University of Qom

Sciences and Techniques of Information Management

Volume 9, Issue 3, 2023

A Feature Selection Method Based on Information Theory and Genetic Algorithm

Mehdi Jabbari
Master, Department of Computer Engineering, Qom University of Technology, Qom, Iran.
jabbari@qut.ac.ir

Jalal Rezaeenour
Professor, Department of Industrial Engineering, University of Qom, Qom, Iran **(Corresponding author).** j.rezaee@qom.ac.ir

Amir Hossein Akbari
P.hD., Student, Faculty of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran. akbari_amir@ind.iust.ac.ir

Abstract

Purpose: When dealing with high-dimensional datasets, dimensionality reduction is a crucial preprocessing step to achieve high accuracy, efficiency, and scalability in classification problems. This research aims to introduce a feature selection method for high-dimensional datasets by employing dimensionality reduction and genetic algorithms.

Method: In this study, an innovative algorithm has been developed to determine the mutual information between features and the target class using a new criterion. In this method, new characteristics are generated through the combination or transformation of the original characteristics. In this manner, the multi-dimensional space is transformed into a new space with fewer dimensions. In addition to considering the new criterion of mutual information, a genetic algorithm has been employed to enhance the speed of the proposed method.

Findings: The performance of this method has been evaluated on datasets of varying dimensions, with the number of features ranging from 13 to 60. The proposed method has been evaluated in comparison to similar methods, focusing on classification accuracy. The results have been promising.

Conclusion: The proposed method has been applied using MRMR, DISR, JMI, and NJMIM methods on various datasets. The average accuracies obtained from the proposed method are 65.32%, 74.51%, 70.88%, and 58.2%, indicating the efficiency of the proposed method. According to the results obtained, the proposed method outperformed DISR, JMI, NJMIM, and MRMR on average, except for the sonar data set, where the sonar data set yielded better results than the proposed method.

Keywords: Feature Selection, Data Pre-Processing, Information Theory, Genetic Algorithm, Classification.

Cite this article: Jabbari, M., Rezaeenour, J. & Akbari, A.H. (2023). A Feature Selection Method Based on Information Theory and Genetic Algorithm. *Sciences and Techniques of Information Management*, 9(3): 7-32. https://doi.org/10.22091/STIM.2023.8708.1877

Received: 2023-06-26 ; **Revised:** 2023-07-15 ; **Accepted:** 2023-07-25 ; **Published online:** 2023-07-28

Article type: Research **Published by:** University of Qom.

علوم و فنون مدیریت اطلاعات
دوره ۹، شماره ۳، ۱۴۰۲

توسعه یک روش انتخاب مشخصه مبتنی بر نظریه اطلاعات و الگوریتم ژنتیک

مهدی جباری
کارشناسی ارشد، گروه مهندسی کامپیوتر، دانشگاه صنعتی قم، قم، ایران. jabbari@qut.ac.ir
جلال رضائی نور
استاد، گروه مهندسی صنایع، دانشگاه قم، قم، ایران (**نویسنده مسئول**). j.rezaee@qom.ac.ir
امیرحسین اکبری
دانشجوی دکتری، دانشکده مهندسی صنایع، دانشگاه علم و صنعت، تهران، ایران. akbari_amir@ind.iust.ac.ir

چکیده

هدف: در مواجهه با مجموعه داده‌های با ابعاد بالا، کاهش بُعد یک گام پیش‌پردازشی مهم برای حصول دقت بالا، کارایی و مقیاس‌پذیری در مسائل کلاسبندی است. هدف تحقیق حاضر ارائه یک روش انتخاب مشخصه در مواجهه با مجموعه داده‌های با ابعاد بالا، با استفاده از کاهش بُعد و الگوریتم ژنتیک است.

روش: در این تحقیق یک الگوریتم ابتکاری توسعه یافته است که با استفاده از یک معیار جدید، اطلاعات متقابل بین ویژگی‌ها و کلاس هدف را مشخص می‌کند. در این روش مشخصه‌های جدید براساس ترکیب یا تبدیل مشخصه‌های اصلی تولید می‌شود و به این ترتیب فضای چند بُعدی، به فضایی جدید با ابعاد کمتر نگاشت پیدا می‌کند. همچنین علاوه‌بر در نظر گرفتن معیار جدید اطلاعات متقابل، از الگوریتم ژنتیک به منظور بهبود سرعت روش پیشنهادی استفاده شده است.

یافته‌ها: عملکرد این روش بر روی مجموعه داده‌هایی با ابعاد مختلف، که تعداد مشخصه‌ها در آن‌ها از ۱۳ تا ۶۰ متفاوت بوده، ارزیابی شده است. ارزیابی روش پیشنهادی در مقایسه با روش‌های مشابه، از لحاظ دقت کلاسبند بررسی شده و نتایج نویدبخشی بدست آمد.

نتیجه‌گیری: روش پیشنهادی با روش‌های MRMR, DISR, JMI, NJMIM در مجموعه داده‌های متفاوت اعمال شده است. متوسط دقت‌های به دست آمده از روش پیشنهادی ۶۵,۳۲، ۷۴,۵۱، ۷۰,۸۸ و ۵۸,۲ درصد می‌باشد، که حاکی از کارآمدی روش پیشنهادی است. طبق نتایج بدست آمده، به جز در مورد مجموعه داده sonar که نتیجه‌ای بهتر از روش پیشنهادی داشته است، متوسط عملکرد روش پیشنهادی بهتر از DISR, JMI, NJMIM و MRMR بوده است.

کلیدواژه‌ها: انتخاب مشخصه، پیش پردازش داده، تئوری اطلاعات، الگوریتم ژنتیک، کلاسبند.

استناد به این مقاله: جباری، م.، رضائی نور، ج.، اکبری، ا.ح. (۱۴۰۲). توسعه یک روش انتخاب مشخصه مبتنی بر نظریه اطلاعات و الگوریتم ژنتیک. *علوم و فنون مدیریت اطلاعات*، ۹(۳): ۷-۳۲. https://doi.org/10.22091/STIM.2023.8708.1877

تاریخ دریافت: ۱۴۰۲/۰۴/۰۵ ؛ **تاریخ اصلاح:** ۱۴۰۲/۰۴/۲۴ ؛ **تاریخ پذیرش:** ۱۴۰۲/۰۵/۰۳ ؛ **تاریخ انتشار آنلاین:** ۱۴۰۲/۰۵/۰۶

نوع مقاله: پژوهشی
ناشر: دانشگاه قم

۱. مقدمه

با گسترش رو به رشد ابزارهای مدرن در ثبت و ذخیره داده‌ها، انبوهی از دادگان در حوزه‌های مختلف دانش مانند تشخیص الگو، بیوانفورماتیک و تحلیل زبان، تولید شده است. این مجموعه داده‌ها علاوه‌بر داشتن مقادیر مفید، شامل داده‌های پرت، حشو[1]، مازاد و بی‌ربط نیز هستند. موضوع داده‌های با ابعاد بالا، در دنیای امروز یک چالش بزرگ است؛ چراکه حجم بالای مشخصه‌ها باعث افزایش پیچیدگی در تحلیل داده‌ها می‌شود. بنابراین، استخراج اطلاعات مفید از مجموعه داده ضروری است؛ چراکه انتخاب ویژگی، منجربه بهبود عملکرد مدل یادگیری ماشین و ارائه درك بهتری از فرآیند یادگیری خواهد شد. بسیاری از مطالعات نشان داده‌اند که حذف مشخصه‌های بی‌ربط و انتخاب متغیرهای حاوی اطلاعات مفید، بهبود چشم‌گیری به عملکرد یادگیری می‌بخشد (Vergara & et al. 2014). از این‌رو برای پردازش در زمان مناسب و با دقت مطلوب، کشف دانش نهفته در دادگان، منوط به ابعاد منطقی مسئله است. کاهش بُعد، یك گام پیش‌پردازشی مهم برای رسیدن به دقت بالا، کارایی و مقیاس‌پذیری در مسائل کلاس‌بندی است. در برخی موارد به علت وجود همبستگی بین داده‌ها، کاهش ابعاد کار دشواری است. این امر تحلیل‌گران را به سمت استفاده از تکنیک‌هایی که دارای توانایی کار با داده‌های با ابعاد مختلف را دارند، سوق می‌دهد. تعداد زیاد مشخصه‌ها و الگوهای یادگیری، در نهایت منجربه پیش‌بینی نادرست در خروجی می‌شود. دو رویکرد کاهش بُعد، استخراج مشخصه[2] و انتخاب مشخصه[3] است. در استخراج مشخصه، مشخصه‌های جدیدی براساس ترکیب یا تبدیل مجموعه مشخصات اصلی تولید می‌شود و به این ترتیب فضای چند بُعدی، به فضایی جدید با ابعاد کمتر نگاشت پیدا می‌کند (Boukharouba & et al., 2018). در حالی‌که انتخاب مشخصه، زیرمجموعه‌ای از مشخصه‌های مرتبط از مجموعه اصلی را انتخاب می‌کند. هدف فرآیند کاهش بُعد، (۱) کمینه کردن مشخصه‌های اضافی[4] و (۲) بیشینه کردن ابعاد مرتبط[5] است (Kramer & et al., 2013).

