسهٔ آن با روش‌های سنتی خوشه‌بندی برای کلاس‌بندی است. به‌منظور ارزیابی عملکرد این الگوریتم، چندین مجموعه داده بیمه استفاده شده و نتایج به‌دست‌آمده با معیارهای مختلفی مانند دقت تحلیل می‌شوند. همچنین، پارامترهای مختلف الگوریتم ژنتیک بررسی‌شده و تأثیر آن‌ها بر عملکرد نهایی الگوریتم مطالعه می‌شود تا بهینه‌ترین تنظیمات برای کلاس‌بندی داده‌ها تعیین شود.

روش‌شناسی: در این پژوهش، به‌منظور تشکیل کروموزوم‌ها، ابتدا تعداد خوشه‌ها مشخص شد. با توجه به اینکه هر مرکز خوشه به اندازهٔ تعداد ویژگی‌های مجموعه داده دارای ویژگی بود، طول هر کروموزوم به‌صورت حاصل‌ضرب تعداد خوشه‌ها در تعداد ویژگی‌ها تعیین شد. برای فرایندهای Mutation، Crossover و Survival از روش‌های نوین و متنوعی بهره گرفته شد. همچنین، معیار ارزیابی مشابه الگوریتم K-means انتخاب شد تا عملکرد خوشه‌بندی بهینه‌سازی شود. این رویکرد نوآورانه به بهبود دقت و کارایی فرایند کلاس‌بندی منجر شد.

یافته‌ها: با اعمال روش توضیح‌داده‌شده در این مقاله برای تشخیص تقلب در ۳ مجموعه دادهٔ بیمه، به نتایج جالب توجهی با ۱۲٪ بهبود در F1 و ۱۰٪ افزایش دقت در مجموعه دادهٔ اول، ۱٪ بهبود F1 و دقت در مجموعه دادهٔ دوم و در نهایت نیز ۱٪ بهبود در F1 و ۲٪ بهبود در دقت مجموعه دادهٔ سوم نسبت به روش K-means و سایر روش‌ها حاصل شده است. با توجه به ۲ کلاس بودن داده‌ها در این مجموعه داده‌ها، مسئله به‌ازای ۲ خوشه با استفاده از الگوریتم حل شده و بهترین برچسب برای هر خوشه با توجه به برچسب‌های واقعی دادگان انتخاب شده و نتیجه به‌صورت نتایج حاصل از مسائل دسته‌بندی ارائه شده است، همچنین بهبود چشمگیری در معیارهایی همچون ARI و سایر معیارهای ارزیابی خوشه‌بندی حاصل شده و پیشرفت چشمگیری نسبت به الگوریتم ژنتیک عادی نیز حاصل شده است.

نتیجه‌گیری: الگوریتم ژنتیک قابلیت حل مسائل پیچیده و بدون راه‌حل قطعی را دارد و می‌تواند در خوشه‌بندی داده‌ها عملکرد بهتری نسبت به روش‌های سنتی مانند K-means داشته باشد. این رویکرد با ترکیب احتمالات و تصادفی بودن، امکان بررسی نقاط بیشتر به‌عنوان مراکز خوشه و بهبود عملکرد خوشه‌بندی را فراهم می‌کند. نتایج نشان می‌دهد که این روش در برخی موارد بهتر از روش‌های معروف عمل می‌کند و ساختار مناسبی برای خوشه‌بندی داده‌ها ارائه می‌دهد.

توجه: مدت‌زمان بحث و انتقاد برای این مقاله تا ۱ ژوئیه ۲۰۲۵ در وب‌سایت IJIR در «نمایش مقاله» باز است.

مقدمه

صنعت بیمه به‌دلیل ماهیتش، به‌راحتی در معرض کلاهبرداری و تقلب قرار می‌گیرد. در بیمهٔ خودرو، بیمه‌گر تمامی خسارت‌هایی را که توسط خودرو یا بار آن به اشخاص ثالث وارد می‌شود، پوشش می‌دهد (Seidi Aghili Abadi *et al.*, 2017).

برای تشخیص کلاهبرداری می‌توان از روش‌های نظارت‌شده و بدون نظارت بهره گرفت(Yousefimehr and Ghatee, 2025). روش‌های نظارت‌شده زمانی که داده‌های برچسب‌خورده موجود باشند، کارآمدند (Ahmadlou *et al.*, 2023). اما در صورت محدودیت یا هزینه‌بر بودن برچسب‌گذاری، روش‌های بدون نظارت مانند خوشه‌بندی K-means و سلسله‌مراتبی برای شناسایی موارد غیرعادی کاربرد دارند(Tajaddodi Nodehi *et al.*, 2023). در این پژوهش، فرض شده که امکان برچسب‌گذاری وجود ندارد و نیاز به یک الگوریتم بدون نظارت است.

الگوریتم ژنتیک که هالند در سال ۱۹۹۲ معرفی کرد، از فرایند تکامل زیستی و نظریهٔ بقای اصلح داروین الهام گرفته است. عناصر اصلی الگوریتم ژنتیک شامل نمایش کروموزوم، انتخاب براساس شایستگی، و عملگرهای زیست‌شناسی مانند انتخاب، جهش، و تبادل هستند. کروموزوم‌ها به‌صورت رشته‌های باینری نمایش داده شده و به‌طور تکراری با استفاده از عملگرهای ژنتیکی جایگزین می‌شوند (Katoch *et al.*, 2021).

الگوریتم k-means روشی محبوب برای خوشه‌بندی داده‌هاست که مجموعه داده‌ها را به k خوشه تقسیم می‌کند. ابتدا k نقطه به‌عنوان مراکز اولیه به‌صورت تصادفی انتخاب می‌شود و هر نمونه داده به نزدیک‌ترین مرکز خوشه اختصاص می‌یابد. مراکز خوشه‌ها تا زمانی که تغییرات ناچیز شوند، به‌روزرسانی می‌شوند. هدف اصلی الگوریتم، کمینه کردن مجموع مربعات فاصله بین نقاط داده و مراکز خوشه‌هاست (Ahmed *et al.*, 2020).

الگوریتم‌های خوشه‌بندی سلسله‌مراتبی با ترکیب یا تقسیم گروه‌های موجود عمل می‌کنند و تعداد خوشه‌ها در ابتدا مشخص نمی‌شود. خوشه‌بندی می‌تواند به دو روش از پایین به بالا یا از بالا به پایین انجام شود و نتیجهٔ آن به‌صورت درخت نمایش داده می‌شود (Shetty and Singh, 2021).

