Industrial Management Journal

Print ISSN: 2008-5885

Online ISSN: 2423-5369

Developing a New Classification Method Based on a Hybrid Machine Learning and Multi Criteria Decision Making Approach

Mahdi Homayounfar
Assistant Prof., Department of Industrial Management, Faculty of Management and Accounting, Rasht Branch, Islamic Azad University, Rasht, Iran. E-mail: homayounfar@iaurasht.ac.ir

Amir Daneshvar
*Corresponding Author, Assistant Prof., Department of Industrial Management, Faculty of Management, Electronic Branch, Islamic Azad University, Tehran, Iran. E-mail: daneshvar.amir@gmail.com

Bijan Nahavandi
Assistant Prof., Department of Industrial Management, Faculty of Management and Economics, Science and Research Branch, Islamic Azad University, Tehran, Iran. E-mail: bnahavandi@gmail.com

Fariba Fallah
MSc., Department of Information Technology Management, Faculty of Management, Electronic Branch, Islamic Azad University, Tehran, Iran. E-mail: faribafallah94@gmail.com

Abstract

Objective: According to the capability of analytical network process (ANP) in analysis of different dependencies and feedback relationships among elements of a decision problem, the current research aims to develop an ANP based method for the benchmark classification problems. Since the essential limitation of ANP is the increase of inconsistency in judgment of decision makers along with increase in problem dimensions, genetic algorithm is used to optimize ANP parameters and improve classification accuracy.

Methods: Considering the objective, this study is a developmental research and in term of data analysis, it's a quantitative and mathematical modeling one. In this research, first a multi criteria decision making problem is developed based on ANP and in form of a classification problem and then the unknown parameters of a super matrix were calculated by machine learning methods. Next, the most proper values of these parameters which include thresholds of each class and the applied coefficients in the super matrix are estimated based on sample's benchmarks or data.The following processes have been conducted through a genetic algorithm. Finally, in order to validate the proposed method, its performance is compared to some frequently used classification methods in the reviewed literature.

Results: The results indicate the very competitive performance of the proposed method compared to known machine learning methods.

Conclusion: Multi-criteria Decision Making Methods (MCDM) are usually used for ranking purposes, however little attention has been paid to their high capabilities. In this paper ANP in combination with genetic algorithm demonstrated an efficient and suitable method in the field of data classification

Keywords: Classification, Analytical Network Process, Machine learning, Genetic Algorithm.

Citation: Homayounfar, M., Daneshvar, A., Nahavandi, B., & Fallah, F. (2019). Developing a New Classification Method Based on a Hybrid Machine Learning and Multi Criteria Decision Making Approach. *Industrial Management Journal,* 11(4), 675-692. (*in Persian*)

Industrial Management Journal, 2019, Vol. 11, No.4, pp. 675-692

DOI: 10.22059/imj.2019.280023.1007586

Received: April 28, 2019; Accepted: August 24, 2019

مدیریت صنعتی
شاپا چاپی: ۵۸۸۵-۲۰۰۸
شاپا الکترونیکی: ۵۳۶۹-۲۴۲۳

ارائه روش طبقه‌بندی جدید با استفاده از رویکرد ترکیبی یادگیری ماشین و تصمیم‌گیری چندمعیاره

مهدی همایون‌فر
استادیار، گروه مدیریت صنعتی، دانشکده مدیریت و حسابداری، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران. رایانامه: homayounfar@iaurasht.ac.ir

امیر دانشور
* نویسنده مسئول، استادیار، گروه مدیریت فناوری اطلاعات، دانشکده مدیریت، واحد الکترونیکی، دانشگاه آزاد اسلامی، تهران، ایران. رایانامه: daneshvar.amir@gmail.com

بیژن نهاوندی
استادیار، گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات تهران، دانشگاه آزاد اسلامی، تهران، ایران. رایانامه: bnahavandi@gmail.com

فریبا فلاح
کارشناس ارشد، گروه مدیریت فناوری اطلاعات، دانشکده مدیریت، واحد الکترونیکی، دانشگاه آزاد اسلامی، تهران، ایران. رایانامه: faribafallah94@gmail.com

چکیده

هدف: از آنجا که در مسائل طبقه‌بندی به تحلیل انواع وابستگی‌ها و روابط بازخوردی میان معیارهای یک مسئله کمتر پرداخته شده است و با توجه به قابلیت فرایند تحلیل شبکه‌ای (ANP) در مدل‌سازی روابط متقابل بین معیارها، هدف این پژوهش ارائه روشی مبتنی بر ANP برای مسائل طبقه‌بندی است. محدودیت اساسی ANP، افزایش ناسازگاری قضاوت تصمیم‌گیرندگان همراه با افزایش ابعاد مسئله است، از این رو به‌منظور بهینه‌سازی پارامترهای مسئله و افزایش صحت طبقه‌بندی، از الگوریتم ژنتیک استفاده خواهد شد.

روش: پژوهش حاضر از نظر هدف، توسعه‌ای و از نظر روش تحلیل داده‌ها کمّی و از نوع مدل‌سازی ریاضی است. در این پژوهش، ابتدا مسئله طبقه‌بندی داده‌ها با در نظر داشتن روابط متقابل معیارها در قالب روش تصمیم‌گیری چندمعیاره ANP تبیین شد. در ادامه، مقدار پارامترهای مسئله، شامل وزن معیارها و آستانه‌های هر کلاس به‌کمک الگوریتم ژنتیک از سوپرماتریس برآورد شد و در نهایت، برای ارزیابی روش پیشنهادی و عملکرد آن، نتیجه با روش‌های پرکاربرد طبقه‌بندی مقایسه شد.

یافته‌ها: نتایج پژوهش‌های مقایسه‌ای روی دیتاست‌های اعتباری با ابعاد مختلف، قابلیت رقابتی بسیار خوب روش پیشنهادی را در مقایسه با روش‌های شناخته‌شده یادگیری ماشینی نشان داد.

نتیجه‌گیری: روش‌های تصمیم‌گیری چندمعیاره، اغلب برای رتبه‌بندی استفاده شده‌اند، این در حالی است که به قابلیت بسیار خوب این روش‌ها در طبقه‌بندی داده‌ها کمتر توجه شده است. فرایند تحلیل شبکه‌ای در ترکیب با الگوریتم ژنتیک، روشی کارا و مناسب در حوزه طبقه‌بندی داده‌ها را به نمایش می‌گذارد.

کلیدواژه‌ها: فرایند تحلیل شبکه‌ای، طبقه‌بندی، یادگیری ماشین، الگوریتم ژنتیک.

استناد: همایون‌فر، مهدی؛ دانشور، امیر؛ نهاوندی، بیژن؛ فلاح، فریبا (۱۳۹۸). ارائه روش طبقه‌بندی جدید با استفاده از رویکرد ترکیبی یادگیری ماشین و تصمیم‌گیری چندمعیاره. *مدیریت صنعتی*، ۱۱(۴)، ۶۷۵- ۶۹۲.

مدیریت صنعتی، ۱۳۹۸، دوره ۱۱، شماره ۴، صص. ۶۷۵- ۶۹۲
DOI: 10.22059/imj.2019.280023.1007586
دریافت: ۱۳۹۸/۰۲/۰۸، پذیرش: ۱۳۹۸/۰۶/۰۲

مقدمه

مسائل تصمیم‌گیری بسیار پیچیده شده‌اند و دیگر نمی‌توان به‌راحتی فرض مستقل بودن معیارها را در نظر گرفت. از این رو کاربرد فرایند تحلیل سلسله‌مراتبی (AHP) به‌عنوان یکی از روش‌های پرکاربرد در محاسبه وزن معیارها که یکی از فرض‌های اساسی آن، عدم وابستگی میان معیارها است، با مشکل مواجه شده است. به‌منظور حل این مسئله، ساعتی[۱] (۱۹۹۶) برای تصمیم‌گیری در مسائل با معیارهای وابسته روشی با عنوان فرایند تحلیل شبکه‌ای (ANP)[۲] ارائه کرد. ANP رابطه سلسله‌مراتبی را با یک نمایش شبکه‌ای برای نشان دادن وابستگی بین معیارها جایگزین کرده و با استفاده از سوپرماتریس مشکل وابستگی و بازخورد میان اجزا را حل کرده است (نیمیرا و ساعتی[۳]، ۲۰۰۴).