انتخاب بهترین مشخصه‌ها براساس محاسبه معیارهای مختلف انجام می‌شود که این معیار میزان تأثیر مشخصه کاندید بر روی کلاس هدف را نشان می‌دهد. همچنین در برخی موارد این معیار برای

1. Redundancy
2. Feature extraction
3. Feature selection
4. Minimizing redundancy
5. Maximizing relevancy

تعیین شباهت مشخصه‌های کاندید متفاوت، مورد استفاده قرار می‌گیرد. در این حالت مشخصه‌هایی انتخاب می‌شوند که کمترین ضریب اثر را بر روی یکدیگر و بیشترین ضریب اثر را با کلاس هدف دارند؛ زیرا مشخصه‌هایی که شباهت بالا و ضریب اثر بالایی روی یکدیگر دارند را می‌توان به یکی از مهم‌ترین آن‌ها کاهش بُعد داد.

در این مطالعه یک روش انتخاب مشخصه مبتنی بر معیار جدید محاسبه اطلاعات متقابل، بین مشخصه‌های کاندید و مشخصه هدف، پیشنهاد شده است. روش پیشنهادی، محدودیت‌های روش‌های انتخاب مشخصه موجود را که باعث انتخاب مشخصه‌های نامرتبط و حشو می‌شود، کاهش داده و باعث افزایش دقت کلاسبند می‌شود. در کنار این معیار، به منظور افزایش سرعت روش پیشنهادی، یک الگوریتم فراابتکاری مبتنی بر الگوریتم ژنتیک مورد استفاده قرار گرفته است. جهت بررسی کارایی، مجموعه داده‌های مختلف پایگاه UCI ابتدا برروی روش پیشنهادی و سپس بر روی کلاسبند *KNN* اعمال شدند و نتایج مقایسه روش‌های مختلف از نقطه‌نظر دقت کلاسبند، گزارش شده است که در سطح قابل قبولی قرار دارند.

۲. ادبیات موضوع

ایده اصلی روش‌های کاهش بُعد، نگاشت داده از یک فضای چند بُعدی به فضایی با ابعاد کمتر است. به طور کلی انتخاب مشخصه با سه هدف اصلی صورت می‌گیرد: ۱.جلوگیری از بیش‌برازش[1]، ۲. ایجاد مدل‌های یادگیری سریع‌تر، ۳. بهبود استخراج دانش از داده‌ها و تفسیرپذیری مدل یادگیری. فرآیند انتخاب مشخصه شامل بخش‌های زیر است (Kwak & et al., 2022).

۱) روال تولیدکننده:[2] این تابع زیرمجموعه کاندید را پیدا می‌کند.

۲) معیار ارزیابی:[3] زیرمجموعه کاندید را براساس یک معیار و روش خاص، ارزیابی می‌کند و خروجی آن یک عدد است. روش‌های مختلف سعی در ساختن زیرمجموعه‌ای دارند که این مقدار را بهینه کند.

۳) شرط خاتمه

۴) اعتبارسنجی:[4] تعیین می‌کند که آیا زیرمجموعه کاندید معتبر است یا خیر.

1. Over fitting
2. Generation procedure
3. Evaluation Function
4. Validation Procedure

محققان روش‌های مختلف انتخاب مشخصه را براساس نوع جستجو و معیار ارزیابی طبقه‌بندی می‌کنند. در روش‌هایی که از جستجوی کامل[1] استفاده می‌کنند، معیار ارزیابی، تمام فضای جواب را برای یافتن پاسخ بهینه جستجو می‌کند. در روش‌های با جستجوی مکاشفه‌ای،[2] با هر بار اجرای الگوریتم، یک مشخصه به مجموعه مشخصه‌های انتخاب شده افزوده یا از آن حذف می‌شود. به دلیل پیچیدگی زمانی کمتر، اجرای این الگوریتم، سریع‌تر و پیاده‌سازی آن ساده‌تر است.

معیار ارزیابی عملکرد، زیرمجموعه کاندید را بررسی کرده و یک مقدار عددی به عنوان میزان بهینگی زیرمجموعه موردنظر در نظر می‌گیرد. این مقدار با مقدار زیرمجموعه قبلی مقایسه می‌شود، اگر زیرمجموعه جدید، بهتر از زیرمجموعه قبلی باشد، به عنوان زیرمجموعه بهینه جایگزین می‌شود. یافتن یک زیرمجموعه بهینه در میان مجموعه ویژگی‌ها، به طور مستقیم به انتخاب تابع ارزیاب بستگی دارد. مقادیری که معیارهای ارزیابی مختلف به یک زیرمجموعه می‌دهند، با هم فرق دارد. معیارهای ارزیابی به پنج دسته تقسیم می‌شوند (Dash & et al., 1997): معیارهای مبتنی بر فاصله،[3] معیارهای مبتنی بر اطلاعات،[4] معیارهای مبتنی بر وابستگی،[5] معیارهای مبتنی بر سازگاری[6]، معیارهای مبتنی بر خطای کلاسبندی.[7]

روش‌های انتخاب مشخصه براساس نوع کارکرد، به دو دسته فیلتر و پوششی تقسیم می‌شوند. روش‌های پوششی در طول الگوریتم یادگیری اعمال می‌شوند (Blum & et al., 1997) و دارای دو ضعف عمده هستند:

۱) پیچیدگی‌های محاسباتی زیادی (بویژه در مواجهه با داده‌های با ابعاد بالا) دارند؛ زیرا ارزیابی هر ترکیبی از مشخصه‌ها، مستلزم انجام کامل فاز آموزش الگوریتم یادگیری است.

۲) فقدان عمومیت؛ یعنی هر تغییر ساده در مدل یادگیری کارایی زیرمجموعه مشخصه را کاهش داده و منجربه بیش‌برازش می‌شود.

روش‌های فیلتر، مستقل از الگوریتم یادگیری عمل کرده و مرتبط‌ترین مشخصه‌ها را انتخاب

1. Complete search
2. Heuristic search
3. Distance measures
4. Information measures
5. Dependence measures
6. Consistency measures
7. Classification Error rate measures

نموده و بقیه را حذف می‌کنند (Guyon & et al., 2003). این روش در داده‌هایی با ابعاد بالا، کاربرد زیادی دارد. روش‌های فیلتر، سریع و ساده عمل کرده و می‌توانند برای هر مسئله رگرسیون، مورد استفاده قرار گیرند.

روش Relief از مشهورترین روش‌های فیلتر است که براساس معیار ارزیابی مبتنی بر فاصله و تابع تولیدکننده مکاشفه‌ای می‌باشد (Kira & et al., 1992). در این روش ابتدا از میان مجموعه آموزشی، یک زیرمجموعه انتخاب می‌شود؛ کاربر بایستی تعداد نمونه‌ها در زیرمجموعه را تعیین کند. الگوریتم به صورت تصادفی یک نمونه از این زیرمجموعه را انتخاب کرده، سپس برای هر یک از مشخصه‌های این نمونه، نزدیک‌ترین برخورد و نزدیک‌ترین شکست را براساس معیار اقلیدسی محاسبه می‌کند. این الگوریتم برای مشخصه‌های دارای خطا یا دارای همبستگی خوب کار می‌کند و پیچیدگی زمانی آن خطی و تابعی از تعداد مشخصه‌ها و تعداد نمونه‌ها است.

الگوریتم مبتنی بر نسبت بهره در این الگوریتم، تابع ارزیابی مبتنی بر اندازه نسبت بهره است. مقدار نسبت بهره، بین یک مشخصه و مشخصه هدف، محاسبه شده و این مقدار برای تمام ویژگی‌ها و دسته‌ها، محاسبه می‌شود. تمام ویژگی‌ها براساس این مقدار رتبه‌بندی می‌شوند (Hall & et al., 2009).

الگوریتم مبتنی بر بهره اطلاعاتی در این الگوریتم، تابع ارزیابی مبتنی بر میزان بهره اطلاعاتی است. مقادیر H(class) و H(class | attribute) با استفاده از تابع بهره اطلاعاتی محاسبه می‌شوند و هرچه مقدار بهره اطلاعاتی بیشتر باشد، آن مشخصه به عنوان مشخصه پر اهمیت‌تر امتیازبندی می‌شود (Hall & et al., 1999).