لازم است توجه داشته باشیم که در مسائل یادگیری بدون ناظر، به‌دلیل نبود داده‌هایی برچسب‌دار، دقت و نتیجهٔ حاصل از مدل به‌طور چشمگیری نسبت به مدل‌هایی که با استفاده از داده‌های برچسب‌دار کار می‌کنند، کمتر است. به همین علت در شرایطی که داده‌های بدون برچسب در اختیار داریم و می‌خواهیم که خود مدل، با توجه به ویژگی‌های داده، بتواند بدون ناظر یاد بگیرد؛ با چالش‌های متنوعی مواجهیم. مثلاً در شرایطی که با داده‌هایی ۲ کلاسه و خطی جداییناپذیر سروکار داشته باشیم، روشK-means قادر به خوشه‌بندی صحیح داده‌ها در ۲ کلاس اصلی داده‌ها نیست (Jain, 2010).

با توجه به روش‌های گوناگون مسئلهٔ خوشه‌بندی، سعی داریم روشی خوشه‌بندی مبتنی بر مرکز (Centroids-based) بر پایهٔ الگوریتم ژنتیک ارائه دهیم که با توجه به عملکرد مناسب K-means در صنعت، عملکردی شبیه و نزدیک به این روش داشته باشد و همچنین با دخیل کردن احتمالات به مدل خود اجازه دهیم که حتی عملکردی بهتر از K-means داشته باشد (Hruschka *et al.*, 2009).

در روش‌های خوشه‌بندی مبتنی بر مرکز، چالش و مسئلهٔ اصلی پیدا کردن مناسب مرکز خوشه‌هاست. روش K-means که معروف‌ترین روش در این دسته است، میانگین داده‌های درون یک خوشه را به‌عنوان مرکز خوشه آن در نظر می‌گیرد (Yong and Xin cheng, 2012). اما در حالات بسیاری بهترین حالتی که می‌توان برای مرکز خوشه در نظر گرفت بدین‌صورت نیست و K-means نیز نتایج خوبی نخواهد داشت و همچنین انتخاب تصادفی مراکز اولیهٔ خوشه‌ها که ممکن است به همگرایی بهینهٔ محلی و نتایج ضعیف منجر شود. این الگوریتم به انتخاب اولیهٔ مراکز حساس است و ممکن است به‌جای یافتن بهینهٔ جهانی، به نتایج محلی بسنده کند. همچنین، استفاده از فاصلهٔ اقلیدسی و فرض کروی بودن خوشه‌ها باعث ناکارآمدی در شناسایی خوشه‌های پیچیده یا هم‌پوشان می‌شود (Ikotun *et al.*, 2023).

همچنین معیارهای متنوعی برای اندازه‌گیری فاصله نیز استفاده می‌شود، مثلاً فاصلهٔ منهتنی یا اقلیدسی (Singh *et al.*, 2013).

حال اگر بخواهیم روشی مبتنی بر مرکز برای خوشه‌بندی ارائه دهیم که مبتنی بر الگوریتم ژنتیک باشد، باید جواب مسئله را که همان مرکز خوشه‌هاست در قالب کروموزوم‌هایی دربیاوریم (Maulik and Bandyopadhyay, 2000) و باید ساختار مسئله را به ساختار مورداستفاده توسط الگوریتم ژنتیک دربیاوریم (Bhatia, 2014).

مواردی مثل تابع برازندگی (fitness function)، عملیات انتخاب (selection)، عملیات تقاطع (crossover) و عملیات جهش (mutation) را بر روی آن تعریف کنیم (Roy and Sharma, 2010).

همچنین باید نحوهٔ تخصیص داده‌ها به هریک از این مراکز خوشه را تعریف کنیم، سپس می‌توانیم این مسئله را با استفاده از الگوریتم ژنتیک حل کنیم و مقایسه‌ای از آن با سایر روش‌های خوشه‌بندی داشته باشیم (Rahman and Islam, 2014).

نتایج نشان می‌دهد که ترکیب الگوریتم ژنتیک با K-means به‌دلیل تشابه در تابع ارزیابی، نسبت به سایر روش‌ها عملکرد بهتری دارد. استفاده از K-means در ابتدا، سرعت همگرایی را افزایش می‌دهد، درحالی‌که الگوریتم ژنتیک برای بهبود نتایج و کاوش گسترده‌تر استفاده می‌شود و این ترکیب در داده‌های کمیاب یا غیرقابل اعتماد، دقت و انعطاف‌پذیری بیشتری ارائه می‌دهد.

می‌توان نوآوری این مقاله را در ساختاربندی دقیق و بهبودیافتهٔ الگوریتم ژنتیک برای مسئلهٔ خوشه‌بندی و تنظیم عملگرهای آن به‌صورتی که بهترین کارایی را داشته باشد، خلاصه کرد که این نوآوری بر روی مجموعه داده‌های بیمه بررسی شده است.

مروری بر پیشینهٔ پژوهش

Lu *et al.* (2016) در روش خود از ترکیب K-means و الگوریتم

ژنتیک برای حل مسئلهٔ چند فروشندهٔ دوره‌گرد بدون تقاطع مسیرها استفاده کردند. K-means برای تقسیم‌بندی رئوس و انتخاب شهر آغازین استفاده شده و الگوریتم ژنتیک به‌صورت موازی در هر زیرمجموعه اجرا می‌شود. این ترکیب هم به اهداف تعیین‌شده می‌رسد و هم به‌دلیل پردازش موازی سرعت بالایی دارد.

روش دیگری را Babaie *et al.* (2016) ارائه کرده است که برای خوشه‌بندی و برنامه‌ریزی بودجهٔ خانوارها، از ترکیب دو روش K-means و ژنتیک استفاده می‌کند. همین‌طور ژنتیک به‌عنوان یک الگوریتم فراابتکاری توانسته سرعت رسیدن به جواب بهینه در الگوریتم K-means را بالا ببرد.

Aibinu *et al.* (2016) روشی ارائه داد تا در زمان ترکیب این دو الگوریتم، کنترل جمعیت انجام شده و همین‌طور عملیات تقاطع، با چندین کروموزوم والد انجام شود. نتایج نشان می‌دهد که این الگوریتم روی مسیریابی ربات‌ها بسیار خوب عمل کرده است

در روش Chang *et al.* (2009)، از الگوریتم ژنتیک برای خوشه‌بندی استفاده شده که در آن هر کروموزوم نمایانگر مرکز خوشه‌هاست. با تعریف بازترتیبی ژن و استفاده از یک عملگر تقاطع مبتنی بر شباهت بین کروموزوم‌ها، فضای جست‌وجو بهتر بررسی می‌شود. نتایج نشان می‌دهد این روش عملکرد بهتری نسبت به K-means داشته و انعطاف‌پذیری بالایی در حل مسائل مختلف دارد

بررسی این مقالات نشان می‌دهد که ترکیب روش خوشه‌بندی با الگوریتم ژنتیک می‌تواند گام مثبتی برای بهبود تشخیص تقلب باشد.