از آنجا که در بسیاری از مسائل واقعی، پیش‌فرض وابستگی متقابل معیارها تضمین نمی‌شود (وانگ، وانگ و کلیر[۴]، ۱۹۹۸)، ANP به‌دلیل مجاز دانستن این روابط متقابل توجه زیادی را به خود جلب کرده است. پژوهش‌های بسیاری بر اساس ANP صورت گرفته که از آن جمله می‌توان به برنامه‌ریزی آمیخته محصول برای ساخت شبه رسانا (چانگ، لی و پیم[۵]، ۲۰۰۵)، انتخاب پروژه‌های سیستم اطلاعاتی (لی و کیم[۶]، ۲۰۰۰)، مدیریت پروژه (میدو پرسلی[۷]، ۲۰۰۲ و میدو سرکیس[۸]، ۱۹۹۹)، ساخت یک مدل بحران مالی (نیمیرا و ساعتی، ۲۰۰۴)، مدیریت استراتژیک برای مدیریت جنگل (ولف‌اسلنر، واکیک و لکسر[۹]، ۲۰۰۵)، عملیات لجستیک معکوس برای کامپیوترهای در پایان چرخه عمر (راوی، شانکار و تیواری[۱۰]، ۲۰۰۵)، ارزیابی سیستم‌های ضبط ویدیوی دیجیتالی (چانگ و همکاران، ۲۰۰۵)، تحلیل SWOT (یوکسل و دادویرن[۱۱]، ۲۰۰۷)، انتخاب حالت حمل‌ونقل (توزکایا و اونوت[۱۲]، ۲۰۰۸) و ارزش‌گذاری اراضی صنعتی شهری (آراگونس بلتان، آزنار، فریس اوناته و گارسیا ملون[۱۳]، ۲۰۰۸) اشاره کرد. با وجود این و با توجه به پژوهش‌های صورت‌گرفته، ANP تاکنون برای طبقه‌بندی الگو به‌کار برده نشده است.

طبقه‌بندی الگو[۱۴] یکی از مسائل تصمیم‌گیری چندمعیاره است که فضای الگوها را به کلاس‌های مختلف تقسیم‌بندی کرده و هر الگوی ورودی را به یک کلاس اختصاص می‌دهد. در سال‌های اخیر، روش‌های متعددی برای طبقه‌بندی ارائه شده‌اند (انگای، هو، ونگ، چن و سان[۱۵]، ۲۰۱۱) که از میان آنها می‌توان به شبکه‌های بیزین، نزدیک‌ترین همسایه، شبکه‌های عصبی، درخت تصمیم، مدل‌های رگرسیونی و الگوریتم‌های تکاملی اشاره کرد (نیکام[۱۶]، ۲۰۱۵). زرین صدف و دانشور (۱۳۹۵) در رابطه با روش‌های تصمیم‌گیری چندمعیاره که در مسائل طبقه‌بندی استفاده شده‌اند، روشی ارائه کردند که با یادگیری مقادیر پارامترهای روش ELECTRE TRI از داده‌های آموزشی با استفاده از الگوریتم بهینه‌سازی تراکم ذرات، آنها را در طبقه‌بندی موجودی‌های انبار به‌کار می‌برد. در پژوهش باکور[۱۷] (۲۰۱۸) روش ترکیبی تاپسیس و ویکور برای مسئله طبقه‌بندی بیماری‌های قلب و تیروئید به‌کار گرفته شده است. نتایج نشان‌دهنده عملکرد

1. Saaty
2. Analytical Network Process (ANP)
3. Niemira & Saaty
4. Wang, Wang & Klir
5. Chung, Lee & Pearn
6. Lee & Kim
7. Meade & Presley
8. Meade & Sarkis
9. Wolfslehner, Vacik & Lexer
10. Ravi, Shankar & Tiwari
11. Yüksel & Dagdeviren
12. Tuzkaya & Önüt
13. Aragonés-Beltrán, Aznar, Ferrís-Oñate & García-Melón
14. Pattern Classification
15. Ngai, Hu, Wong, Chen & Sun
16. Nikam
17. Baccour

بهتر روش ارائه‌شده در مقایسه با روش‌های موجود مانند ماشین بردار پیشتیبان، لجستیک رگرسیون، ماشین بردار پشتیبان با تابع کرنل خطی و شبکه عصبی بودند. در پژوهش کارتال، اوزتکین، گوناسکاران و کبی[1] (۲۰۱۶)، مسئله طبقه‌بندی آیتم‌های موجود در سه دسته A، B وC با ترکیب روش‌های یادگیری ماشین و روش‌های تصمیم‌گیری چند معیاره VIKOR، AHP و SAW بررسی شد که ابتدا روش‌های چندمعیاره برای طبقه‌بندی درست موجودی‌ها به‌کار گرفته شدند و سپس به‌منظور پیش‌بینی کلاس موجودی‌ها از روش‌های یادگیری ماشین شامل شبکه بیزین، شبکه عصبی و ماشین بردار پشتیبان استفاده شد. در مقاله کو، پنگ و وانگ[2] (۲۰۱۴) به اعتبارسنجی روش‌های طبقه‌بندی داده‌های مالی با استفاده از روش‌های تصمیم‌گیری چندمعیاره شامل تاپسیس، ویکور و تحلیل پوششی داده‌ها پرداخته شد. در این مقاله روش‌های K میانگین، الگوریتم پارتیشن‌بندی گراف، روش تکرار ـ تنصیف[3]، بر اساس شاخص‌های مختلف ارزیابی شدند. دانشور، زندیه و ناظمی (۱۳۹۴) روش جدیدی برای اعتبارسنجی مشتریان اعتباری بانک‌ها مبتنی بر مدل ELECTRE TRI ارائه کردند که در آن اطلاعات مورد نیاز برای تصمیم‌گیری را به‌جای استفاده از پرسش‌نامه از درون داده‌های تاریخی بر اساس تصمیم‌های قبلی تصمیم‌گیرندگان و با استفاده از الگوریتم‌های فراابتکاری استخراج کرده و آنها را به‌عنوان ورودی به مدل ELECTRE TRI داده و مشتریان را طبقه‌بندی کردند. دانشور، همایون‌فر و فرهمندنژاد (۱۳۹۸) با ترکیب الگوریتم K میانگین و تکنیک پرامتی، یک روش جدید خوشه‌بندی چندمعیاره ارائه دادند. در مطالعه آنها، پارامترهای مسئله، پروفایل‌های جداکننده خوشه‌ها هستند که برای بهینه‌سازی آنها از الگوریتم ژنتیک استفاده شده است. دانشور، همایون‌فر و اخوان (۱۳۹۸) در پژوهشی به توسعه یک روش طبقه‌بندی دیتاست‌های نامتوازن پرداختند. به‌دلیل پیچیدگی فرایند کسب اطلاعات از تصمیم‌گیرنده، در مطالعه آنها از الگوریتم فرا ابتکاری NSGA II برای استنتاج مقدار پارامترها استفاده شده است.

تا جایی که پژوهش‌های ما نشان می‌دهد، روش‌های طبقه‌بندی موجود در ادبیات پژوهش تاکنون وابستگی متقابل معیارها را بررسی نکرده‌اند و اگر بتوان این محدودیت را مرتفع کرد تصمیم‌گیری به شرایط واقعی بسیار نزدیک‌تر خواهد شد. از این رو، این مقاله طبقه‌بند مبتنی بر ANP را به‌منظور پرداختن به وابستگی متقابل معیارها توسعه می‌دهد. یکی از مشکل‌های استفاده از ANP، پیچیدگی تعریف سوپرماتریس است. انجام مقایسه‌های زوجی و تشکیل سوپرماتریس به دانش تخصصی و فرد خبره نیاز دارد که در ماتریس‌های مقایسه‌های زوجی بزرگ، اغلب ناسازگاری در قضاوت را به‌دنبال خواهد داشت. برای غلبه بر این مشکل، در این مقاله با استفاده از الگوریتم ژنتیک به تخمین درایه‌های سوپرماتریس و حدود آستانه پرداخته می‌شود. در هر تکرار، بردار وزن به‌دست‌آمده در ماتریس تصمیم موجود ضرب شده و ارزش هر آلترناتیو به دست می‌آید. مقدار ارزش محاسبه‌شده هر آلترناتیو با حدود آستانه کلاس‌ها مقایسه شده و بر این اساس تخصیص آلترناتیو به کلاس‌ها صورت می‌گیرد. در نهایت، تابع برازش به‌کاررفته عبارت است از حداکثرسازی صحت طبقه‌بندی[4] یعنی نسبت تعداد آلترناتیوهای درست طبقه‌بندی‌شده به تعداد کل آلترناتیوهای موجود.