باتیتی[1] (۱۹۹۴) و لوئیس[2] (۱۹۲۲) اولین محققانی بودند که پیشنهاد استفاده از اطلاعات متقابل در انتخاب مشخصه مبتنی بر بهره اطلاعاتی را دادند. لوئیس (۱۹۹۲) از اطلاعات متقابل برای انتخاب مشخصه و استخراج مشخصه در تقسیم‌بندی متن استفاده کرد. لوییس (۱۹۹۲) کاربرد معیار اطلاعات متقابل برای ارزیابی مجموعه‌ای از مشخصه‌های سودمند با استفاده از کلاس‌بند KNN را بررسی کرد. بیش از ۲۰ سال بعد از ارائه تئوری اطلاعات متقابل، روش‌های انتخاب مشخصه مختلفی براساس اطلاعات متقابل ارائه گردید و به‌طور گسترده روی انواع مختلف داده‌ها به‌کار گرفته شده است.

1. Battiti
2. Lewis

الگوریتم انتخاب مشخصه مبتنی بر اطلاعات متقابل[۱] از زیرمجموعه‌های روش مبتنی بر بهره اطلاعات، توسط باتیتی (۱۹۹۶) ارائه شد (Blum & et al., 1997). اساس این الگوریتم روی بیشینه کردن اطلاعات متقابل بین متغیر کاندید شده و متغیر هدف، و حداقل کردن حشو بین متغیرهای کاندید شده و متغیرهای انتخاب شده پیشین است. الگوریتم سیلوسی[۲] در سال ۲۰۰۵ توسط سیلوسی (Xu & et al., 2011) ارائه شد. کواک و همکاران[۳] (۲۰۰۲) الگوریتمی به نام MIFS-U برای حل محدودیت‌ها و بهبود MIFS پیشنهاد داد تا اطلاعات متقابل بین مشخصه‌های ورودی و کلاس هدف، بهتر از MIFS محاسبه شود. هر دو نسخه MIFS و MIFS-U یک مشکل مشترک داشتند: با افزایش تعداد مشخصه‌های انتخاب شده، جداسازی مشخصه‌های مرتبط، افزایش می‌یافت (Kwak & et al., 2002). لونگ و همکاران[۴] (۲۰۰۵) روش انتخاب مشخصه مبتنی بر اطلاعات متقابل را ارائه داد که نام آن mRmR[۵] بود. این روش افراط را بین مشخصه‌ها به مینیمم می‌رساند و ارتباط بین مشخصه‌های انتخاب شده و کلاس هدف را ماکزیمم می‌کرد (Peng & et al., 2005).

الگوریتم سلیوسی[۶] بر مبنای این فرض ساخته شده است که سری‌های ویژگی‌ها و توابع هدف، از نظر آماری مستقل هستند. اگرچه الگوریتم MIFS به اثرات تعداد متغیرهای انتخاب شده، توجه نمی‌کند، اما وقتی تعداد متغیرهای انتخاب شده، افزایش می‌یابد، اثر ویژگی‌ها بر تابع هدف کاهش یافته و مقادیر کمتر عددی را شامل می‌شود. برای دوری کردن از این شرایط، الگوریتم MMIFS توسط امیری و همکاران (۲۰۱۱) ارائه شد.

یک روش انتخاب مشخصه غیرنظارتی[۷] نیز توسط فلورت[۸] (۲۰۰۴) ارائه شد تا افراط میان مشخصه‌ها را از بین ببرد. از معیار جدیدی به نام حداکثر شاخص فشرده‌سازی اطلاعات[۹] برای محاسبه تشابه بین دو مشخصه تصادفی، در فرآیند انتخاب مشخصه استفاده شد. در روش حداکثرسازی اطلاعات متقابل،[۱۰] مشخصه‌ها به‌ترتیب نزولی و براساس میزان اطلاعات متقابل بین

1. Mutual Information Feature Selection-MIFS
2. Celluci
3. Kwak & Choi
4. Peng, Long & Ding
5. Max-Relevance and Min-Redundancy
6. Cellucci Algorithm
7. Unsupervised
8. Fleuret
9. Maximal information compression index
10. Mutual Information Maximization

مشخصه و کلاس هدف، مرتب می‌شوند. سپس مشخصه‌هایی با بالاترین میزان اطلاعات متقابل را انتخاب می‌نماید.

یانگ و مودی[1] (۲۰۰۰) روشی به نام (JMI)[2] پیشنهاد دادند. برخلاف سایر روش‌ها، این روش برای تخمین اهمیت یک مشخصه JMI و اثر مکمّل[3] بین مشخصه کاندید شده و مشخصه‌های قبلاً انتخاب شده و کلاس هدف را در نظر می‌گرفت. در این روش، مشخصه کاندید شده که اثر مکمّل JMI با سایر مشخصه‌های انتخاب شده ماکزیمم باشد، انتخاب شده و به زیرمجموعه اضافه می‌شود. JMIM[4] و NJMIM[5] در سال ۲۰۱۵ ارائه شدند. JMIM مسئله افراط در انتخاب مشخصه حل می‌کند (Hicks & et al., 2015). این روش از معیارهای ماکزیمم کردن مینیمم و JMI برای حذف و کاهش مشخصه‌های نامربوط و حشو استفاده می‌کند.JMIM مشابه روش CMIM است. مهم‌ترین تفاوت بین آن‌ها این است که CMIM، CMI بین مشخصه کاندید شده و کلاس هدف را ماکزیمم می‌کند (با در اختیار داشتن مشخصه‌های از قبل انتخاب شده)، در حالی‌که JMIM مشخصه‌هایی را انتخاب می‌کند که MIJ مشخصه‌های قبلاً انتخاب شده را افزایش می‌دهد. NJMIM نوعی JMIM است که MI را نرملایز می‌کند.

MIFS-ND نوعی MIFS است که با یک روش بهینه‌سازی تلفیق شده است. این روش، اطلاعات متقابل بین مشخصه کاندید شده و کلاس هدف را محاسبه می‌کند، همچنین متوسط اطلاعات متقابل بین مشخصه کاندید شده و مشخصه‌های قبلاً انتخاب شده را محاسبه می‌کند. یک الگوریتم ژنتیک برای انتخاب مشخصه‌هایی که اطلاعات متقابل آن‌ها با کلاس هدف ماکزیمم بوده و متوسط اطلاعات متقابل آن‌ها با سایر مشخصه‌های انتخاب شده مینیمم است، به‌کار گرفته شده است (هوگوا و همکاران، ۲۰۱۴). آنها در سال ۲۰۱۶ این روش را با روش فازی ارتقاء داده و روشی بنام FMIFS-ND معرفی نمودند. همچنین از طبقه‌بندی KNN-ND برای تخمین نتایج استفاده کردند. تفاوت این طبقه‌بندی با سایر طبقه‌بندی‌هایKNN در این است که KNN-ND میزان مشابهت بین هر دو نمونه را مبتنی بر وزن هر مشخصه، محاسبه می‌کند، در حالی‌که KNN وزن مشخصه‌ها را در نظر نمی‌گیرد (Hoque & et al., 2016).

1. Yang & Moody
2. Joint Mutual information
3. Complementary Effect
4. Joint Mutual Information Maximization
5. Normalized Joint Mutual Information

ژائو و همکاران[1] (۲۰۱۸) یک روش انتخاب مشخصه خطی مبتنی بر اطلاعات متقابل به نام DCSF[2] معرفی کردند. در این روش CMI بین مشخصه‌های انتخاب شده و کلاس هدف، و مشخصه کاندید شده، محاسبه می‌شود. این روش تغییرات اطلاعاتی داینامیک بین مشخصه‌های انتخاب شده و کلاس را در نظر می‌گیرد.

با توجه به مطالعات پیشین، در این پژوهش تلاش شده است تا رابطه‌ای در راستای استفاده از بهره اطلاعاتی ارائه شود. در این رابطه سعی شده است تا رابطه بین ویژگی‌ها و همچنین تابع هدف به صورت صحیح و نزدیک به واقعیت ارائه شود. از طرفی، همان‌گونه که بیان شد، یکی از مهم‌ترین مشکلات روش‌های زیرشاخه بهره اطلاعاتی این است که با بالا رفتن تعداد ویژگی‌ها، مقایسات زوجی ویژگی‌ها و توابع هدف هر نمونه، دارای بار محاسباتی بالایی است. از این‌رو برای رفع اشکال نیز از الگوریتم مبتنی بر جمعیت ژنتیک استفاده شده است، تا از بررسی تمامی پاسخ‌ها که زمان زیادی را می‌طلبد، خودداری شود و بتوان با بررسی سیستماتیک پاسخ‌های مناسب، به سمت پاسخ بهینه حرکت کرد. در ادامه به بیان دقیق الگوریتم مذکور پرداخته شده است.