روش‌شناسی پژوهش

در این بخش به بررسی روش پیشنهادی می‌پردازیم. شکل ۱ عملکرد کلی روش ارائه‌شده را که مشابه الگوریتم ژنتیک است، نشان می‌دهد.

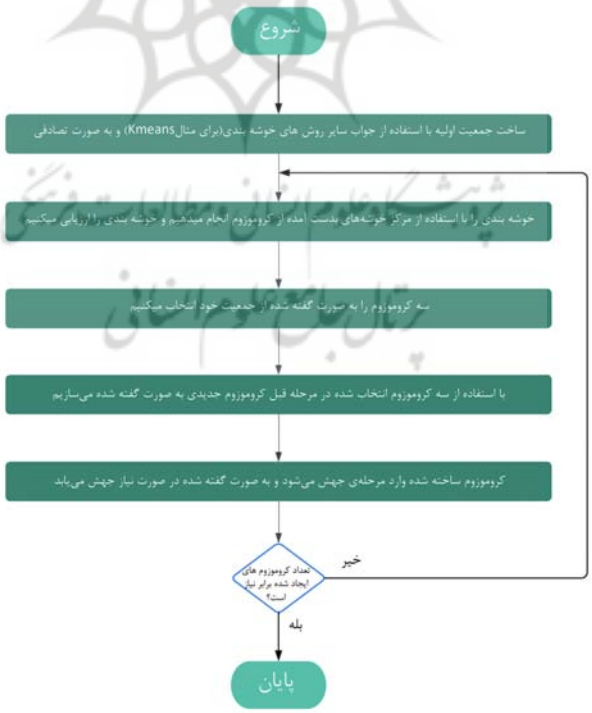
در خوشه‌بندی مبتنی بر مراکز، لازم بود داده‌های موجود را درون خوشه‌هایی قرار دهیم و هر خوشه را به‌گونه‌ای انتخاب کنیم که داده‌های درون آن بیشترین تراکم و نزدیکی به یکدیگر را داشته باشند (Sonia Sharma and Shikha Rai, 2014). برای انجام این کار، نیاز بود تا مراکز خوشه‌ها را بیابیم و هر داده را به نزدیک‌ترین خوشه وصل کنیم.

در الگوریتم ژنتیک، باید پاسخ مسئله را درون کروموزوم‌ها ذخیره می‌کردیم و تلاش می‌کردیم با استفاده از عملیات تقاطع و جهش کروموزوم‌هایی با کارایی بهتر تولید کنیم و بهترین کروموزوم را به‌عنوان پاسخ نهایی برگردانیم (Katoch *et al.*, 2021).

برای ذخیره‌سازی مراکز خوشه‌ها درون کروموزوم‌ها، هر مرکز خوشه به‌عنوان یک مختصات هندسی در فضای n بعدی ذخیره می‌شود که n تعداد ویژگی‌های مورد بررسی است. با داشتن k خوشه، هر کروموزوم باید k مختصات n بعدی را ذخیره کند. بنابراین، کروموزوم‌ها به‌صورت یک آرایهٔ k×n در نظر گرفته شدند که هر خانه آن یک ژن محسوب می‌شود. بنابراین تعداد ژن‌های لازم برای هر کروموزوم، از رابطهٔ زیر به دست می‌آید:

تعداد ژن کروموزوم (Kn) = تعداد خوشه (K) × تعداد ویژگی (n)

در این پژوهش، با هدف تشخیص داده‌های تقلبی، تعداد خوشه‌ها



شکل ۱. مراحل کلی روش ارائه‌شده
Figure 1. General steps of the method

به‌دلیل دوکلاسه بودن داده‌ها برابر با ۲ در نظر گرفته شده است. بااین‌حال، بسته به نوع مجموعه داده و نیاز مسئله، می‌توان تعداد خوشه‌ها را تغییر و با برچسب‌های موجود در مسئله تطبیق داد.

گرچه درصورتی‌که از تعداد برچسب‌های نهایی هیچ اطلاعی نداشته باشیم، می‌توان خود K را نیز به‌عنوان پارامتر یک ژن درون کروموزوم‌ها قرار داد و بر آن اساس به ادامهٔ کار پرداخت (Maulik and Bandyopadhyay, 2000).

با توجه به اینکه مختصات در این مسئله به‌صورت اعداد اعشاری است، ساختار کروموزوم‌ها به‌صورت آرایه‌ای از اعداد اعشاری تعریف شدند. سپس، نیاز بود معیاری برای سنجش کیفیت کروموزوم‌ها تعریف شود تا بتوان هر کروموزوم را رتبه‌بندی کرد. گفته شد که اگر مرکز خوشه به‌گونه‌ای تعیین شود که نقاط نزدیک به آن خوشه متراکم باشند، بدین معناست که خوشه‌بندی مناسبی انجام شده است (Chiang *et al.*, 2006).

بدین‌ترتیب، مراکز خوشه‌ها از کروموزوم استخراج شدند، داده‌های موجود به نزدیک‌ترین مرکز خوشه متصل شدند و مجموع فاصلهٔ اقلیدسی نقاط از مرکز هر خوشه محاسبه شد. بدیهی است که هرچه این مجموع فاصلهٔ اقلیدسی کمتر باشد، کروموزوم بهتر خواهد بود. بنابراین، خوب بودن کروموزوم نسبتی عکس با این مقدار دارد. در نتیجه، معکوس این مجموع فاصله از مراکز خوشه‌ها به‌عنوان معیار ارزندگی معرفی شد.

$$\text{ارزندگی} = \frac{1}{\sum_{k=1}^{N_c}\sum_{j=1}^{N_k}\sqrt{\sum_{i=1}^{F}(C_{ki}-P_{ij})^2}}$$

که N_c تعداد خوشه‌ها، N_k تعداد داده‌های درون خوشهٔ K، F تعداد ویژگی‌های مجموعه داده و C_K مرکز خوشهٔ K است.