1. Kartal, Oztekin, Gunasekaran & Cebi
2. Kou, Peng & Wang
3. Repeated-Bisection Method
4. Classification Accuracy (CA)

روش‌شناسی پژوهش

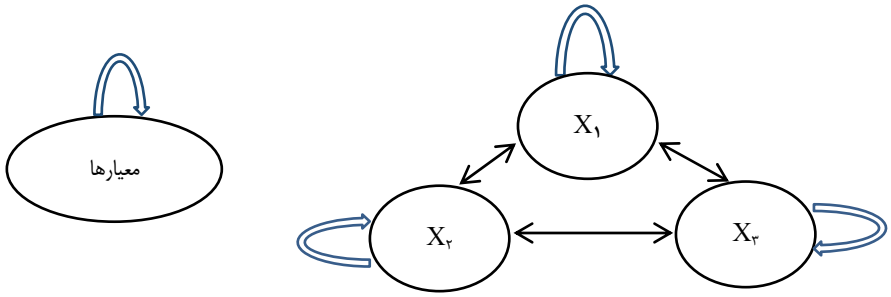
پژوهش حاضر از نظر هدف، توسعه‌ای و از نظر روش تحلیل داده‌ها کمی و از نوع مدل‌سازی ریاضی است.

فرایند تحلیل شبکه‌ای در مسائل طبقه‌بندی الگو

با افزایش پیچیدگی مسائل تصمیم‌گیری، دیگر به‌راحتی نمی‌توان فرض مستقل بودن معیارها را در نظر گرفت. بدین منظور برای تصمیم‌گیری در مسائلی با عناصر وابسته، روشی با عنوان فرایند تحلیل شبکه‌ای (ANP) معرفی شده است (نیمیرا و ساعتی، ۲۰۰۴) که ارتباط‌های پیچیده میان عناصر تصمیم را از طریق جایگزینی ساختار سلسله‌مراتبی با ساختار شبکه‌ای در نظر می‌گیرد و با استفاده از سوپرماتریس، مشکل وابستگی و بازخور میان اجزا را حل می‌کند (ساعتی[1]، ۲۰۰۱). مراحل پیاده‌سازی این روش عبارت است از:

۱. ساخت مدل و ساختاربندی مسئله: در این مرحله ساختار مسئله در قالب شبکه‌ای از روابط ترسیم می‌شود.

۲. تعیین مقایسه‌های زوجی و بردارهای اولویت: مشابه با AHP و ANP مقایسه‌های زوجی معیارها را با توجه به اهمیت نسبی آنها در مقایسه با معیار کنترلشان انجام می‌دهد. اهمیت نسبی هر معیار در مقایسه با معیار کنترل را می‌توان از ماتریس مقایسه زوجی به دست آورد. مقایسه‌های زوجی بیانگر روابط میان عناصر موجود در جزء هستند. تصمیم‌گیرندگان به یک سری از مقایسه‌های زوجی که در آنها دو معیار از حیث نحوه تأثیرگذاری بر معیار کنترل خود مقایسه می‌شوند، پاسخ می‌دهند (مید و سرکیس، ۱۹۹۹). سپس می‌توان از ماتریس مقایسه زوجی یک بردار اولویت نسبی را به دست آورد.



شکل ۱. نمایش شبکه‌ای متشکل از سه معیار

به‌طور مثال، در شکل ۱، C_{ij} بیانگر اهمیت نسبی x_i نسبت به x_j است. اگر C_{11} صفر نباشد، یک کمان را می‌توان از x_i به x_j رسم کرد، در غیر این صورت، بین x_i و x_j ارتباطی وجود ندارد. بردار اولویت نسبی $(C_{31}, C_{21}, C_{11})^T$ به‌ترتیب بیانگر اهمیت نسبی x_1، x_2 و x_3 نسبت به x_1 است و می‌توان آنها را از سه مقایسه زوجی میان x_i (یعنی x_1 به x_2، x_2 به x_3 و x_1 به x_3) به دست آورد. $(C_{32}, C_{22}, C_{12})^T$ و $(C_{33}, C_{23}, C_{13})^T$ را می‌توان به شیوه مشابه به دست آورد.

۳. تشکیل سوپرماتریس: سوپرماتریس مشابه با ماتریس احتمال انتقال در زنجیره مارکوف است (ساعتی، ۲۰۰۱). با جایگزین کردن بردارهای اولویت نسبی در ستون‌های مناسب یک سوپرماتریس، وزن اولویت سراسری هر

1. Saaty

معیار مسئله تصمیم‌گیری را می‌توان از سوپرماتریس محدودکننده‌ای که از ضرب سوپرماتریس در خودش تا زمان تثبیت ستون‌ها حاصل شده است، به دست آورد. این سوپرماتریس یک ماتریس مربع است. به‌طور مثال، سوپرماتریس متناظر با شکل ۱ را می‌توان به شرح ذیل تولید کرد:

رابطه ۱)
$$\begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix}$$

که در آن،

رابطه ۲)
$$C_{11} + C_{21} + C_{31} = 1$$

رابطه ۳)
$$C_{12} + C_{22} + C_{32} = 1$$

رابطه ۴)
$$C_{13} + C_{23} + C_{33} = 1$$

در این صورت، سوپرماتریس محدودکننده W^{2K+1} می‌تواند تعامل‌های میان اجزا را در بر گیرد و مقادیر موزون ثابت بلندمدت را نشان دهد (چانگ و همکاران، ۲۰۰۵) و K بیانگر یک عدد دلخواه بزرگ است. یافته‌ها حاکی از آن است که تمامی ستون‌های W^{2K+1} یکسان هستند. w_i نشانگر وزن معیار iام است که خود این معیار با x_i نشان داده می‌شود و W^{2K+1} به شرح ذیل مفروض است:

رابطه ۵)
$$W^{2K+1} = \begin{pmatrix} W_1 & W_1 & W_1 \\ W_2 & W_2 & W_2 \\ W_3 & W_3 & W_3 \end{pmatrix}$$

و W_1، W_2 و W_3 شرط ذیل را برآورده می‌کنند:

رابطه ۶)
$$\sum_{i=1}^{3} w_i = 1$$

به بیان دیگر، سوپرماتریس محدودکننده، وزن‌های W_1، W_2 و W_3 را به‌ترتیب برای معیارهای x_1، x_2 و x_3 مدل به دست می‌آورد.

از آنجا که بردارهای اولویت در سوپرماتریس توسط الگوریتم ژنتیک و نه از طریق مقایسه‌های زوجی توسط پرسش‌نامه به دست می‌آید، روش پیشنهادی مقایسه‌های زوجی را در مرحله ۲ الگوریتم ANP حذف می‌کند.

طبقه‌بند مبتنی بر تحلیل شبکه‌ای با استفاده از الگوریتم ژنتیک

در مسئله طبقه‌بندی مبتنی بر فرایند تحلیل شبکه‌ای، هدف تعیین کلاس مناسب برای هر الگوی n بعدی ورودی به‌صورت $x_i = (x_1, x_2, \ldots, x_n)$ است. با هر الگوی ورودی مانند یک آلترناتیو در ANP برخورد شده و با استفاده از وزن‌های به‌دست‌آمده از بخش قبل، ارزش آن آلترناتیو با استفاده از رابطه ۷ محاسبه می‌شود:

رابطه ۷)
$$u(x) = \sum_{i=1}^{n} w_i x_i$$

که در آن مقادیر x_i بین صفر و ۱ نرمال‌سازی شده‌اند. باید توجه کرد که x_iها به معیارهای مختلف یک آلترناتیو مربوط هستند که در مقیاس‌های مختلف اندازه‌گیری می‌شوند. از این رو، به‌منظور سازگاری قضاوت‌ها، برای نرمال‌سازی داده‌ها از رابطه ۸ استفاده می‌شود و در رابطه ۷ به‌جای پارامتر x_i از نرمال‌شده آن یعنی x'_i استفاده خواهد شد:

رابطه ۸)
$$x'_i = \frac{x_i - min_i}{max_i - min_i}$$

x'_i مقدار نرمال‌شده x_i و max_i حداکثر و min_i حداقل مقدار معیار iام است که از روی داده‌های آموزش و آزمایش قابل تشخیص هستند. با در نظر گرفتن x'_i مطابق با رابطه ۸ که نتیجه می‌دهد $0 \leq x'_i \leq 1$ و همچنین رابطه $\sum w_i = 1$، به‌راحتی تضمین می‌شود $0 \leq u(x) \leq 1$ و در آن x به‌صورت $x = (x'_1, x'_2, \ldots, x'_n)$ است. با مشخص شدن مقدار $u(x)$ برای هر آلترناتیو و مقایسه آن با حد آستانه کلاس‌ها، طبقه‌بندی آلترناتیوها صورت می‌گیرد (دامپوس و زوپونیدیس[1]، ۲۰۰۲).