۳. روش پیشنهادی

در این قسمت ابتدا یک معیار جدید برای انتخاب مشخصه‌های مهم و تأثیرگذار در الگوریتم MIFS معرفی شده است. پس از ارائه معیار پیشنهادی، برای کاهش زمان حل، این معیار توسط الگوریتم ژنتیک به صورت معیار اندازه‌گیری تابع برازندگی، در راستای انتخاب مشخصه‌های تأثیرگذار عمل می‌نماید. با پیاده‌سازی این روش می‌توان سرعت همگرایی الگوریتم‌های MIFS را افزایش داد.

۳-۱. معیار انتخاب مشخصه

اطلاعات متقابل و یا به اختصار MI برای اولین بار در سال ۱۹۴۸ توسط شانن[3] معرفی شد. MI یک معیار کمّی است که میزان اطلاعات مشترک بین دو متغیر تصادفی را اندازه‌گیری می‌کند. هر اندازه MI بین دو متغیر بیشتر باشد، به این معنا است که به یکدیگر شبیه‌تر و وابسته‌تر هستند. این معیار به طور گسترده برای تعیین پارامتر مهمی به نام تأخیر زمانی در سری‌های زمانی غیرخطی مورد

1. Gao, Hu & Zhang
2. Dynamic Change of Selected Feature with the class
3. Shannon

استفاده قرار گرفته و توضیح ریاضی آن در ادامه آمده است. با در نظر گرفتن یک سری زمانی غیرخطی مانند $\{S(t_i)\}$ $(i = 1, 2, \dots, N)$، سری‌های دیگری که با تأخیر زمانی τ ایجاد می‌شوند، به این صورت خواهد بود:

$$\begin{matrix} (S(t_1), S(t_1+\tau), \dots, S(t_1+(d-1)\tau) \\ \vdots \\ (S(t_N), S(t_N+\tau), \dots, S(t_N+(d-1)\tau) \end{matrix} \quad (۱)$$

MI بین سری زمانی $S = \{S(t_1), S(t_2), \dots, S(t_N)\}$ و $Q = \{S(t_1+\tau), S(t_2+\tau), \dots, S(t_N+\tau)\}$ که به صورت $I(S,Q)$ نمایش داده می‌شود، تعداد بیت‌هایی است که سری S می‌تواند با در اختیار داشتن اطلاعات سری Q پیش‌بینی کند، که به صورت زیر تعریف می‌شود:

$$I(S;Q) = H(S) + H(Q) - H(S,Q) \quad (۲)$$

H(S) و H(Q) به ترتیب آنتروپی، Q و S و $H(S,Q)$ آنتروپی متقابل S و Q است. برای محاسبه MI بین این دو سری زمانی مقادیر عددی S و Q روی دو بردار عمودی به نامSQ plane نمایش داده می‌شود (شکل ۱).

فاصله بین مقادیر ماکزیمم و مینیمم به تعداد مقادیر $\{q_i\}, (i = 1,2, \dots.)$ تقسیم می‌شود. این تقسیم براساس استانداردها و شاخص‌های مختلفی صورت می‌گیرد. توزیع احتمالی نقاط در SQplane یعنی $P(q_i)$ به جفت- داده‌هایی در q_i اشاره دارد که با تقسیم بر اندازه سری زمانی (یعنی تعداد مقادیر) N، به دست می‌آید. تابع توزیع احتمالاتی (s_i, q_i)، $p(s_i, q_i)$، میزان نقاط در (s_i, q_i) تقسیم بر N است.

$$H(Q) = -\sum P(q_i) log_2 P(q_i) \quad (۳)$$

$$H(s) = -\sum P(s_i) log_2 P(s_i) \quad (۴)$$

$$H(Q,S) = -\sum P(s_i, q_i) log_2 P(s_i, q_i) \quad (۵)$$

در نهایت MI بین S و Q به صورت زیر محاسبه می‌شود:

$$I(Q,S) = \sum_i \sum_j P_{s,q}(s_i, q_i) log_2 \frac{P_{sq}(s_i, q_i)}{P_s(s_i) P_q(q_i)} \quad (۶)$$

سلیوسی[1] در سال ۲۰۰۵ این معیار را توسعه داد. او فرض کرد که سری‌های S و Q مستقل هستند و معیار محاسباتی جدیدی برای MI نوشت:

1. Cellucci

$$I(Q,S)=\sum_{i=1}^{N_E}\sum_{j=1}^{N_E}P(s_i,q_j)log_2(N_E^2P(s_i,q_j)) \quad (۷)$$

که N_E بزرگ‌ترین عدد صحیحی است که در معادله زیر صدق می‌کند:

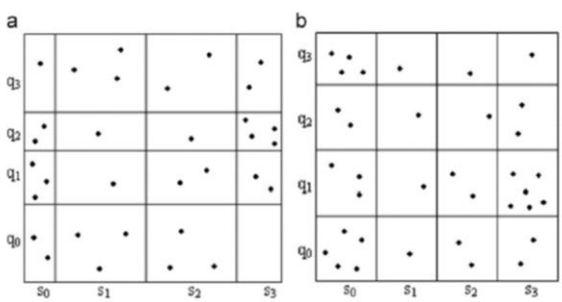
$$N_E \le \frac{N^{1/2}}{5} \quad (۸)$$

کمی بعد در سال ۲۰۱۰، سلیوسی این الگوریتم را توسعه داده و سه معیار مختلف برای محاسبه MI پیشنهاد داد. این معیار برای مجموعه داده‌های بزرگ، در زمان کمتری محاسبات را انجام می‌داد. در این روش SQ plane به مقادیر یکسان احتمالاتی روی هر دو بردار تقسیم می‌شود و فقط به ترتیب قرار گرفتن مقادیر عددی در سری‌های زمانی بستگی دارد، و نه به خود مقادیر عددی؛ یعنی هر تغییر در مقادیر عددی سری‌های زمانی، فقط تا زمانی پذیرفته است که ترتیب مقادیر عددی حفظ شود. در این روش ابتدا تمام مقادیر عددی سری‌های زمانی S و Q مرتب می‌شوند و بدون توجه به اینکه حاوی چه مقادیری هستند، به‌ازای آن‌ها دو سری A و B ایجاد می‌شود که به ترتیب عدد صحیح 1 تا N در آن قرار می‌گیرد. بنابراین، محاسبه توزیع احتمال روی A و B به مراتب ساده‌تر از S و Q است. در مجموعه داده A(i), B(i) (i=1,…N) برای تعیین مکان (A(i), B(i)) روی SQ plane ماتریس C(m,n) محاسبه می‌شود که به آن احتمال وقوع گفته می‌شود. m بزرگ‌ترین عدد صحیح کوچک‌تر مساوی $1+A(i)/N_1$ است و n بزرگ‌ترین عدد صحیح کوچک‌تر و یا مساوی $1+B(i)/N_1$ است.

در این پژوهش از معیار زیر برای محاسبه MI بین مشخصه‌ها استفاده شده است:

$$I(Q,S)=\sum_{m=1}^{k}\sum_{n=1}^{k}\frac{C(m,n)}{N^*}\log_2[\frac{C(m,n)N^*}{N_1^2}] \quad (۹)$$

که N^* مساوی N_1k است و N_1 برابر بزرگ‌ترین عدد صحیح کوچک‌تر مساوی $\frac{N}{N_1}$



شکل ۱- نمودار SQ-plan دو سری Q و S

جهت افزایش سرعت معیار پیشنهادی، این معیار با یک الگوریتم فراابتکاری مبتنی بر الگوریتم ژنتیک ترکیب شده است. در این حالت می‌توان مسائلی با ابعاد بزرگ را در زمان مناسب حل کرد.

نمای کلی روش حل پیشنهادی در شکل (۲) نشان داده شده است.

C-MIFS Algorithm

Input: Set F of n features
Output: Set S of k features
For feature $f \in F$
Calculate $I(c;f)$
End
Find first feature f that maximizes $I(c;f)$;
Set $F \leftarrow F\backslash\{f\}$
Set $S \leftarrow \{f\}$
While$|S| < k$do
Choose feature$f$ as the one that maximizes:

$$I(Q,S) = \sum_{m=1}^{k}\sum_{n=1}^{k}\frac{C(m,n)}{N^*}log_2\left[\frac{C(m,n)N^*}{N_1^2}\right]$$

Which

$$N^* = N_1.k$$

$$k\ is\ the\ \max integer\ which\ satisfying\ k \leq {}^{N}/_{N_1}$$

End

شکل ۲- نمای کلی روش پیشنهادی

در محاسبات، فرض بر این است که $N_1 = N$ و در نتیجه k=1 خواهد بود.