در استفاده از الگوریتم ژنتیک، نیاز بود تعدادی کروموزوم برای تشکیل جمعیت اولیه تولید کنیم (Poikolainen *et al.*, 2015). روش‌های متعددی برای تولید این جمعیت وجود دارد. برای ساخت کروموزوم‌ها ابتدا بازهٔ اعداد هر ویژگی تعیین شد و سپس یک عدد تصادفی در آن بازه برای هر ژن انتخاب گردید. به‌این‌ترتیب، جمعیت اولیه به‌صورت تصادفی ایجاد شد و سپس با استفاده از تابع ارزیابی، کیفیت هر کروموزوم سنجیده و ذخیره شد.

علاوه‌براین، می‌توانستیم برای محدودتر کردن بازه و استفاده از دقت روش K-means، بازهٔ انتخاب اعداد برای ژن i را به $[-2\times c_i, 2\times c_i]$ محدودتر کنیم، که در اینجا c_i بعد i ام مرکز خوشهٔ ارائه‌شده توسط K-means است. این اقدام به بهبود دقت و کارایی الگوریتم کمک می‌کرد.

علاوه‌بر تولید تصادفی جمعیت اولیه، می‌توان تعداد بسیار محدودی کروموزوم با استفاده از این بازه‌ها تولید کرد تا کروموزوم‌های آینده بهبود یابند و کیفیت جامعهٔ اولیه ارتقا یابد (Kudova, 2007). پس از ساخت جمعیت اولیه، نیاز بود تعدادی والد از بین کروموزوم‌ها انتخاب شوند و فرزندانی با استفاده از آن‌ها تولید شوند (Jebari, 2013). در بخش انتخاب والد، دو والد با بهترین ارزندگی و یک والد دیگر به‌صورت تصادفی از جمعیت فعلی برای تکثیر انتخاب شدند. سپس، نیاز بود روشی برای نحوهٔ عملیات تقاطع معرفی کنیم تا بتوانیم با استفاده از آن از والدها فرزندانی ایجاد کنیم. روش‌های متنوعی برای تقاطع وجود دارد، مانند تقاطع تک‌نقطه‌ای، تقاطع چندنقطه‌ای، تقاطع یکنواخت و امثال آن (Zainuddin and Abd Samad, 2020).

عملیات تقاطع به این صورت انجام شد که هر ژن با احتمالی خاص از والد ۱، ۲ یا ۳ انتخاب می‌شد. این روش به‌عنوان تقاطع یکنواخت تصادفی معرفی می‌شود که هر ژن ۵۰ درصد احتمال دارد از والد ۱ (بهترین کروموزوم)، ۳۳ درصد احتمال دارد از والد ۲ (دومین بهترین کروموزوم) و ۱۲ درصد احتمال داشت از والد ۳ (یک کروموزوم تصادفی) انتخاب شود. این احتمالات به‌صورت تجربی تعیین شده و قابل تغییر هستند. در نهایت، روشی برای جهش نیز لازم است (De Falco *et al.*, 2002).

ما از ترکیب روش‌های جهش تصادفی استفاده کردیم. به این صورت که ابتدا، با توجه به نرخ جهش، اگر کروموزومی وارد فرایند جهش می‌شد، هر ژن آن به‌صورت تصادفی دوباره با توجه به بازه‌ای که مقادیر آن ویژگی در آن قرار داشت، با عددی تصادفی از آن بازه جایگزین می‌شد.

روش دیگری که برای جهش در این گزارش استفاده شده، جهش جایگزینی است. ازآنجاکه فقط دو خوشه داشتیم، یک موقعیت به‌صورت تصادفی در کروموزوم انتخاب شد و سپس طبق فرمول زیر، دو ژن با هم جابه‌جا شدند:

$$(i+n)\%(k*n)$$

در اینجا، n تعداد ویژگی‌ها، i جایگاه انتخابی و k تعداد خوشه‌هاست. این ترکیب از روش‌های جهش، به افزایش تنوع ژنتیکی و بهبود کارایی الگوریتم کمک می‌کرد. در این فرمول، اگر بخواهیم بعدی از مرکز خوشه را با بعدی دیگر از مرکز خوشهٔ دیگر هم عوض کنیم، ابتدا یک موقعیت تصادفی در کروموزوم انتخاب می‌شود. سپس، برای تعویض آن با ژن متناظر در خوشه دیگر، تعداد خوشه‌ها را در تعداد ویژگی‌های مجموعه داده ضرب کرده و باقی‌ماندهٔ آن موقعیت تصادفی انتخاب شده، به‌علاوه تعداد بعدها را نسبت به آن محاسبه می‌کنیم تا ژن متناظر را به دست آوریم و جای دو ژن را عوض کنیم.

ابتدا جمعیت اولیه تولید و ارزش هر فرد محاسبه شد. والدین انتخاب و فرزندان از طریق ترکیب و جهش ایجاد شدند. بهترین کروموزوم به‌عنوان جواب نهایی انتخاب شد و نقاط به نزدیک‌ترین مرکز خوشه متصل شدند. کروموزوم‌ها براساس تعداد خوشه‌ها تعیین شده‌اند و با تغییر تعداد خوشه‌ها، طول آن‌ها نیز تغییر می‌کند. در این پژوهش، ساختار کروموزوم‌ها براساس تعداد خوشه‌ها تعیین شده است و با تغییر تعداد خوشه‌ها، طول کروموزوم‌ها نیز تغییر می‌کند. برای یافتن تعداد بهینهٔ خوشه‌ها، می‌توان یک ژن خاص برای این پارامتر تعریف کرد و عملگرهای الگوریتم ژنتیک را به‌طور جداگانه برای آن تنظیم نمود که به‌دلیل مشخص بودن تعداد خوشه‌ها از قبل، این پارامتر در نظر گرفته نشده است.

نتایج و بحث

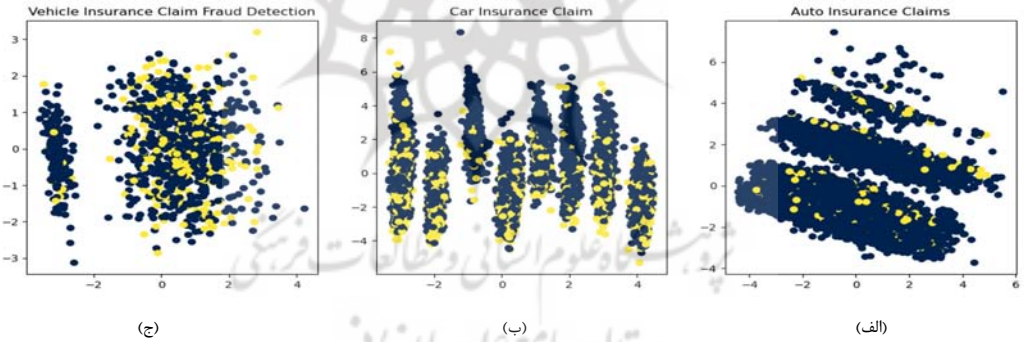
برای کاهش ابعاد داده‌ها از روش تحلیل مؤلفه‌های اصلی (Principal Component Analysis) استفاده شده است. کاهش ابعاد از چندین بعد به دو بعد به‌منظور ساده‌سازی و بصری‌سازی داده‌ها انجام گرفت تا امکان تحلیل و مشاهده الگوها، به‌ویژه در تشخیص تقلب فراهم شود. با انتخاب دو مؤلفهٔ اصلی، امکان تجسم داده‌ها در فضای دوبعدی فراهم آمد که باعث می‌شود الگوها و روابط بین داده‌ها راحت‌تر قابل شناسایی و تفسیر باشد. در شکل ۲ هر نقطه نشان‌دهندهٔ یک نمونه از مجموعه داده است که با استفاده از مؤلفه‌های اصلی جدید ترسیم شده است.