برای نمونه در یک مسئله دوکلاسه، اگر ارزش آلترناتیو $u(x)$ کمتر از حد آستانه Th باشد، الگو به کلاس ۱ و در غیر این صورت به کلاس ۲ متعلق خواهد بود. برای تعمیم به مسائل mکلاسه، m −۱ حد آستانه تعریف می‌شود (مسام[2]، ۱۹۸۸):

رابطه ۹)
- اگر $u(x) < Th_1$ الگوی متعلق به کلاس ۱
- اگر $Th_{i-1} < u(x) < Th_i$ الگو متعلق به کلاس i
- اگر $u(x) > Th_{m-1}$ الگوی متعلق به کلاس m

چگونگی محاسبه پارامترهای یک مسئله طبقه‌بندی با استفاده از روش ANP از روی داده‌های الگو با استفاده از الگوریتم ژنتیک در ادامه نشان داده شده است. این داده‌ها شامل پارامترهای مربوط به درایه‌های سوپرماتریس و نقاط برش هستند. شبه کد الگوریتم به‌کاررفته در این پژوهش در شکل ۲ نشان شده است.

ورودی‌ها
- اندازه جمعیت (N_{pop})
- اندازه جمعیت حذف‌شده (N_{del})
- ماکزیمم تعداد تکرارها (Max_{It})
- نرخ تقاطع (c_r)
- نرخ جهش (m_r)

خروجی:
- جواب نزدیک بهینه با ماکزیمم دقت طبقه‌بندی (CA)

کدگذاری کروموزوم: ژن‌های متناظر با سوپرماتریس و ژن‌های متناظر با نقاط برش

$b_1 b_2 \ldots b_{n^2} b_{n^2+1} b_{n^2+2} \ldots b_{n^2+c-1}$

شروع GA

اندیس تکرارها را برابر ۱ قرار بده $GAIter = 1$

1. Doumpos & Zopounidis
2. Massam

جواب اولیه (P_0) را تولید کن: یک آرایه از اعداد تصادفی باینری با احتمال ۰/۵ و یک عدد تصادفی حقیقی بین صفر و ۱ تولید کن $Rand \in [0,1]$.

(While)تا وقتی که ($Itr \leq Max_{Itr}$) انجام بده.

تولید فرزندان:

ارزش برازش کروموزوم‌های جمعیت فعلی (P_t) را ارزیابی کن.

عملیات انتخاب را با استفاده از چرخه رولت انجام بده و کروموزوم والد را انتخاب کن.

والدین را در G_t ذخیره کن.

تقاطع را به‌کار بگیر:

تا رسیدن به مقدار (Max_{It})

$G_{t(j)}$ و $G_{t(j+1)}$ را با عملگر تقاطع تک‌نقطه‌ای ترکیب کن.

فرزندان را در P^1_t ذخیره کن.

جهش را به‌کار بگیر:

یک کروموزوم فرزند جدید با تغییر تصادفی روی کروموزوم انتخاب شده تولید کن. یک نرخ جهش، به تمامی موقعیت‌ها مرتبط است. برای موقعیت انتخابی، یک عدد تصادفی تولید کن.

فرزندان را در P^2_t ذخیره کن.

ارزیابی:

P_t را با P^1_t و P^2_t را ادغام و در P'_t ذخیره کن.

به‌صورت تصادفی تعداد N_{del} کروموزوم را از P'_t حذف کن.

به تعداد N_{del} تا از بهترین کروموزوم‌ها را به P'_t اضافه کن تا جمعیت نسل بعدی (P''_t) تشکیل شود.

کروموزوم‌ها را به‌صورت نزولی مرتب کن.

فرزند تولیدشده را مطابق با CA ارزیابی کن.

به تکرارها یکی اضافه کن $GAIter = GAIter + 1$

پایان حلقه While

شکل ۲. شبه کد الگوریتم به‌کاررفته در پژوهش

فرموله‌سازی مسئله با استفاده از الگوریتم ژنتیک

اجزای اصلی الگوریتم ژنتیک در حل این مسئله به شرح ذیل هستند:

نمایش جواب: برای ساختن طبقه‌بندی مبتنی بر ANP، طبقه‌بند پیشنهادی، GAها را به‌منظور تعیین درایه‌های سوپرماتریس و حدود آستانه کلاس‌ها به‌کار می‌گیرد. تعداد کل پارامترهای لازم برای کدگذاری کروموزوم $n^2 + c - 1$ است که n^2 پارامتر مربوط به درایه‌های سوپرماتریس شامل n معیار و $c - 1$ متناظر با تعداد حدود آستانه مربوط به تعریف C کلاس مدنظر است. طبقه‌بند پیشنهادی توسط کروموزوم باینری $b_1 b_2 \dots b_{n^2} b_{n^2+1} b_{n^2+2} \dots b_{n^2+c-1}$ برای یک مسئله C کلاسه نشان داده می‌شود؛ $b_1 b_2 \dots b_{n^2}$ ژن‌های متناظر با درایه‌های سوپرماتریس و $(0 \leq j \leq C - 1)$ b_{n^2+j} ژن‌های متناظر با حد آستانه t_j است. در جمعیت اولیه، به‌طور تصادفی و با احتمال ۰/۵ به هر ژن مقدار ۱ یا ۰ داده می‌شود. هر ژن کروموزم را می‌توان به‌طور مستقیم به‌عنوان یک مقدار حقیقی که بین ۰ تا ۱ تغییر می‌کند، رمزگشایی کرد. سپس مجموع اعداد تصادفی تولیدشده برای سطر ماتریس W محاسبه شده و اعداد هر سطر بر مجموع اعداد آن سطر تقسیم می‌شوند. با استفاده از این روش n^2 عنصر ماتریس W محاسبه می‌شود. سپس با به توان رساندن ماتریس W تا

زمانی که ماتریس W^{2K+1} هم‌گرا شود، ضرایب اهمیت ماتریس (مقادیر w_i) به ازای iها مشخص خواهد شد. به‌علاوه، در تعیین آستانه‌های مربوط به کلاس‌های مختلف، به‌طور مشخص آستانه مربوط به کلاس iام کوچک‌تر از آستانه مربوط به کلاس i − ۱ است. از طرف دیگر با توجه به اینکه $0 \leq u(x) \leq 1$، به‌منظور تولید ژن‌های مربوط به آستانه‌های کلاس‌ها در تولید جواب‌های اولیه، می‌توان C − ۱ عدد تصادفی بین صفر و ۱ تولید کرد و بعد از مرتب‌سازی این اعداد به‌صورت نزولی، آنها را به‌ترتیب برابر Th_1، Th_2، ... و Th_{m-1} قرار داد. همچنین، با توجه به اینکه در پژوهش حاضر عملگرهای تقاطع و جهش مورد استفاده روی اعداد باینری به‌کار گرفته خواهند شد، هر یک از اعداد رشته جواب به یک عدد اعشاری در نمایش باینری تبدیل می‌شوند. بدین منظور تعداد بیت‌های لازم برای نمایش دودویی هر یک از اعضای رشته جواب یعنی τ_3 با استفاده از رابطه $2^{\tau_3-1} < \tau_1 \times 10^{\tau_2} < 2^{\tau_3}$ مشخص می‌شوند که در آن دامنه تغییر، τ_2 دقت مد نظر (تعداد اعشار) و $\tau_3 = 10$ تعداد بیت‌های مورد استفاده برای نمایش اعداد رشته کروموزوم در نظر گرفته می‌شوند.

تابع برازش: تابع برازش الگوریتم ژنتیک برای مسئله طبقه‌بندی را می‌توان به‌صورت بیشینه‌سازی نسبت طبقه‌بندی‌های درست تعریف کرد. در پژوهش حاضر، صحت طبقه‌بندی[1] یعنی نسبت تعداد نمونه‌هایی که کلاس آنها به‌درستی پیش‌بینی شده به تعداد کل آلترناتیوهای طبقه‌بندی‌شده، به‌عنوان تابع برازش در نظر گرفته شده است. جدول ۱ ماتریس اغتشاش برای یک مسئله دوکلاسه را نشان می‌دهد و بر اساس آن صحت طبقه‌بندی مطابق با رابطه ۹ تعریف می‌شود.