۳-۲. الگوریتم ژنتیک جهت انتخاب مشخصه

الف) الگوریتم ژنتیک: با توجه به این‌که برای ابعاد بزرگ مسئله، روش‌های دقیق کارا نیست، و نمی‌توان از تمام مشخصه‌ها جهت طبقه‌بندی استفاده کرد و زمان حل به صورت نمایی بالا می‌رود، بنابراین، از الگوریتم‌های فراابتکاری برای حل مسئله در ابعاد بزرگ استفاده می‌شود. الگوریتم ژنتیک به عنوان یکی از روش‌های تصادفی بهینه‌سازی شناخته شده، توسط جان هالند در سال ۱۹۶۷ ابداع شده است. بعدها این روش با تلاش‌های گلدبرگ (۱۹۸۹)، توسعه یافته و امروزه نیز به واسطه توانایی‌های آن، جای مناسبی در میان دیگر روش‌ها دارد.

الگوریتم ژنتیک یکی از انواع الگوریتم‌های تکاملی است که از علم زیست‌شناسی مثل وراثت، جهش و ترکیب الهام گرفته شده است. در الگوریتم ژنتیک، ابتدا به صورت تصادفی یا الگوریتمیک، چندین جواب برای مسئله تولید می‌شود. این مجموعه جواب را جمعیت اولیه می‌نامیم. هر جواب را یک کروموزوم می‌نامیم. سپس با استفاده از عملگرهای الگوریتم ژنتیک، و پس از انتخاب

کروموزوم‌های بهتر، کروموزوم‌ها را با هم ترکیب کرده و جهشی در آنها ایجاد می‌کنیم. در نهایت نیز جمعیت فعلی را با جمعیت جدیدی که از ترکیب و جهش در کروموزوم‌ها حاصل می‌شود، ترکیب می‌کنیم. این فرایند تا زمان برقراری شرایط توقف ادامه می‌یابد.

ب) نحوه تولید جواب اولیه: تعداد جواب اولیه برابر با ۱۰۰ است. جهت تولید جواب اولیه، ابتدا تعداد کل مشخصه‌های مورد نیاز را در نظر می‌گیریم. سپس تعدادی مشخصه به صورت تصادفی از بین تمام مشخصه‌ها انتخاب می‌شوند. تعداد مشخصه‌های انتخاب شده برابر با یک عدد تصادفی بین ۱/۲ و ۲/۳ تعداد کل مشخصه‌ها است. به عنوان مثال، اگر ۶۰ مشخصه موجود باشد، تعداد مشخصه‌های انتخاب شده برابر با یک عدد تصادفی بین ۳۰ و ۴۰ عدد است که این تعداد مشخصه، به صورت تصادفی از بین تمام مشخصه‌ها انتخاب می‌شود.

ج) نمایش کروموزوم: در الگوریتم ژنتیک، هر کروموزوم نشان‌دهنده یک نقطه در فضای جستجو و یک راه‌حل ممکن برای مسئله مورد نظر است. کروموزوم انتخابی از ۲ رشته تشکیل شده است. رشته اول برابر با تعداد مشخصه‌هایی بوده که باید انتخاب شوند، و رشته دوم نیز بیانگر یک چیدمان تصادفی از کل مشخصه‌ها است که با استفاده از رشته اول، مشخصه‌های انتخابی را معین می‌کنیم. برای مثال فرض کنید که تعداد ۱۰ مشخصه داریم. شکل (۳) کروموزوم استفاده را نشان می‌دهد.

رشته اول	۷									
رشته دوم	۸	۳	۲	۱	۵	۹	۱۰	۴	۶	۷

شکل ۳- نحوه نمایش کروموزوم

به‌منظور حل هر مسئله با استفاده از الگوریتم ژنتیک، ابتدا باید یک تابع برازندگی برای آن مسئله تعریف شود. برای هر کروموزوم، این تابع عددی را بازمی‌گرداند که نشان‌دهنده شایستگی یا توانایی فردی آن کروموزوم است. نحوه محاسبه تابع برازندگی به صورت زیر است. شاخص زیر برای کروموزوم‌های مسئله لحاظ می‌شود.

د) تابع برازندگی: به‌منظور حل هر مسئله با استفاده از الگوریتم ژنتیک، ابتدا باید یک تابع برازندگی برای آن مسئله تعریف شود. برای هر کروموزوم، این تابع عددی را بازمی‌گرداند که نشان‌دهنده شایستگی یا توانایی فردی آن کروموزوم است. نحوه محاسبه تابع برازندگی به صورت زیر است. شاخص زیر برای کروموزوم‌های مسئله لحاظ می‌شود.

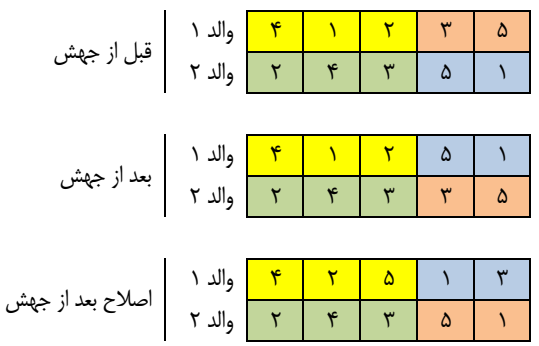
$$I(Q,S)=\sum_{m=1}^{k}\sum_{n=1}^{k}\frac{C(m,n)}{N}log_2\left[\frac{C(m,n)}{N}\right] \qquad (۱۰)$$

برای هر کروموزوم، معیار بالا اندازه‌گیری شده و به عنوان تابع برازندگی آن کروموزوم در نظر گرفته می‌شود.

ر) مکانیزم انتخاب: مکانیزم انتخاب در نظر گرفته شده، چرخه رولت[1] است. در این روش کروموزوم‌هایی که عدد برازش (تناسب) بیشتری داشته باشند، انتخاب می‌شوند. در واقع به نسبت عدد برازش برای هر جواب، یک احتمال تجمعی نسبت می‌دهیم و با این احتمال است که شانس انتخاب هر جواب تعیین می‌شود.

ز) ساختار روش تقاطع: به‌وسیله چرخه رولت، تعداد مشخصی کروموزوم انتخاب می‌شود. برای عملیات تقاطع، از روش تقاطع نقطه‌ای استفاده شده است.

تقاطع: در این حالت عملیات تقاطع روی رشته دوم کروموزوم انجام می‌شود. برای رشته دوم به‌صورت جداگانه یک نقطه به‌صورت تصادفی انتخاب شده و سپس عملیات تقاطع انجام می‌شود. پس از اعمال تقاطع، برای آنکه مشخصه تکراری نداشته باشیم، مشخصه‌های موجود یک بار دیگر به صورت تصادفی در انتهای کروموزوم چیده شده و سپس موارد تکراری حذف می‌شوند. شکل (۴) عملیات تقاطع روی دو والد را نشان می‌دهد.



شکل ٤ – عملیات تقاطع در رشته اول

ط) ساختار روش جهش: استفاده از عملگر جهش، از بهینگی زودرس و افتادن در بهینگی محلی جلوگیری می‌کند. برای جهش از روش تعویض تصادفی استفاده شده است.

جهش ۱: عمل جهش در این حالت بر روی رشته دوم کروموزوم انجام می‌شود. در این حالت یک کروموزوم به صورت تصادفی از بین کروموزوم‌های موجود انتخاب می‌شود. سپس از عمل

1. Roulette wheel

تعویض تصادفی دوتایی جهت عمل جهش استفاده می‌شود. شکل (۵) نحوه انجام عمل جهش را نشان می‌دهد.

شکل ۵- عمل جهش در روی کروموزوم

جهش ۲: در این حالت عمل جهش بر روی رشته اول کروموزوم انجام می‌شود. یک کروموزوم به صورت تصادفی انتخاب شده و رشته اول آن با یک عدد تصادفی تعویض می‌شود.

هـ) تولید مجدد: در هر تکرار، جواب‌های جدیدی که از طریق عملگرهای ژنتیک تولید شده‌اند، با جواب‌های اولیه ترکیب شده و به اندازه جمعیت اولیه، بهترین جواب‌ها که به واسطه ترتیب‌بندی کلیه جواب‌ها در هر تکرار حاصل می‌شوند، نگه داشته می‌شوند.

ی) شرط توقف الگوریتم: شرط توقف الگوریتم، ترکیبی از تعداد تکرار و همگرا شدن است. در واقع بیشینه تعداد تکرار ۴۰۰ است که از تکرار ۱ تا ۱۰۰ عملیات تقاطع و جهش، فقط بر روی رشته دوم اعمال می‌شود. از تکرار ۱۰۰ تا ۲۰۰ الگوریتم، تنها جهش بر روی رشته اول کروموزوم زده می‌شود و از تکرار ۲۰۰ تا ۳۰۰ الگوریتم، تنها تقاطع و جهش بر روی رشته اول و دوم کروموزوم به صورت همزمان زده می‌شود. برای همگرا شدن، اگر ۱۰۰ عدد از بهترین جواب‌هایی که الگوریتم یافت، مشابه باشند، الگوریتم متوقف می‌شود. درواقع هرکدام از این دو شرط که زودتر برقرار شود، الگوریتم متوقف خواهد شد.