جدول ۱ اطلاعات مربوط به سه مجموعه دادهٔ مختلف در پژوهش را نشان می‌دهد. هر مجموعه داده شامل تعداد کل نمونه‌ها، نمونه‌های مثبت (تقلبی) و منفی (غیرتقلبی)، تعداد ویژگی‌ها و توضیحاتی دربارهٔ محتوای مجموعه داده است. به‌طور خاص، ویژگی‌های هر مجموعه داده شامل اطلاعاتی مانند تاریخ، مبلغ ادعا، وضعیت بیمه‌نامه و جزئیات خودرو است که برای تحلیل تقلب و تشخیص آن استفاده شده‌اند.

این مجموعه داده‌ها اطلاعاتی برای تشخیص تقلب در درخواست‌های بیمهٔ وسایل نقلیه فراهم می‌کنند. مجموعه دادهٔ Auto insurance شامل ویژگی‌های وسایل نقلیه، تصادفات و بیمه‌نامه‌ها و هدف آن شناسایی تقلب در درخواست‌های بیمه‌ای با نرخ تقلب ۰,۰۶ است. در مجموعه دادهٔ Vehicle insurance، نرخ تقلب ۰,۲۵ است، این مجموعه داده شامل ویژگی‌هایی مانند سن بیمه‌شده و مدت‌زمان عضویت در بیمه است. در Car insurance، با نرخ تقلب ۲۷ درصد، ویژگی‌های اقتصادی و خانوادگی بیمه‌شدگان بررسی می‌شود.

با استفاده از الگوریتم خوشه‌بندی K-means، ژنتیک ساده، خوشه‌بندی سلسله‌مراتبی و روش ارائه‌شده مجموعه داده‌ها را به دو خوشه تقسیم و سپس آن‌ها را مقایسه می‌کنیم. در روش پیشنهادی با نرخ جهش ۴۰ درصد و پس از تولید ۲ هزار کروموزوم و با جامعهٔ اولیه ۱۰۰ نفر و تولید ۵ کروموزوم از بازهٔ مرکز خوشه‌های K-means همانند توضیحات بر روی مجموعه داده‌ها اعمال شد.

در جدول ۲ مشاهده می‌کنیم که روش پیشنهادی در هر سه مجموعه داده در معیار ARI (Adjusted Rand Index) عملکرد بهتری دارد. همچنین، در معیارهای Fowlkes–Mallows و NMI (Normalized Mutual Information)، این روش بهبودهایی نشان می‌دهد. با اینکه هیچ‌یک از این معیارها به داده‌های کلاس تقلب توجه خاصی ندارند، پس از خوشه‌بندی می‌توان دو حالت برچسب‌گذاری را برای خوشه‌بندی خود در نظر گرفت و حالتی با دقت بیشتر را انتخاب کرد و سپس با استفاده از روش‌های ارزیابی مسائل دسته‌بندی با تمرکز بر داده‌های تقلب آن را ارزیابی کرد. این روش ارزیابی با تمرکز بر داده‌های تقلب برای مقایسه عملکرد مدل‌ها به کار می‌رود. همچنین، معیار F1 تعریف‌شده روی داده‌های تقلب نشان‌دهندهٔ عملکرد بهتر روش ارائه‌شده در مقایسه با دیگر روش‌هاست. نتایج حاصل را می‌توانید در جدول ۳ بررسی کنید.



شکل ۲. مصورسازی مجموعه داده‌ها با کمک کاهش بعد (نقاط زرد نمونه‌های مثبت (تقلبی) و سایر نقاط نمونه‌های غیرتقلبی هستند): (الف) مجموعه دادهٔ Auto Insurance؛ (ب) مجموعه دادهٔ Car Insurance؛ (ج) مجموعه دادهٔ Vehicle Insurance

Figure 2. Visualization of datasets using dimension reduction (The yellow points are positive (fraud) samples and the other points are normal samples): (a) Auto Insurance dataset; (b) Car Insurance dataset; (c) Vehicle Insurance dataset

جدول ۱. مشخصات سه مجموعه دادهٔ استفاده‌شده

Table 1. Information of the three datasets used

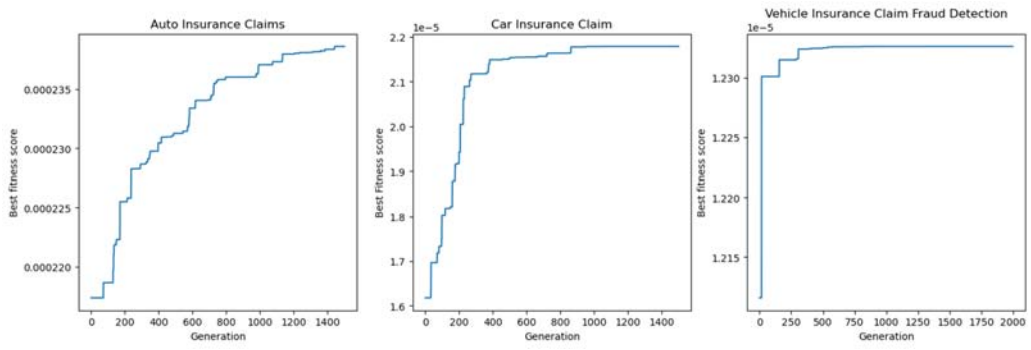
مجموعه داده Dataset	نمونه Sample	نمونه‌های منفی Negative samples	نمونه‌های مثبت Positive samples	ویژگی Feature	توضیحات Description
Auto insurance	1000	753	247	39	شامل تاریخ، مبلغ ادعا و وضعیت بیمه‌نامه است.
Car insurance	10302	7556	2746	27	شامل مشخصات مشتری، جزئیات خودرو و نوع حادثه است.
Vehicle insurance	15420	14497	923	32	شامل ویژگی‌هایی مانند مبلغ ادعا و مشخصات بیمه‌نامه است.