جدول ۱. ماتریس اغتشاش برای مسئله طبقه‌بندی دوکلاسه

	درست (T)	غلط (F)
مثبت (P)	درست مثبت (TP)	غلط مثبت (FP)
منفی (N)	درست منفی (TN)	غلط منفی (FN)

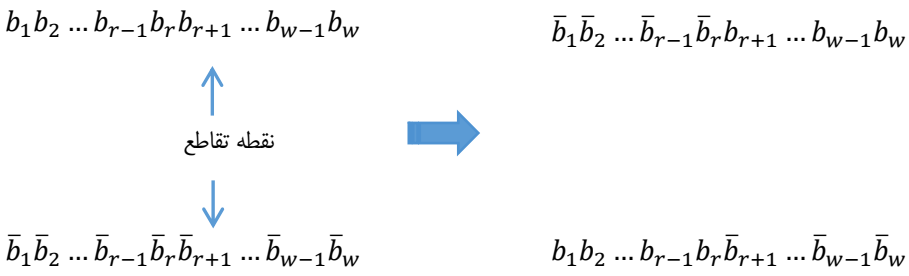
$$CA = \frac{TP + TN}{TP + TN + FP + FN}$$ رابطه ۱۰)

انتخاب اعضای جمعیت برای جهش یا انتقال به نسل بعد: بعد از تشکیل جمعیت (جواب‌های) اولیه و محاسبه تابع برازش برای هر یک از آنها، عملگرهای تولید مثل روی جواب‌های موجود اعمال می‌شوند تا جواب‌های جدید تولید شوند. بدین منظور، ابتدا جمعیت مد نظر برای انجام عملیات تولید مثل مشخص می‌شود و سپس عملیات تقاطع و جهش روی آنها صورت می‌گیرد تا فرزندان (جواب‌های) جدید ایجاد شوند. به‌منظور انتخاب والدین و جمعیتی که از یک تکرار به تکرار بعد منتقل می‌شود، از یک عملگر انتخاب نخبه‌گرا استفاده شده است. ابتدا، احتمال بقای اعضای جمعیت به‌عنوان یکی از پارامترهای الگوریتم در نظر گرفته شده و سپس، جمعیتی که برای تکرار بعد زنده می‌ماند مشخص می‌شود. به‌منظور حفظ نخبه‌گرایی، این جمعیت از بین اعضای جمعیت که برازش بهتری دارند انتخاب می‌شود.

1. Classification Acuuracy (CA)

به‌منظور حفظ تعداد اعضای جمعیت در هر تکرار، به‌میزان تفاوت بین تعداد اعضای جمعیت در هر تکرار و جمعیت زنده‌مانده، به تولید فرزندان جدید و انتقال آنها به نسل بعد نیاز است. این فرزندان در نتیجه عملگرهای آمیزش بین دو والد ایجاد می‌شوند. از این رو، در مرحله بعد باید والدینی که برای تولید فرزندان انتخاب می‌شوند مشخص شوند. سپس اعضای جمعیت زنده‌مانده بر حسب تابع برازش آنها به‌طور صعودی مرتب‌سازی می‌شوند. با توجه به ایده الگوریتم ژنتیک مبنی بر تکثیر بیشتر اعضای با تابع برازش بیشتر برای تکرارها یا نسل‌های بعد، احتمال انتخاب اعضای برابر با موقعیت آنها در کروموزوم تقسیم بر مجموع موقعیت‌های اعضای جمعیت خواهد بود. بعد از مشخص شدن احتمال انتخاب اعضا به‌عنوان والد، از عملگر تورنمنت برای انتخاب والدین هر فرزند جدید استفاده می‌شود. این عملگر، متناسب با احتمال انتخاب اعضا به‌عنوان پدر یا مادر، والدین جمعیت جدید که در نتیجه عملگر آمیزش انتخاب می‌شوند را مشخص می‌کند.

عملگرهای تقاطع و جهش: فرض کنید که در نتیجه عملگر انتخاب، دو والد P_1 و P_2 برای تولید نسل انتخاب شوند. در نتیجه به‌کارگیری عملگر آمیزش، دو فرزند جدید از این دو والد تولید می‌شوند. از آنجا که نمایش جواب دو والد به‌صورت یک رشته باینری است، از عملگر تقاطع تک‌نقطه‌ای روی رشته باینری برای ایجاد فرزندان جدید در این پژوهش استفاده می‌شود. آنچه در ادامه آمده است، نمونه‌ای از عملیات تقاطع تک‌نقطه‌ای است که روی دو کروموزوم به طول w اجرا شده است:



شکل ۳. نمونه‌ای از عملیات تقاطع تک‌نقطعی روی دو کروموزوم

در نهایت عملگر جهش یکنواخت روی فرزندان تولیدشده به‌کار گرفته می‌شود. بدین منظور، تعداد کل بیت‌های فرزندان تولیدشده در احتمال جهش ضرب شده و این حاصل‌ضرب، تعداد بیت‌هایی که روی آنها عملیات جهش انجام می‌شود را مشخص خواهد کرد. سپس به تعداد بیت‌های مشخص‌شده، بیت‌هایی که جهش روی آنها انجام می‌شود به‌صورت تصادفی انتخاب و با احتمال برابر انتخاب می‌شوند. در نتیجه عملگر جهش روی آنها در صورتی که بیت انتخاب‌شده دارای مقدار ۱ باشد مقدار آن به صفر تغییر می‌یابد و در صورتی که مقدار آن صفر باشد، مقدار آن به ۱ تبدیل خواهد شد. در حقیقت با این روش مقدار بیت‌ها بین حد پایین (۰) و حد بالای (۱) آنها جابه‌جا می‌شوند.

الگوریتم روش پیشنهادی

الگوریتم یادگیری مبتنی بر GA، که مشخصه‌های پارامتری مناسب برای طراحی طبقه‌بند پیشنهادی را تعیین می‌کند به شرح ذیل نوشته می‌شود:

مرحله ۱. نرمال‌سازی معیارها

هر آلترناتیو، نرمال‌سازی را مطابق رابطه ۸ برای تمامی معیارها انجام دهد تا معیارها بی‌مقیاس شوند.

مرحله ۲. تولید جمعیت اولیه

تعداد N_{pop} کروموزوم باینری به‌صورت $b_1b_2 \dots b_{n^2}b_{n^2+1}b_{n^2+2} \dots b_{n^2+c-1}$ برای جمعیت اولیه به‌صورت تصادفی تولید کن ($b_1b_2 \dots b_{n^2}$ متناظر با درایه‌های سوپر ماتریس و $(0 \le j \le C-1)$ b_{n^2+j} ژن‌های متناظر با نقاط برش t_j است).

مرحله ۳. محاسبه مقادیر برازش

برای هر کروموزوم موجود در جمعیت فعلی:

۱. بردارهای اولویت محلی[1] را به‌منظور ایجاد سوپرمارتریس تولید کن.

۲. سوپرماتریس محدودکننده را تولید کن: اگر سوپرماتریس ماتریس مربعی غیرمنفی[2] بود، در آن صورت سوپرماتریس به توان ۲K+۱ برسد (K= ۵۰) تا سوپرماتریس محدود به دست آید.

۳. وزن‌های نهایی را به‌طور مستقیم از سوپرماتریس محدود بگیر.

۴. ارزش هر آلترناتیو را با استفاده از رابطه ۷ محاسبه کن.

۵. نقاط برش[3] را از داخل هر کروموزوم رمزگشایی کن.

۶. ارزش هر آلترناتیو را با نقاط برش مقایسه کن، سپس آلترناتیو را مطابق رابطه ۹ به کلاس مناسب تخصیص بده.

۷. تابع برازش را به‌منظور محاسبه دقت کلاسه‌بندی انجام‌شده با استفاده از رابطه ۱۰ محاسبه کن.

مرحله ٤. تولید کروموزوم‌های جدید

تعداد N_{pop} کروموزوم جدید با استفاده از تقاطع تک‌نقطه‌ای[4] و جهش یکنواخت[5] با توجه به نرخ‌های ذکرشده در جدول ۲ از جمعیت فعلی تولید کن.

مرحله ٥. اجرای استراتژی نخبه‌گرایی

به‌صورت تصادفی تعداد N_{del} کروموزوم را از جمعیت فعلی حذف کن، سپس به تعداد N_{del} تا از بهترین کروموزم‌ها را به جمعیت فعلی اضافه کن تا جمعیت نسل بعدی تشکیل شود.

مرحله ٦. آزمون خاتمه

اگر شرط از پیش تعیین‌شده برای توقف برآورده نشود به مرحله ۳ برگرد. بدین معنا که اگر شرط از پیش تعیین‌شده برای توقف برآورده نشد، عملیات ژنتیک را تکرار کن. در غیر این صورت، به‌روزرسانی نسل خاتمه یابد. در این مقاله الگوریتم زمانی خاتمه می‌یابد که تعداد تکرارها از حد معینی ($Max_{Iteration}$) تجاوز نکند.

1. Priority vectors
2. Primitive
3. Cutoff points
4. Single point crossover
5. Uniform mutation

یافته‌های پژوهش

پارامترهای الگوریتم ژنتیک

در این بخش، برای آموزش و ارزیابی الگوریتم، داده‌ها به دو دسته آموزش و آزمایش تقسیم می‌شوند. الگوریتم ژنتیک با استفاده از داده‌های آموزش مقادیر پارامترها را تخمین زده و پارامترهای بهینه را تعیین می‌کند، سپس داده‌های آزمایش به طبقه‌بند اعمال می‌شود تا کلاس آن را تشخیص دهد. سایر پارامترهای در نظر گرفته‌شده در الگوریتم ژنتیک برای پیاده‌سازی در نرم‌افزار متلب به شرح جدول ۲ هستند.