۳-۳. تنظیم پارامترها به روش تاگوچی

روش تاگوچی انحراف‌های ممکن از مقدار هدف را همراه با تابع زیان، مدل‌بندی می‌کند.

در الگوریتم ژنتیک، دو پارامتر تعداد جمعیت و همچنین بیشینه تعداد تکرار الگوریتم، به عنوان پارامترهای کلیدی شناسایی شدند. این دو به عنوان پارامترهایی تاثیرگذار در زمان حل مسئله و تابع هدف نهایی مسئله اثر دارند. افزوده و یا کاسته شدن این دو پارامتر می‌تواند پاسخ نهایی را دستخوش تغییر کند. لذا این دو پارامتر جهت بررسی زمان حل و تابع هدف تحلیل می‌شوند.

جدول ۱- سطوح در نظر گرفته شده برای الگوریتم

پارامتر الگوریتم	سطح ۱	سطح ۲	سطح ۳
تعداد اعضای جمعیت	۱۰۰	۱۵۰	۲۰۰
تعداد تکرار الگوریتم	۵۰	۱۰۰	۱۵۰

پس از در نظر گرفتن تابع هدف و زمان حل، و با طراحی آزمایش تاگوچی، نتایج به شرح جدول (۲) می‌باشد.

جدول ۲- نتایج آزمایشات

پارامتر الگوریتم	سطح پارامتر تعداد جمعیت	سطح پارامتر تعداد تکرار الگوریتم	تابع هدف	زمان حل (ثانیه)
آزمایش ۱	۱	۱	۱۴.۶۲	۱۲.۳۶
آزمایش ۲	۱	۲	۱۵.۳۶	۱۲.۴۸
آزمایش ۳	۱	۳	۱۵.۵۱	۱۳.۶۲
آزمایش ۴	۲	۱	۱۵.۳۹	۱۲.۸۱
آزمایش ۵	۲	۲	۱۵.۴۸	۱۳.۷۱
آزمایش ۶	۲	۳	۱۵.۸۶	۱۴.۰۱
آزمایش ۷	۳	۱	۱۵.۶۲	۱۴.۶۳
آزمایش ۸	۳	۲	۱۵.۹۱	۱۴.۸۹
آزمایش ۹	۳	۳	۱۶.۰۱	۱۵.۳۶

معادلات رگرسیونی به صورت زیر است.

obj = 1 + 4.2 A + 5.1 B

time = -28.6667 + 50.3333 A + 82.5 B

در این دو معادله، A و B به ترتیب نشان‌دهنده تعداد اعضای جمعیت و تعداد تکرار الگوریتم می‌باشند. با حل این معادلات و با در نظر گرفتن محدودیت محدوده حد بالا و پایین برای هر متغیر کنترلی می‌توان مقدار بهینه آن را بدست آورد.

۴. نتایج

در این بخش بهترین زیرمجموعه انتخاب شده از مشخصات، توسط روش‌های مختلف انتخاب مشخصه از نظر دقت مدل کلاسبند، ارزیابی می‌شوند. عملکرد کلاسبند با معیارهای sensitivity, specificity, accuracy, precision و area under the curve (AUC) بررسی می‌شود. این موارد به صورت زیر تعریف می‌شوند:

$$Sensitivity = \frac{TP}{TP + FN} \qquad (۱۱)$$

$$Specifity = \frac{TN}{TN + FP} \qquad (۱۲)$$

$$Precision = \frac{TP}{TP + FP} \qquad (۱۳)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (۱۴)$$

ماتریس درهم ریختگی، یک ماتریس مربعی است که چگونگی عملکرد الگوریتم کلاسبندی را با توجه به مجموعه داده ورودی، به تفکیک انواع دسته‌های مسأله کلاسبندی، نمایش می‌دهد. این ماتریس به صورت زیر است:

	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

هر یک از عناصر ماتریس به شرح ذیل می‌باشد:

TN: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها را بدرستی منفی تشخیص داده است.

TP: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم کلاسبندی نیز دسته آن‌ها را بدرستی مثبت تشخیص داده است.

FP: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم کلاسبندی، دسته آن‌ها را به اشتباه مثبت تشخیص داده است.

FN: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم کلاسبندی، دسته آن‌ها را به اشتباه منفی تشخیص داده است.

۴-۱. نتایج مقایسه روش پیشنهادی با الگوریتم‌های انتخاب مشخصه کلاسیک

به منظور ارزیابی عملکرد مدل انتخاب مشخصه، از مجموعه داده‌های مختلف استفاده شده است. جدول (۳) جزئیات این داده‌ها را نشان می‌دهد.

جدول ۳- جزئیات مجموعه داده‌ها

شماره	عنوان مجموعه داده	تعداد کلاس‌ها	مشخصه‌ها	نمونه‌ها
1	Wine	3	13	178
2	Iris	3	4	150
3	Liver	2	6	345
4	Sonar	2	60	208

این مجموعه داده‌ها بر روی *Chi-squared, Gain Ratio, Information Gain* و Relief و روش پیشنهادی اعمال شده است. جدول (۴) ترتیب انتخاب هر مشخصه را از پراهمیت‌ترین (دارای وزن بیشتر)، به کم اهمیت‌ترین (دارای وزن کمتر)، توسط هر الگوریتم نشان می‌دهد. این جزئیات صرفاً جهت نمونه عملکرد روش‌های انتخاب مشخصه، برای مجموعه داده‌های پرکاربرد wine و iris آورده شده است.

جدول ٤- ترتیب انتخاب مشخصه‌ها

مجموعه داده	Info gain	relief	Gain ratio	Chi-squared	Proposed
Iris	4,3,1,2	4,3,1,2	4,3,1,2	4,3,1,2	3,4,2,1
Sonar	11,12,9,10,13,48, 49,51,47,45	12,11,10,45,48, 9,36,49,13	11,12,49,19,17, 9,47,45,30,13	12,11,10,13,49, 36,9,46,21,48	12,11,10,9,13,48, 49, 51,47,46,36,52

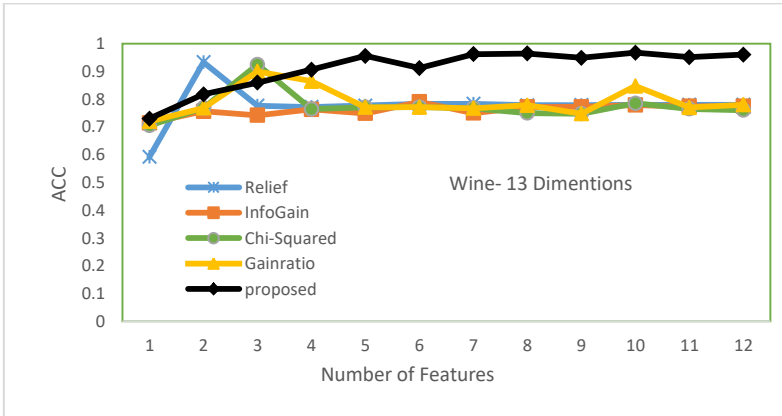
نسبت مجموعه آموزشی و آزمایشی برای کلاسبند KNN (K=3) ۷۵ به ۲۵ درصد انتخاب شده است. از اپراتور k-fold cross validation برای اعتبارسنجی استفاده شده است. پس از ۱۰ بار تکرار، نتایج به‌عنوان دقت نهایی کلاسبندی در نظر گرفته شد.

برطبق شکل‌های (۶ و۷) روش پیشنهادی نسبت به روش‌های سنّتی، عملکرد نسبتاً بهتری در طبقه‌بندی مجموعه داده‌های Wine, Liver, Iris, و Sonar داشته است. در شکل (۸ و ۹) عملکرد روش پیشنهادی با روش‌های JMI, NJMIM, MRMR, DISR که مبتنی بر تئوری اطلاعات هستند، مقایسه شده است.

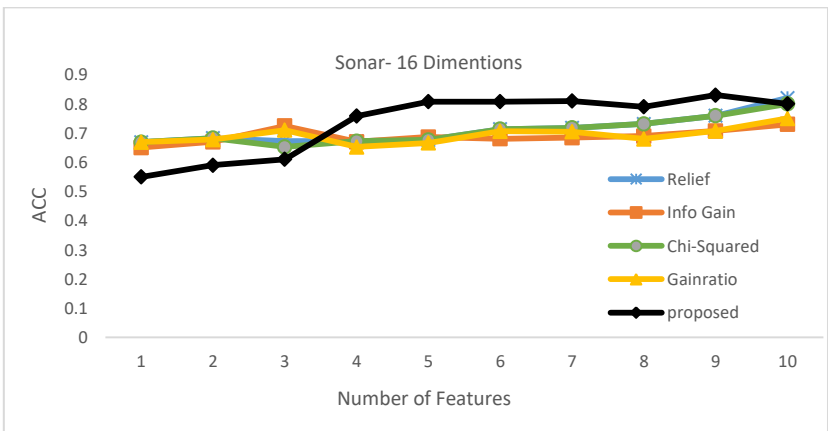
میانگین بهترین دقت‌های به دست آمده در الگوریتم‌های مختلف، و با در نظر گرفتن تعداد متفاوت مشخصه‌ها، در جدول (۵) نشان داده شده است.