جدول ۲. عملکرد روش ارائه‌شده با استفاده از معیارهای آماری ارزیابی مسائل خوشه‌بندی
Table 2. The performance of the proposed method is evaluated using statistical metrics for clustering problems.

خوشه‌بندی سلسله‌مراتبی Hierarchical clustering	ژنتیک	K-Means	ژنتیک + K-Means	معیار Metric	مجموعه داده Dataset
0.1438	0.0988	0.1426	0.0391	Silhouette	Auto insurance
-0.0672	0.0130	-0.0681	0.0269	ARI	
0.0329	0.0023	0.0358	0.0028	NMI	
0.6398	0.7782	0.6351	0.7090	Fowlkes–Mallows	
137.22	12.394	137.96	34.657	Calinski–Harabasz	
0.2083	0.1470	0.2137	0.2091	Silhouette	Car insurance
0.0108	0.0100	0.0094	0.0136	ARI	
0.0066	0.0020	0.0063	0.0094	NMI	
0.5576	0.7696	0.5565	0.5586	Fowlkes–Mallows	
2978.74	108.87	3098.97	2991.93	Calinski–Harabasz	
0.0443	0.5524	0.0590	0.0563	Silhouette	Vehicle insurance
-0.0270	0.0005	0.0002	0.0016	ARI	
0.0065	0.0000	0.0003	0.0010	NMI	
0.7256	0.9393	0.6662	0.6677	Fowlkes–Mallows	
677.91	286.59	898.10	850.97	Calinski–Harabasz	

جدول ۳. نتایج حاصل از اجرای روش‌های گوناگون خوشه‌بندی در مقایسه با روش پیشنهادی با تمرکز بر داده‌های با برچسب مثبت
Table3. The results of applying various clustering methods, in comparison with the proposed method, with a focus on positively labeled data.

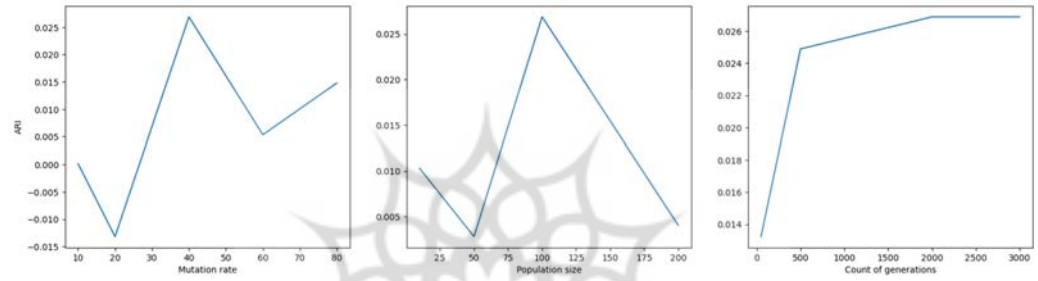
خوشه‌بندی سلسله‌مراتبی Hierarchical clustering	ژنتیک	K-Means	ژنتیک + K-Means	معیار Metric	مجموعه‌داده Dataset
0.09	0.36	0.09	0.31	صحت (Precision)	Auto insurance
0.07	0.03	0.07	0.15	بازخوانی (Recall)	
0.08	0.06	0.08	0.20	F1	
0.60	0.75	0.60	0.70	دقت (Accuracy)	
0.31	0.39	0.31	0.32	صحت (Precision)	Car insurance
0.55	0.03	0.55	0.57	بازخوانی (Recall)	
0.40	0.05	0.40	0.41	F1	
0.55	0.73	0.55	0.56	دقت (Accuracy)	
0.03	0.07	0.06	0.07	صحت (Precision)	Vehicle insurance
0.15	0.00	0.52	0.53	بازخوانی (Recall)	
0.05	0.01	0.11	0.12	F1	
0.69	0.94	0.51	0.53	دقت (Accuracy)	



(ج) (ب) (الف)

شکل ۳. نمودار همگرایی روش ارائه‌شده براساس نسبت جمعیت به بهترین امتیاز برازندگی: (الف) مجموعه داده Auto Insurance؛ (ب) مجموعه داده Car Insur-ance؛ (ج) مجموعه دادهٔ Vehicle Insurance

Figure 3. Convergence diagram of the presented method based on the ratio of the generation to the best fitness score: (a) Auto Insurance dataset; (b) Car Insurance dataset; (c) Vehicle Insurance dataset



(ج) (ب) (الف)

شکل ۴. نتیجهٔ تغییر پارامترهای روش ارائه‌شده در عملکرد مدل روی مجموعه دادهٔ Auto Insurance؛ (الف) براساس تعداد نسل؛ (ب) اندازهٔ جمعیت؛ (ج) نرخ جهش.

Figure 4. The impact of changing the parameters of the proposed method on the model's performance on Auto Insurance data-set; (a) Based on Count of generation; (b) Population size; (c) Mutation rate.

در مجموعه دادهٔ اول، روش پیشنهادی توانسته است F1 را به میزان ۱۲٪ و دقت کل را ۱٪ نسبت به K-Means افزایش دهد. در مجموعه دادهٔ دوم، به‌دلیل غیرخطی بودن داده‌ها، K-Means عملکرد ضعیفی داشته است، زیرا خوشه‌بندی با دو حالت نمی‌تواند به‌خوبی کلاس‌ها را تفکیک کند. روش خوشه‌بندی سلسله‌مراتبی نیز به‌دلیل تفکیک بهتر داده‌های میانی، عملکرد بهتری نسبت به K-Means داشته است. باوجوداین، روش پیشنهادی توانسته در معیارهای F1، دقت (Accuracy)، صحت (Precision) ۱٪ و بازخوانی (Recall) ۲٪ بهبود داشته باشد. در مجموعه دادهٔ سوم نیز، همانند مجموعه دادهٔ دوم، بهبود محسوسی در تمامی معیارها نسبت به K-Means مشاهده شده است. همچنین، روش پیشنهادی در هر سه مجموعه داده، عملکرد بهتری نسبت به روش ژنتیک عادی داشته است.