جدول ۲. پارامترهای مورد استفاده در الگوریتم ژنتیک

پارامتر	مقدار
حداکثر تعداد تکرار الگوریتم (Max_{It})	۳۰۰
تعداد اعضای جمعیت در هر تکرار (N_{pop})	۳۰
تعداد اعضای حذف‌شده جمعیت (N_{Del})	۸
احتمال تقاطع (pcrossover)	۰/۴
احتمال جهش (pmutation)	۰/۱۵

دیتاست‌های مورد استفاده

روش پیشنهادی روی پنج دیتاست پیاده‌سازی می‌شود، سه دیتاست مربوط به داده‌های بانکی استرالیا، آلمان، ژاپن و برگرفته از انباره داده‌های یادگیری ماشینی دانشگاه کالیفرنیا اروایین[1] هستند. دیتاست ایران و حاوی اطلاعات مربوط به طبقه‌بندی مشتریان بدحساب و خوش‌حساب از یک بانک خصوصی و دیتاست لهستان مربوط به داده‌های ۱۲۰ شرکت طی یک دوره زمانی دوساله است (مارکز، گارسیا و سانچز[2]، ۲۰۱۲). در جدول ۳ خلاصه مشخصه‌های دیتاست‌های استفاده‌شده نشان داده شده است.

جدول ۳. مشخصات دیتاست‌ها مورد استفاده

دیتاست	تعداد داده‌ها	تعداد معیارها	داده‌های کلاس ۱	داده‌های کلاس ۲
استرالیا	۶۹۰	۱۴	۳۰۷	۳۸۳
آلمان	۱۰۰۰	۲۴	۷۰۰	۳۰۰
ژاپن	۶۵۳	۱۵	۲۹۶	۳۵۷
لهستان	۲۴۰	۳۰	۱۱۲	۱۲۸
ایران	۱۰۰۰	۲۷	۹۵۰	۵۰

ارزیابی عملکرد روش پیشنهادی

دیتاست ایران از ۱۰۰۰ آلترناتیو و ۲۷ معیار تشکیل شده است. ما این دیتاست را به‌عنوان یک مثال در نظر می‌گیریم تا به‌طور خلاصه نحوه تعیین بردارهای اولویت و تشکیل سوپرماتریس را نشان دهیم. نمایش شبکه‌ای مشابه با شکل ۱

1. http://archive.ics.uci.edu/ml
2. Marqués, Garcia & Sánchez

است با فقط یک جزء[1] به نام معیارها که از ۲۷ معیار مختلف، $C_1, C_2, C_3 \dots, C_{27}$ تشکیل شده است. سپس ۲۷ بردار اولویت محلی شامل $(C_{11}, C_{21}, \dots, C_{27,1})^T$، $(C_{12}, C_{22}, \dots, C_{27,2})^T$، ...، $(C_{1,27}, C_{2,27}, \dots, C_{27,27})^T$ ، به‌صورت اتوماتیک با استفاده از الگوریتم ژنتیک تولید می‌شوند. در نتیجه سوپرماتریس متناظر زیر، ساخته می‌شود:

$$W = \begin{pmatrix} C_{11} & \cdots & C_{1,27} \\ \vdots & \ddots & \vdots \\ C_{27,1} & \cdots & C_{27,27} \end{pmatrix}$$

$$j = 1, 2, \dots, 27$$

$$C_{1j} + C_{21j} + \cdots + C_{27j} = 1$$

اگر سوپرماتریس ماتریس مربعی غیرمنفی باشد، آنگاه سوپرماتریس محدود متناظر آن می‌تواند در ادامه با محاسبه W^{101} تولید شود. همان‌طور که پیش‌تر ذکر شد، سوپرماتریس محدود (یعنی W^{101}) اهمیت نسبی یا وزن هر معیار را به دست می‌دهد.

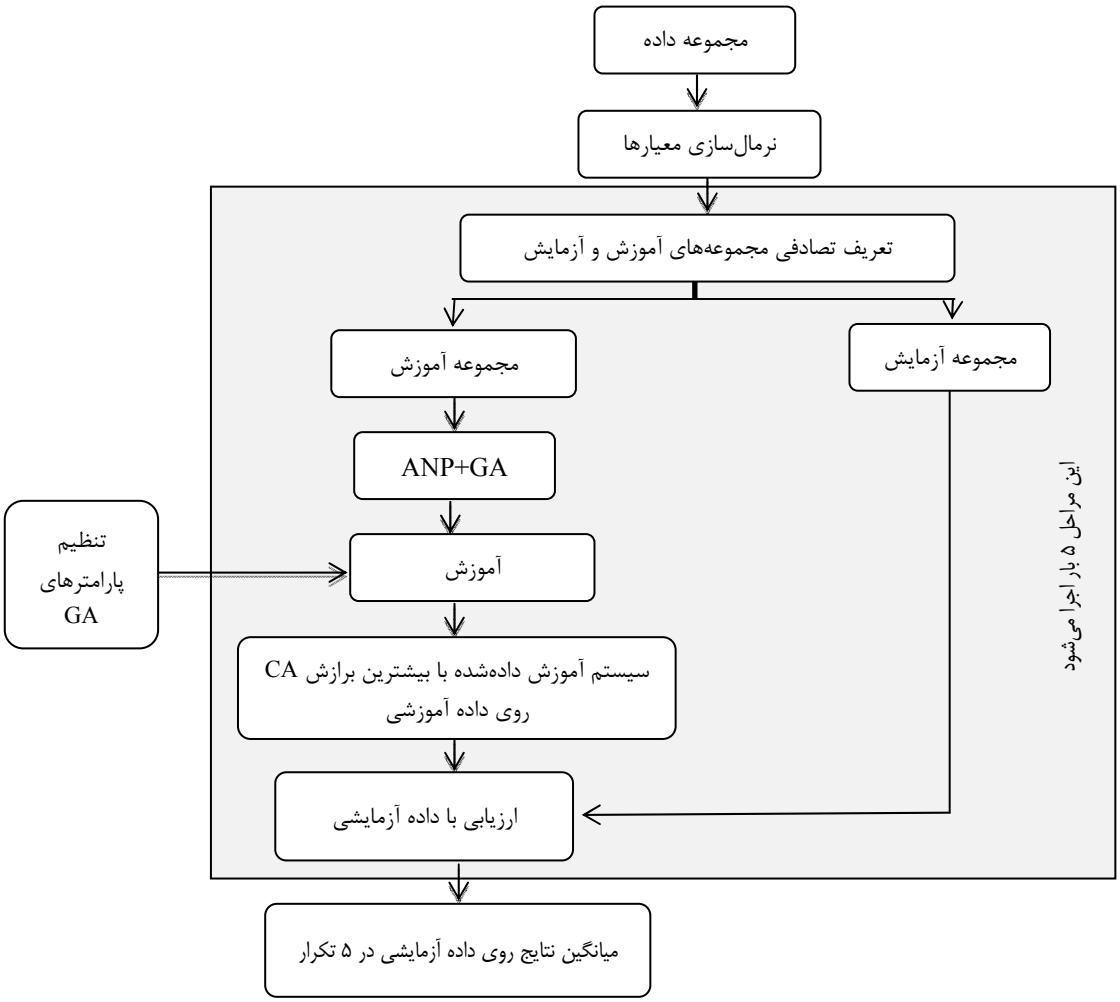
عملکرد روش پیشنهادی در دو مرحله ارزیابی می‌شود. در مرحله نخست روش پیشنهادی روی هر یک از دیتاست‌ها آزمایش می‌شود. در این پژوهش از روش اعتبارسنجی تقاطعی k بخشی[2] (K=۵) طبقه‌بندی‌شده استفاده می‌شود. رویه این کار به این صورت است که دیتاست مورد بررسی به پنج فولد مجزا تقسیم‌بندی شده و هر بار یک فولد برای آزمایش و فولدهای باقی‌مانده برای آموزش استفاده شدند. از این رو همه فولدها حداقل یک بار به‌عنوان داده آزمایشی استفاده می‌شوند. این رویه همچنین باعث می‌شود که عملکرد الگوریتم‌های مقایسه‌شده روی همه داده‌ها به‌طور یکسان ارزیابی شوند.

ما از نمونه‌های آموزشی برای بهینه‌سازی پارامترهای الگوریتم ژنتیک استفاده می‌کنیم. در شکل ۴، فرایند بهینه‌سازی پارامترهای الگوریتم ژنتیک و محاسبه دقت طبقه‌بندی (CA) روی داده‌های آزمایش نشان داده می‌شود. این رویه برای دیتاست ایران انجام شده و در جدول ۴ آمده است.