جدول ٥- میانگین دقت کلاسبندی در الگوریتم‌های مختلف (N بهترین دقت)

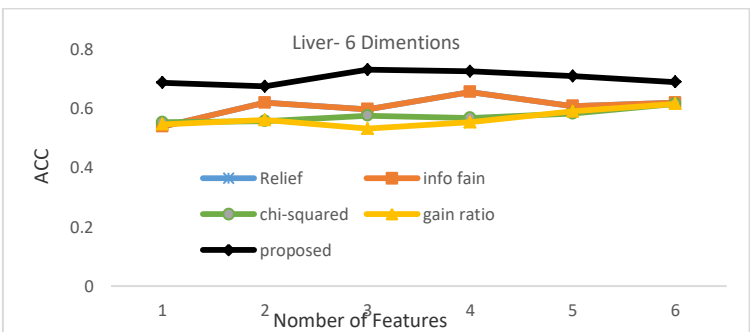
مجموعه داده	Chi squared		Gain Ratio		Information Gain		Relief		proposed		بهبود نسبت به (۱)	بهبود نسبت به (۲)	بهبود نسبت به (۳)	بهبود نسبت به (٤)
	دقت	N	دقت	N	دقت	N	دقت	N	دقت	N				
sonar	70	10	70	10	69	3	71	10	73	9	3	3	4	2
Wine	77	10	79	3	76	6	77.5	2	91	7	14	12	15	13.5
liver	57	6	57	6	60	6	60	6	70	6	13	13	10	10
Iris	92	2	91	2	93	2	92	2	94	2	2	3	1	2



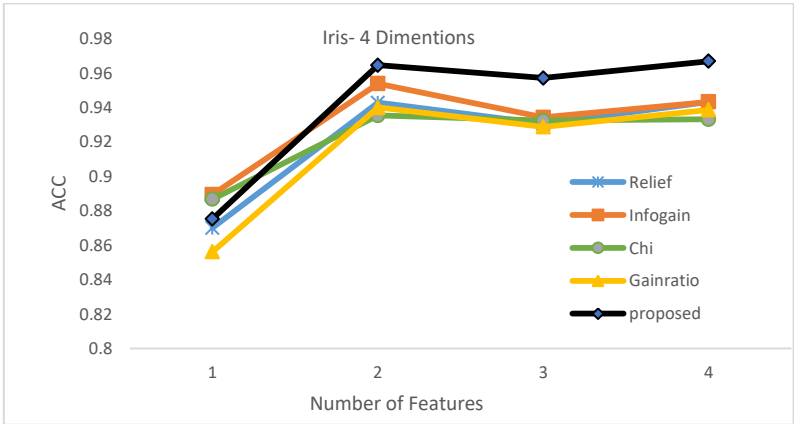
شکل ٦- نمودار دقت کلاسبندی در داده Wine



شکل ۷- نمودار دقت کلاسبندی در داده Sonar



شکل ۸- نمودار دقت کلاسبندی در داده Liver



شکل ۹- نمودار دقت کلاسبندی در داده Iris

برطبق نمودار و جدول‌های بدست آمده، عملکرد مدل پیشنهادی نسبت به الگوریتم‌های کلاسیک انتخاب مشخصه بهتر بوده و بسته به مجموعه داده بین مدل پژوهش حاضر و مدل‌های قدیمی، از ۲ تا ۱۰ درصد اختلاف دقت وجود دارد.

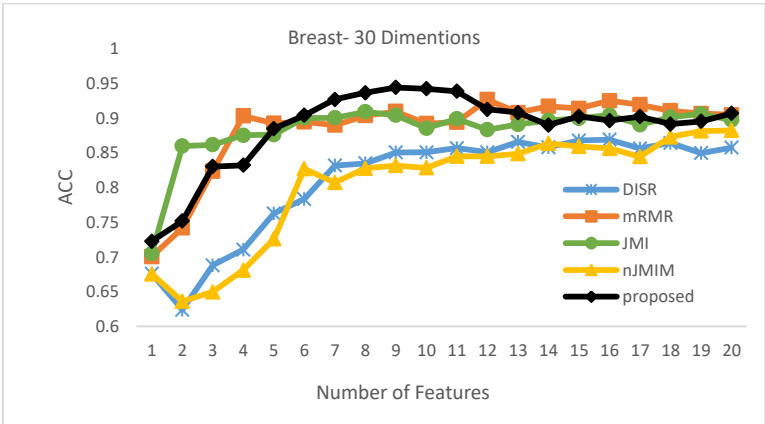
۴-۲. مقایسه نتایج با الگوریتم‌های انتخاب مشخصه مبتنی بر اطلاعات متقابل

در این قسمت، روش پیشنهادی با چهار روش مبتنی بر نظریه اطلاعات به نام‌های mRmR, DISR, JMI, NJMIM مقایسه شده است. مجموعه داده‌ها در جدول (۶) آمده است.

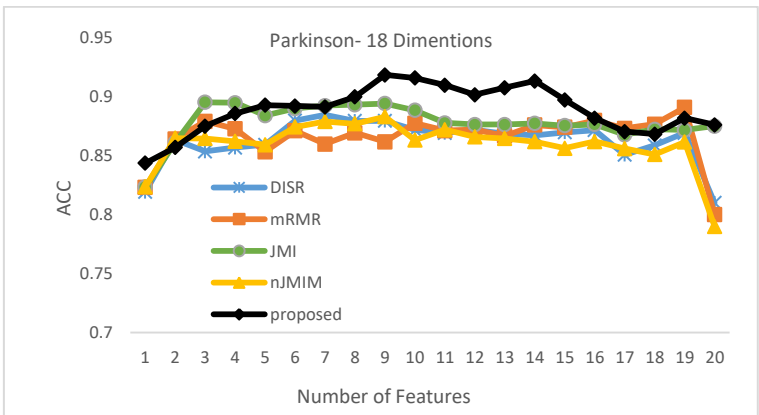
همان‌طور که در شکل‌های (۱۰، ۱۱، ۱۲) شاهده می‌شود، در مجموعه داده‌های سنسور گاز، پارکینسون و سرطان سینه بهترین دقت به دست آمده است. متوسط دقت به ترتیب ۷۶، ۸۸٫۶ و ۸۸٫۸ درصد بوده است. اگرچه روی داده‌های سونار متوسط، دقت روش پیشنهادی از تمام مدل‌ها پایین‌تر بوده، اما دقت مدل بعد از پنج مشخصه بالا رفته و به بهترین دقت یعنی ۸۰ درصد می‌رسد.

جدول ٦- جزئیات مجموعه داده‌های مورد استفاده

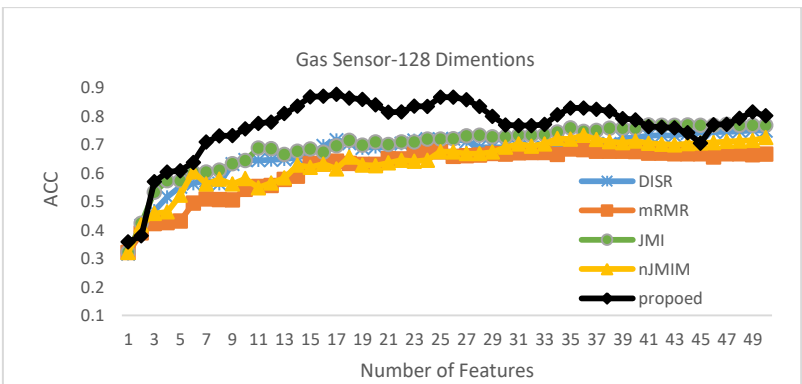
No.	Dataset Title	Classes	Features	Samples
1	sonar	2	60	208
2	Lymphoma	9	4026	96
3	Breast	2	30	569
4	parkinson	2	18	196
5	Gas sensor	6	128	13874



شکل ۱۰- دقت کلاسبندی در مجموعه داده Breast



شکل ۱۱- دقت کلاسبندی در مجموعه داده Parkinson



شکل ۱۲- دقت کلاسبندی در مجموعه داده Gas Sensor

جدول (۷) خلاصه نتایج دقت طبقه‌بندی داده‌ها را نشان می‌دهد.