در مورد همگرایی روش ارائه‌شده نمودار همگرایی سه مجموعه داده به جواب بهینه در شکل ۳ قابل مشاهده است. در هر جمعیت بهترین برازندگی بین کروموزوم‌های حال حاضر جمعیت پیدا می‌شود و جواب بهینه در آن لحظه در نظر گرفته می‌شود. مشاهده می‌شود که به‌مرور و با افزایش جمعیت به جواب بهتری می‌رسیم و از نقطه‌ای به بعد بهبودی رخ نداده و مدل به جوابی بهینه همگرا شده است.

برای بهینه‌سازی روش پیشنهادی، پارامترهایی که باید بادقت تنظیم شوند، شامل نرخ جهش، اندازهٔ جمعیت اولیه و تعداد نسل‌ها هستند. این پارامترها نقش بسیار مهمی در یافتن راه‌حل بهینه دارند.

نرخ جهش تأثیر زیادی بر عملکرد الگوریتم و معیار ARI دارد. افزایش آن تا ۴۰ درصد باعث بهبود ARI و تنوع کروموزوم‌ها می‌شود، اما جهش بالاتر از ۴۰ درصد کارایی را کاهش می‌دهد و مانع همگرایی مناسب می‌شود (شکل ۴). نرخ‌های جهش پایین نیز می‌تواند الگوریتم را در کمینهٔ موضعی بیندازد. اندازهٔ جمعیت اولیه نیز اهمیت دارد؛ جمعیت کوچک جست‌وجو را محدود می‌کند و جمعیت بزرگ‌تر از ۱۰۰ کروموزوم، هزینهٔ محاسباتی را افزایش می‌دهد و همگرایی را کند می‌کند که اندازهٔ بهینهٔ جمعیت حدود ۱۰۰ کروموزوم است. تعداد نسل‌های بیشتر نیز به همگرایی کمک می‌کند و الگوریتم در ۲۰۰۰ نسل به نتایج خوبی می‌رسد، اما افزایش بیشتر تأثیر چندانی ندارد.

به‌طور کلی، تنظیم دقیق پارامترها برای بهینه‌سازی عملکرد الگوریتم ژنتیک ضروری است، هرچند ممکن است این پارامترها برای مسائل مختلف نیاز به تغییر داشته باشند.

جمع‌بندی و پیشنهادها

این پژوهش با استفاده از الگوریتم ژنتیک، روشی برای یافتن مراکز خوشه‌ها در مسائل خوشه‌بندی ارائه می‌دهد. این روش با ترکیب جواب‌های مختلف، جواب‌های تصادفی و معیارهای ارزیابی چندگانه، به بهبود نتایج خوشه‌بندی کمک می‌کند. همچنین، وارد کردن احتمالات و تصادفی بودن به الگوریتم، امکان بررسی گسترده‌تر مراکز خوشه‌ها را فراهم و از گیر افتادن در نقاط محدود جلوگیری می‌کند، که موجب بهبود عملکرد الگوریتم می‌شود.

همان‌طور که در بخش نتایج مشاهده شد، این روش می‌تواند بهتر از سایر روش‌های معروف مانند K-means و خوشه‌بندی سلسله‌مراتبی عمل کند و همچنین می‌تواند ساختاربندی درست و قابل قبولی برای خوشه‌بندی داده‌ها با استفاده از الگوریتم ژنتیک ارائه دهد.

همچنین پیشنهاد می‌شود با بهره بردن و ترکیب سایر معیارهای ارزیابی رایج در مسئلهٔ خوشه‌بندی و سایر امتیازها و همچنین ترکیب و استفاده از نتایج خروجی چندین مدل همانند و سایر روش‌های مبتنی بر مرکز خوشه می‌تواند نتایج و خروجی جالب‌تری را نیز به ارمغان بیاورد. همچنین می‌توان تعداد مراکز خوشه را نیز به کروموزوم‌های الگوریتم اضافه کرد و آن را نیز به‌صورت پارامتر در نظر گرفت.

مشارکت نویسندگان

مهدی قطعی و بهنام یوسفی مهر: ارائهٔ مفهوم و طرح پژوهش، بازنگری محتوای کیفی پژوهش؛ بهنام یوسفی مهر، سینا مرادی و یاسمین تفکر: جمع‌آوری، تحلیل و تفسیر داده‌ها؛ ساجد توکلی، سینا مرادی و یاسمین تفکر: پیش‌نویس مقاله؛ بهنام یوسفی مهر و ساجد توکلی: تحلیل، تفسیر و تحلیل آماری داده‌ها؛ مهدی قطعی: اخذ حمایت‌های مالی، اداری و فنی پژوهش.

تشکر و قدردانی

نویسندگان از نظرات داوران که باعث ارتقای کیفیت مقاله می‌شوند، کمال تشکر را دارند.

تعارض منافع

نویسندگان اعلام می‌کنند که هیچ تضاد منافعی در خصوص انتشار تحقیق ثبت‌شده وجود ندارد. علاوه‌براین، موارد اخلاقی از جمله سرقت ادبی، رضایت آگاهانه، رفتار نادرست، جعل و/یا جعل داده‌ها، انتشار مضاعف و یا سوءرفتار به‌طور کامل ازسوی نویسندگان رعایت شده است.

یادداشت ناشر

ناشر نشریهٔ *پژوهشنامه بیمه* با توجه به مرزهای حقوقی در نقشه‌های منتشرشده بی‌طرف باقی می‌ماند.

منابع

Ahmadlou, Y.; Pourebrahimi, A.; Tanha, J.; Rajabzadeh, A., (2023). Presenting a hybrid model for identifying claims of suspicious damages in agricultural insurance. J. Insur. Res., 12(1): 63-78. **(16 pages)** [In Persian].

Ahmed, M.; Seraj, R.; Islam, S.M.S., (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electron, 9(8): 1295.

Aibinu, A.M.; Salau, H.B.; Rahman, N.A.; Nwohu, M.N.; Akachukwu, C.M., (2016). A novel clustering based genetic algorithm for route optimization. Eng. Sci. Technol. Int. J., 19(4): 2022-2034. **(12 Pages)**

Babaie, S.S.; Omid Mahdi, E.E.; Firoozan, T., (2016). A Novel Combined Approach of k-Means and Genetic Algorithm to Cluster Cultural Goods in Household Budget. In Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015 Springer India.

Bhatia, S., (2014). New improved technique for initial cluster centers of K means clustering using Genetic Algorithm. In International Conference for Convergence for Technology-2014. IEEE.

Chang, D.X.; Zhang, X.D.; Zheng, C.W., (2009). A genetic algorithm with gene rearrangement for K-means clustering. Pattern Recognition, 42(7): 1210-1222 **(12 Pages)**.