جدول ٤. اجرای الگوریتم روی دیتاست ایران با روش پنج فولد

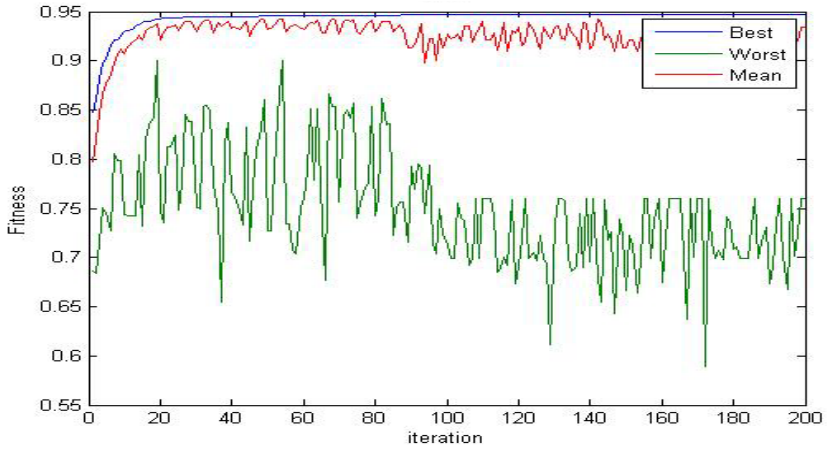
فولد	آموزش	آزمایش
فولد ۱	۹۶/۰۴	۹۵/۸۴
فولد ۲	۹۶/۰۴	۹۴/۸۲
فولد ۳	۹۶/۱۶	۹۵/۵۹
فولد ۴	۹۶/۰۳	۹۴/۶۹
فولد ۵	۹۶/۲۹	۹۵/۸۸
میانگین	۹۶/۱۱	۹۵/۳۷

1. Component
2. K-fold cross validation



شکل ٤. فرایند بهینه‌سازی پارامترهای الگوریتم ژنتیک

نمودار هم‌گرایی بهترین، بدترین و یکی از اجراهای مستقل الگوریتم پیشنهادی روی دیتاست ایران در شکل ۵ مشاهده می‌شود.



شکل ۵. نمودار هم‌گرایی اجرای الگوریتم روی دیتاست ایران

در مرحله دوم عملکرد روش پیشنهادی بر اساس صحت طبقه‌بندی یعنی درصد آلترناتیوهای درست طبقه‌بندی شده به کل آلترناتیوهای طبقه‌بندی، با پنج طبقه‌بند یادگیری ماشینی شامل طبقه‌بندی‌کننده بیز ساده (NBC)، پرسپترون چندلایه‌ای (MLP)، رگرسیون لجستیک (LR)، ماشین بردار پشتیبان (SVM) و درخت تصمیم (C4.5) مقایسه می‌شود.

جدول ۵ میانگین صحت طبقه‌بندی هر یک از روش‌ها با اعمال روی دیتاست‌های مختلف را نشان می‌دهد. نتایج عملکردی پنج طبقه‌بند اول روی دیتاست‌ها از مقاله (مارکز و همکاران، ۲۰۱۲) استخراج شده است. بر این مبنا، روش پیشنهادی به‌طور میانگین از عملکرد خوبی برخوردار است.

جدول ۵. مقادیر صحت طبقه‌بندی روی دیتاست‌ها

طبقه‌بند	استرالیا	آلمان	ژاپن	ایران	لهستان
NBC	۸۰/۷۲	۷۴/۹۰	۸۲/۰۸	۹۳/۲۰	۶۸/۳۳
Log R	۸۴/۹۳	۷۵/۷۰	۸۷/۲۹	۹۴/۲۰	۷۲/۷۰
MLP	۸۳/۰۴	۷۲/۴۰	۸۳/۳۰	۹۴/۸۰	۷۲/۹۲
SVM	۸۵/۰۷	۷۶/۰۰	۸۶/۳۷	۹۵/۰۰	۷۱/۲۵
C4.5	۸۵/۵۱	۷۳/۳۰	۸۴/۲۲	۹۴/۵۰	۶۸/۷۵
GA-ANP	۶۰/۱۴	۷۵/۱۰	۶۴/۶۲	۹۵/۳۷	۷۲/۹۶

بر مبنای پیشنهاد دمسار[1] (۲۰۰۶)، آزمون فریدمن و سایر آزمون‌های پیشرفته همچون آزمون نمنی برای تعیین عملکرد کلی هر طبقه‌بند روی کلیه دیتاست‌ها استفاده می‌شوند.

در این بخش با استفاده از آزمون فریدمن که به‌منظور مقایسه داده‌های رتبه‌ای استفاده می‌شود، به مقایسه آماری طبقه‌بندهای به‌کارگرفته‌شده پرداخته شده است. به‌منظور اجرای این آزمون، نخست رتبه هر یک از طبقه‌بندها بر اساس مقدار صحت آنها در هر یک از دیتاست‌ها مشخص می‌شود و با توجه به اینکه شش روش بررسی شده‌اند، به بهترین آنها در هر دیتاست رتبه ۱ و به بدترین آنها رتبه ۶ اختصاص خواهد یافت. رتبه هر یک از طبقه‌بندهای مد نظر در جدول ۶ آورده شده است.

جدول ٦. رتبه طبقه‌بندهای مختلف ٦ روی دیتاست‌های مختلف

طبقه‌بند	استرالیا	آلمان	ژاپن	ایران	لهستان	میانگین رتبه
NBC	۵	۴	۵	۶	۶	۵/۲
Log R	۳	۲	۱	۵	۳	۲/۸
MLP	۴	۶	۴	۳	۲	۳/۸
SVM	۲	۱	۲	۲	۴	۲/۲
C4.5	۱	۵	۳	۴	۵	۳/۶
GA-ANP	۶	۳	۶	۱	۱	۳/۴

1. Demsar

نتایج حاصل از پیاده‌سازی آزمون فریدمن نشان می‌دهد که الگوریتم‌های GA-ANP و NBC به‌طور میانگین در مقایسه با سایر روش‌ها به‌غیر از ماشین بردار پشتیبان عملکرد بهتری دارند. مقدار P-value آزمون فریدمن برابر ۰/۰۲۵ به دست آمد که نشان‌دهنده رد فرض برابری عملکرد طبقه‌بندها در سطح معناداری ۰/۰۵ است.

آزمون نمنی به‌منظور مقایسه زوجی روش‌های مختلف روی دیتاست‌های مختلف استفاده شده است. در این آزمون اگر میانگین رتبه دو روش روی دیتاست‌های مختلف، اختلافی بیشتر از یک مقدار بحرانی (CD)[1] داشته باشد، آنگاه عملکرد دو روش تفاوت معناداری خواهد داشت. مقدار بحرانی در این روش با توجه به رابطه $CD = \sqrt{K(K+1)/6D}$ محاسبه می‌شود که در آن K و D به‌ترتیب تعداد روش‌ها و تعداد دیتاست‌ها هستند. مقایسه میانگین رتبه GA-ANP با سایر روش‌ها در جدول ۷ نشان داده شده است.

جدول ۷. تفاوت میانگین رتبه روش GA-ANP با سایر طبقه‌بندها

روش	طبقه‌بندی‌کننده ساده بیز	پرسپترون چندلایه‌ای	رگرسیون لجستیک	ماشین بردار پشتیبان	درخت تصمیم
اختلاف میانگین رتبه‌ها	۳/۲	۱/۴	۰/۸	۰	۱/۲

با توجه به اینکه در آزمون مد نظر مقدار CD برابر ۱/۱۸ است، روش GA-ANP در مقایسه با روش‌های طبقه‌بندی‌کننده ساده بیز، پرسپترون چندلایه‌ای، درخت تصمیم و رگرسیون لجستیک اختلاف معنادار و عملکرد به‌مراتب بهتری دارد.

نتیجه‌گیری و پیشنهادها

در پژوهش حاضر به‌منظور مدل‌سازی مسئله طبقه‌بندی از ANP استفاده شده است. مراحل این روش شامل تشکیل سوپرماتریس اولیه، تشکیل سوپرماتریس نهایی، نرمال‌سازی امتیاز نمونه‌ها در شاخص‌ها، تعیین شاخص طبقه‌بندی نمونه‌ها و در نهایت مقایسه با آستانه‌های تعریف‌شده کلاس‌ها معرفی شدند. همچنین پارامترهای مورد استفاده در روش فرایند تحلیل شبکه‌ای شامل سوپرماتریس اولیه و سطوح آستانه، توسط الگوریتم ژنتیک از روی داده‌های نمونه به‌عنوان ورودی‌های الگوریتم ژنتیک تخمین زده شد. پیاده‌سازی روش پیشنهادی روی دیتاست‌های اعتباری نشان داد که الگوریتم ژنتیک در تکرارهای مختلف به بهبود شاخص صحت طبقه‌بندی آلترناتیوها منجر می‌شود. به‌علاوه، این الگوریتم از قابلیت هم‌گرایی و مقدار تابع برازش خوبی برای مسئله طبقه‌بندی برخوردار است.