جدول ۷- میانگین دقت کلاسبندی در الگوریتم‌های مختلف (N بهترین دقت)

مجموعه داده	mRMR		DISR		JMI		NJMIM		proposed		بهبود نسبت به (۱)	بهبود نسبت به (۲)	بهبود نسبت به (۳)	بهبود نسبت به (٤)
	دقت	N	دقت	N	دقت	N	دقت	N	دقت	N				
sonar	75	4	76.4	4	76.3	7	76.9	8	73.5	9	0	0	0	0
Breast	88.3	12	81	13	88.2	2	80.4	20	88.5	9	0.2	7.5	0.3	8.1
Parkinson	86.5	19	86.2	7	87.7	3	85.9	9	88.8	9	2.3	2.6	1.1	2.9
Gas sensor	61.3	21	66.8	17	69.4	11	63.9	33	76.7	17	15.4	9.9	7.3	13.8
ALL	**82.11**		**80.6**		**81.2**		**82.8**		**87.15**					

مقایسه نتایج با متوسط دقت کلاسبندی KNN با و بدون استفاده از الگوریتم‌های انتخاب مشخصه، در جدول (۸) نشان داده شده است. میزان موثر بودن روش پیشنهادی با این نتایج قابل مشاهده است.

جدول ۸- بهترین دقت کلاسبندیKNN با استفاده و بدون استفاده از الگوریتم‌های انتخاب مشخصه

مجموعه داده	با تمام مشخصه‌ها/ بهترین دقت	DISR	mRMR	JMI	NJMIM	proposed
Gas sensor	60.3	74.5	67	76	73	86
Parkinson	76.1	87	89	89	87	91
sonar	86	76.4	75	76.3	76.9	75.5
Breast	85	86	92.6	90	88	94
مجموعه داده	**با تمام مشخصه‌ها/ بهترین دقت**	**Relief**	**Info Gain**	**Chi-Squared**	**Gain ratio**	**proposed**
Wine	78.24	92	79	92	90	96
Iris	85.51	93.9	95	93.8	93.9	96
Sonar	86	82	73	80	75	83
Liver	56	62	62	61	61	73.5

متوسط دقت‌های به دست آمده از خروجی‌های روش پیشنهادی 65.32% ,74.51% ,70.88% و 58.20% می‌باشد. mRMR از نقطه‌نظر دقت به دست آمده، دومین مدل است، MRI, JMI, DISR در درجات بعدی قرار دارند. طبق نتایج بدست آمده، به جز در مورد مجموعه داده sonar که متوسط عملکرد روش پیشنهادی بهتر از DISR, JMI, NJMIM و مشابه mRmR بوده است، در مورد مجموعه داده‌های دیگر، دقت روش پیشنهادی بهتر از همه روش‌ها بوده؛ در برخی موارد نیز با بیش از

10% اختلاف است. از طرفی همانگونه که بیان شد، در انتخاب مشخصه انتخاب و ارائه معیار مناسب در راستای کاهش ابعاد، مجموعه داده حائز اهمیت است. از این‌رو در پژوهش حاضر معیار مناسبی جهت بیان میزان تأثیر بر روابط ویژگی‌ها، با تابع هدف مناسب ارائه شده است. نتایج نشان داده نیز حاکی از صحت‌سنجی نهایی ارائه این معیار مناسب بوده که دارای بالاترین دقت نسبت به سایر روش‌های موجود در ادبیات است.

۵. نتیجه‌گیری

در مواجهه با مجموعه داده‌های با ابعاد بالا، کاهش بُعد، یك گام پیش‌پردازشی مهم برای حصول دقت بالا، کارایی و مقیاس‌پذیری در اغلب مسائل اقتباس دانش از میان داده‌ها است. در این تحقیق برای کاهش بُعد داده‌ها، ابتدا رابطه‌ای ارائه شده است که می‌تواند روابط بین ویژگی‌ها و توابع هدف را مبتنی بر واقعیت در نظر بگیرد و از طرفی نیز پیچیدگی محاسبه را کاهش دهد. رابطه مذکور از روابط زیرمجموعه بهره اطلاعاتی است که در آن روابط بین ویژگی‌ها و تابع هدف، به صورت مقایسات زوجی تشخیص داده می‌شود. همچنین با توجه به زمان‌بر بودن روش پیشنهادی در ابعاد بالا، به دلیل افزایش تعداد ویژگی‌های اصلی و اولیه مسئله و افزایش مقایسات زوجی بین ویژگی‌ها و همچنین توابع هدف آن‌ها، از یک الگوریتم فراابتکاری مبتنی الگوریتم ژنتیک استفاده شده است، تا زمان انتخاب مشخصه‌های منتخب را کاهش دهد. پس از محاسبه معیار ارائه شده و تعیین ویژگی‌هایی که بیشترین تأثیر را در تشخیص مؤلفه هدف مسئله دارند، مشخصه‌های منتخب از مجموعه اصلی ویژگی‌ها تعیین می‌گردد. سپس ویژگی‌های انتخاب شده وارد کلاسبند KNN شده، تا دقت کلاسبندی داده‌ها با بعد انتخاب شده تعیین و اعتبارسنجی گردد. روش پیشنهادی با روش‌های mRMR, DISR, JMI, NJMIM در مجموعه داده‌های متفاوت اعمال شده است. متوسط دقت‌های به دست آمده از خروجی‌های روش پیشنهادی ۶۵٫۳۲ و ۷۴٫۵۱ و ۷۰٫۸۸ و ۵۸٫۲ درصد می‌باشد، که حاکی از کارآمدی روش پیشنهادی است. طبق نتایج بدست آمده، به جز در مورد مجموعه داده sonar که نتیجه‌ای بهتر از روش پیشنهادی داشته است، متوسط عملکرد روش پیشنهادی بهتر از DISR, JMI, NJMIM و مشابه mRmR بوده است. در مورد مجموعه داده‌های دیگر، متوسط دقت روش پیشنهادی بهتر از همه روش‌ها بوده است. روش پیشنهادی فوق می‌تواند در ترکیب با الگوریتم‌های یادگیری ماشین، دارای عملکرد بهتری شود. همچنین می‌توان از ترکیب روش‌های فراابتکاری جهت بهبود مسئله استفاده کرد.

References

Amiri, F. & et al. (2011). Mutual information-based feature selection for intrusion detection systems. *Journal of network and computer applications*, 33(4): 1184-119.

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4): 537-550.

Blum, A.L. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1): 245–271.

Boukharouba, A. & Bennia, A. (2017). Novel feature extraction technique for the recognition of handwritten digits. *Applied Computing and Informatics*, 13(1): 19-26.

Cellucci, C.J., Albano, A.M. & Rapp, P.E. (2005). Statistical validation of mutual information calculations: comparisons of alternative numerical algorithms. *Physical Review*, E 71: 1-14.

Dash, M. & Liu, H. (1997). Feature Selection for Classification. *Intelligent Data Analysis*, No. 1: 131-156.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5: 1531-1555.

Gao, W., Hu, L. & Zhang, P. (2018). Class-specific mutual information variation for feature selection. *Pattern Recognition*, 79: 328-339.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3: 1157-1182.

Hall, M.A. (1999). *Correlation-based Feature Selection for Machine Learning*. P.hD. Thesis. Philosophy at The University of Waikato, Hamilton, NewZealand.

Hicks, Y., Setchi, R. & Bennasar, M. (2015). Feature selection using Joint Mutual Information Maximisation. *Expert Systems with Applications*, 42(22): 8520-8532.

Hoquea, N., Bhattacharyyaa, D.K. & Kalitab, J.K. (2014). MIFS-ND: A Mutual Information-based Feature Selection Method. *Expert systems with applications*, 41(14): 6371-6385.

Kira, K. & Rendell, L.A. (1992). *The feature selection problem: Traditional methods and a new algorithm*. In: Proceedings of Ninth National Conference on Artificial Intelligence: 129–134.

Kramer, O. (2013). *Dimensionality Reduction with Unsupervised Nearest Neighbors*. Intelligent Systems Reference Library: 33-52. Springer-Verlag Berlin Heidelberg.

Kramer, O. (2013). *Dimensionality Reduction with Unsupervised Nearest Neighbors*. Intelligent Systems Reference Library 51: 33-52. Springer-Verlag Berlin Heidelberg.

Kwak, N. & Choi, C.-H. (2002). Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, 13(1): 143–159.

Lewis, D.D. (1992). *Feature selection and feature extraction for text categorization*. In: Proceedings of speech and natural language workshop, Morgan Kaufmann: 212–217.

Peng, H., Long, F. & Ding, C. (2005). Feature selection based on mutual information criteria of max dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8): 1226–1238.

Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3): 379–423.

Vergara, J.R. & Estevez, P.A. (2014). A review of feature selection methods based on mutual information. *Neural Comput Appl*, 24: 175-186.

Yang, H.H. & Moody, J. (2000). Data visualization and feature selection: New algorithms for nongaussian data. *Adv Neur In*. 12: 687-693.