Chiang, S.; Chu, S.C.; Hsin, Y.C.; Wang, M.H., (2006). Genetic distance measure for K-modes algorithm. Int. J. Innovative Comput. Inf. Control, 2(1): 33-40 **(7 Pages)**.

De Falco, I.; Della Cioppa, A.; Tarantino, E., (2002). Mutation-based genetic algorithm: performance evaluation. Appl. Soft Comput., 1(4): 285-299 **(14 Pages)**.

Hruschka, E.R.; Campello, R.J.; Freitas, A.A., (2009). A survey of evolutionary algorithms for clustering. IEEE Trans. Syst. Man, Cybern, Part C (applications and reviews), 39(2): 133-155 **(22 Pages)**.

Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J., (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Inf. Sci., 622. 178-210 **(32 Pages)**.

Jain, A.K., (2010). Data clustering: 50 years beyond K-means. Pattern Recognit. Let., 31(8): 651-666 **(15 Pages)**.

Jebari, K.; Madiafi, M., (2013). Selection methods for genetic algorithms. Int. J. Emerging Sci., 3(4): 333-344 **(11 Pages)**.

Katoch, S.; Chauhan, S.S.; Kumar, V., (2021). A review on genetic algorithm: Past, present, and future. Multimedia tools Appl., 80: 8091-8126 **(35 Pages)**.

Kudova, P., (2007). Clustering genetic algorithm. In 18th International Workshop on Database and Expert Systems Applications (DEXA 2007). IEEE.

Lu, Z.; Zhang, K.; He, J.; Niu, Y., (2016). Applying k-means clustering and genetic algorithm for solving mtsp. In Bio-inspired Computing–Theories and Applications: 11th International Conference, BIC-TA 2016, Xi'an, China, 2017, Revised Selected Papers, Part II 11. Springer Singapore.

Maulik, U.; Bandyopadhyay, S., (2000). Genetic algorithm-based clustering technique. Pattern Recognit., 33(9): 1455-1465 **(10 Pages)**.

Poikolainen, I.; Neri, F.; Caraffini, F., (2015). Cluster-based population initialization for differential evolution frameworks. Inf. Sci., 297: 216-235 **(19 Pages)**.

Rahman, M.A.; Islam, M.Z., (2014). A hybrid clustering technique combining a novel genetic algorithm with K-Means. Knowl. Based Syst., 71: 345-365 **(20 Pages)**.

Roy, D.K.; Sharma, L.K., (2010). Genetic k-means clustering algorithm for mixed numeric and categorical data sets. Int. J. Artif. Intell. Appl., 1(2): 23-28 **(5 Pages)**.

Seidi Aghil Abadi, Z.; Sehhat, S.; Salehi, R., (2017). Investigation and analysis of fraudulent factors in the third-party civil liability car Insurance (Third-party insurance-physical damage). J. Insur. Res., 7(1): 13-26. **(14 pages)** [In Persian].

Shetty, P.; Singh, S., (2021). Hierarchical clustering: A survey. Int. J. Appl. Res., 7(4): 178-181 **(3 Pages)**.

Singh, A.; Yadav, A.; Rana, A., (2013). K-means with three different distance metrics. Int. J. Comput. Appl., 67(10): 13-17. **(5 pages)**

Sonia, S.; Rai, S., (2012). Genetic k-means algorithm - implementation and analysis. Int. J. Recent Tech and Eng., 1(2): 1-4 **(4 pages).**

Tajaddodi Nodehi, M.; Hosseini Khatibani, S.; Yazdinejad, M.; Zolfi, S., (2024). Predicting people's health insurance costs using machine learning and ensemble learning methods. J. Insur. Res., 13(1): 1-14 **(14 pages)** [In Persian].

Yong, Y.; Xin cheng, G., (2012, July). A new minority kind of sample sampling method based on genetic algorithm and K-means cluster. In 2012 7th International Conference on Computer Science & Education (ICCSE). IEEE.

Yousefimehr, B.; Ghatee, M., (2025). A distribution-preserving method for resampling combined with LightGBM-LSTM for sequence-wise fraud detection in credit card transactions. Expert Syst. Appl., 262: 125661.

Zainuddin, F.; Abd Samad, M.F., (2020). A review of crossover methods and problem representation of genetic algorithm in recent engineering applications. Indones. J. Electr. Eng. Comput. Sci., 19(3): 759-769. **(11 pages)**

معرفی نویسندگان

AUTHOR(S) BIOSKETCHES

بهنام یوسفی مهر، دانشجوی دکتری گروه علوم کامپیوتر، دانشکده علوم ریاضی و کامپیوتر، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)، تهران، ایران

- Email: *behnam.y2010@aut.ac.ir*
- ORCID: 0009-0003-0954-635X
- Homepage: https://math.aut.ac.ir/

مهدی قطعی، استاد گروه علوم کامپیوتر، دانشکده علوم ریاضی و کامپیوتر، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)، تهران، ایران

- Email: *ghatee@aut.ac.ir*
- ORCID: 0000-0002-9558-8286
- Homepage: https://aut.ac.ir/cv/2174/%D9%85%D9%87%D8%AF%DB%8C-%D9%82%D8%B7%D8%B9%DB%8C?slc_lang=fa&&cv=2174&mod=scv

سینا مرادی، دانشجوی کارشناسی گروه علوم کامپیوتر، دانشکده علوم ریاضی و کامپیوتر، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)، تهران، ایران

- Email: *sina.moradi@aut.ac.ir*
- ORCID: 0009-0008-1612-7627
- Homepage: https://math.aut.ac.ir/

یاسمین تفکر، دانشجوی کارشناسی گروه علوم کامپیوتر، دانشکده علوم ریاضی و کامپیوتر، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)، تهران، ایران

- Email: *Yasi.tafakor@aut.ac.ir*
- ORCID: 0009-0001-7105-6277
- Homepage: https://math.aut.ac.ir/

ساجد توکلی، دانشجوی کارشناسی ارشد، گروه علوم کامپیوتر، دانشکده علوم ریاضی و کامپیوتر، دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)، تهران، ایران

- Email: *sajedtavakoli@aut.ac.ir*
- ORCID: 0009-0009-8548-2347
- Homepage: https://math.aut.ac.ir/

HOW TO CITE THIS ARTICLE

Yousefimehr, B., Ghatee, M., Moradi, S., Tafakor, T, Tavakoli, S., (2025). Detecting car insurance fraud using improved clustering with genetic algorithm. J. Insur. Res., 14(2): 109-118.

DOI: 10.22056/ijir.2025.02.02

URL: https://ijir.irc.ac.ir/article_160337.html