برای ارزیابی روش پیشنهادی، عملکرد آن با تعدادی از الگوریتم‌های یادگیری ماشین شامل طبقه‌بندی ساده بیز، پرسپترون چندلایه‌ای، رگرسیون لجستیک، ماشین بردار پشتیبان و درخت تصمیم مقایسه شد. نتایج مقایسه با استفاده از آزمون فریدمن نشان داد که روش‌های نام‌برده عملکرد یکسانی ندارند. با توجه به رد فرضیه یکسانی عملکرد این روش‌ها، برای مقایسه زوجی عملکرد طبقه‌بندها از آزمون نمنی استفاده شد. بر اساس نتایج این آزمون، روش پیشنهادی در مقایسه با روش‌های طبقه‌بندی‌کننده ساده بیز، پرسپترون چندلایه‌ای و درخت رگرسیون عملکرد بهتری دارد.

1. Critical Difference

در خصوص مرحله ۱ (یعنی نرمال‌سازی) در الگوریتم یادگیری مبتنی بر GA پیشنهادی، حداکثر و حداقل مقادیر نرمال‌سازی برای هر معیار به ترتیب برابر با ۱ و ۰ است. افزون بر روش نرمال‌سازی ذکرشده در بخش قبلی، می‌توان روش‌های دیگری را در نظر گرفت. روش‌های مختلف نرمال‌سازی ممکن است بر عملکرد طبقه‌بندی تأثیر بگذارند. مقاله حاضر بر موضوع یادشده در رابطه با نرمال‌سازی تمرکز ندارد و این موضوع را می‌توان در پژوهش‌های آتی بررسی کرد.

به‌طور کلی برای حل هر مسئله بهینه‌سازی در ابتدا با مجموعه‌ای از جواب‌های تصادفی سروکار داریم. بنابراین ترجیح بر این است که در تکرارهای اولیه الگوریتم، جست‌وجوی سراسری بیشتری در کل فضای جست‌وجو انجام شود. هر چه الگوریتم به پیش می‌رود، کیفیت جواب‌ها بهتر شده و الگوریتم با راه‌حل‌های نزدیک به بهینه[1] سروکار دارد. در چنین حالتی به‌منظور افزایش سرعت و دقت، بهتر است جست‌وجوی محلی پیرامون جواب (های) نزدیک به بهینه انجام شود. پیشنهاد می‌شود به‌منظور دستیابی به یک تبادل[2] بهتر دقت- سرعت، ابتدا جست‌وجوی سراسری با استفاده از الگوریتم مبتنی بر جمعیت ژنتیک انجام شود. بعد از دستیابی به جواب نزدیک به بهینه، به‌منظور صرفه‌جویی در زمان، افزایش سرعت الگوریتم و دستیابی به دقت بالاتر، الگوریتم زمان بر مبتنی بر جمعیت، متوقف و ادامه فرایند بهینه‌سازی بر عهده الگوریتم تک‌جمعیتی همچون تبرید شبیه‌سازی‌شده که قابلیت جست‌وجوی محلی بالایی دارد، قرار گیرد.

منابع

دانشور، امیر؛ زندیه، مصطفی؛ ناظمی، جمشید (۱۳۹۴). یک روش تکاملی برای طبقه‌بندی اعتباری مبتنی بر رویکرد تجمیع‌زدایی ترجیحات. *مطالعات مدیریت صنعتی*، ۱۳ (۴)، ۱-۳۴.

دانشور، امیر؛ همایون‌فر، مهدی؛ اخوان، الهام (۱۳۹۸). توسعه روش طبقه‌بندی دیتاست‌های نامتوازن با استفاده از الگوریتم‌های تکاملی چندهدفه. *مطالعات مدیریت صنعتی*، ۱۷ (۴)، ۱۶۱-۱۸۳.

دانشور، امیر؛ همایون‌فر، مهدی؛ فرهمندنژاد، آنیا (۱۳۹۸). توسعه یک روش هوشمند خوشه‌بندی چندمعیاره مبتنی بر پرامیتی. *چشم‌انداز مدیریت صنعتی*، ۹ (۴)، ۴۱-۶۱.

زرین‌صدف، مسعود؛ دانشور، امیر (۱۳۹۵). روش کارای یادگیری ترجیحات مبتنی بر مدل ELECTRE TRI به‌منظور طبقه‌بندی چندمعیار موجودی. *مجله مدیریت صنعتی*، ۸ (۲)، ۱۹۱-۲۱۶.

References

Aragonés-Beltrán, P., Aznar, J., Ferrís-Oñate, J., & García-Melón, M. (2008). Valuation of urban industrial land: an analytic network process approach. *European Journal of Operational Research*, 185 (1), 322–339.

Baccour, L. (2018). Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets. *Expert Systems with Applications*, 99, 115-125.

Chung, S.H., Lee, H.I., & Pearn, W.L. (2005). Analytic network process (ANP) approach for product mix planning in semiconductor fabricator. *International Journal of Production Economics*, 96, 15–36.

1. Near Optimal
2. Trade off

Daneshvar, A., Homayounfar, M., & Akhavan, E. (2020). Developing a classification Method for Imbalanced Dataset Using Multi-Objective Evolutionary Algorithms. *Industrial Management Studies*, 17 (4), 161-183. (*in Persian*)

Daneshvar, A., Homayounfar, M., & Farahmandnejad, A. (2020). Developing an Intelligent Multi Criteria Clustering Method Based on PROMETHEE. *Journal of Industrial Management Perspective*, 9 (4), 41-61. (*in Persian*)

Daneshvar, A., Zandieh, M., & Nazemi, J. (2015). An evolutionary method for credit scoring; Preference Disaggregation approach. *Industrial Management Studies*, 13 (4), 1-34. (*in Persian*)

Demsar, J. (2006), Statistical comparisons of classifiers over multiple datasets. *Journal of Machine Learning Research,* 7, 1–30.

Doumpos, M., Zopounidis, C. (2002). *Multicriteria Decision Aid Classification Methods*. Kluwer, Dordrecht.

Kartal, H., Oztekin, A., Gunasekaran, A., & Cebi, F. (2016). An integrated decision analytic framework of machine learning with multi-criteria decision making for multi-attribute inventory classification. *Computers & Industrial Engineering*, 101, 599-613.

Kou, G., Peng, Y., & Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences*, 275, 1-12.

Lee, J.W., Kim, S.H. (2000). Using analytic network process and goal programming for interdependent information system project selection, *Computers and Operations Research*, 27, 367–382.

Marqués, A.I., Garcia, V., & Sánchez, J.S. (2012). Exploring the behavior of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, 39 (11), 10244-10250.

Massam, B.H. (1988). *Multicriteria Decision Making Techniques in Planning*, Pergamon, NY.

Meade, L.M., Presley, A. (2002). R&D project selection using the analytic network process, *IEEE Transactions on Engineering Management*, 49 (1), 59–66.

Meade, L.M., Sarkis, J. (1999). Analyzing organizational project alternatives for agile manufacturing processes: an analytical network approach. *International Journal of Production Research*, 37(2), 241–261.

Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50 (3), 559-569.

Niemira, M.P., Saaty, T.L. (2004). An analytic network process model for financial-crisis forecasting. *International Journal of Forecasting*, 20 (4), 573-587.

Nikam, S.S. (2015). A Comparative Study of Classification Techniques in Data Mining Algorithms. *Computer Science and Technology*, 8 (1), 13-19.

Ravi, V., Shankar, R., Tiwari, M.K. (2005). Analyzing alternatives in reverse logistics for end-of-life computers: ANP and balanced scorecard approach. *Computers & industrial engineering,* 48 (2), 327-356.

Saaty, T.L. (1996). *The Analytic Network Process,* RWS Publications, Pittsburgh.

Saaty, T.L. (2001). Analytic network process. *Encyclopedia of Operations Research and Management Science. Springer*, 28-35.

Tuzkaya, U.R., Önüt, S. (2008). A fuzzy analytic network process based approach to transportation-mode selection between Turkey and Germany: a case study, *Information Sciences*, 178 (15), 3133–3146.

Wang, W., Wang, Z., Klir, G.J. (1998). Genetic algorithms for determining fuzzy measures from data, *Journal of Intelligent and Fuzzy Systems*, 6, 171–183.

Wolfslehner, B., Vacik, H., & Lexer, M.J. (2005). Application of the analytic network process in multicriteria analysis of sustainable forest management, *Forest Ecology and Management*, 207, 157–170.

Yüksel, I., Dagdeviren, M. (2007). Using the analytic network process (ANP) in a SWOT analysis – a case study for a textile firm, *Information Sciences,* 177 (16), 3364–3382.

Zarrin Sadaf, M., Daneshvar, A. (2016). An efficient preference learning method based on ELECTRE TRI model for multi-criteria inventory classification. *Industrial Management Journal*, 8 (2), 191-216. (*in Persian*)