ارائهٔ یک چارچوب توزیع‌شده برای انتخاب ویژگی چندمتغیره

منا شریف‌نژاد*، محسن رحمانی و حسین غفاریان
گروه کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اراک

چکیده

در بسیاری از مسائل یادگیری ماشین، انتخابِ ویژگی‌های مرتبط و اجتناب از ویژگی‌های افزونه، برای بهبود کارایی انتخاب ویژگی ضروری است. در اکثر رویکردهای موجود، از الگوریتم‌های فیلتر چندمتغیره برای این منظور استفاده می‌شود که در آن‌ها تعامل با طبقه‌بند نادیده گرفته‌می‌شود. این مقاله با ارائه یک چارچوب، ترکیب روش‌های نهفته با روش‌های فیلتر چندمتغیره را پیشنهاد می‌دهد تا با درنظرگرفتن تعامل با طبقه‌بند در انتخاب ویژگی‌ها، این مشکل را برطرف کند. در چارچوب پیشنهادشده، ارتباط بین هر ویژگی و برچسب‌های رده به‌وسیلهٔ الگوریتم‌های نهفته محاسبه و افزونگی بین ویژگی‌ها از طریق الگوریتم‌های فیلتر چندمتغیره بررسی می‌شود. در فرایند انتخاب ویژگی پیشنهادشده، به‌جای استفاده از همه مجموعه‌داده‌ها، از توزیع افقی آن‌ها استفاده شده‌است. این خصوصیت برای مجموعه‌داده‌هایی که دارای نمونه‌های زیادی هستند و نیز در محیط‌هایی که داده‌ها متمرکز نیستند، باعث کاهش زمان اجرای فرایند انتخاب ویژگی شده‌است. کیفیت روش ما با استفاده از شش مجموعه‌داده ارزیابی‌شده است. نتایج ثابت می‌کنند که چارچوب پیشنهادشده، می‌تواند دقت طبقه‌بندی را در مقایسه با روش‌های صرفاً مبتنی بر فیلتر چندمتغیره بهبود دهد. همچنین سرعت اجرا می‌تواند در مقایسه با روش‌های متمرکز بهبود یابد.

کلمات کلیدی: انتخاب ویژگی فیلتر چندمتغیره، انتخاب ویژگی نهفته، طبقه‌بندی، توزیع‌شدگی

New Framework for Distributed Multivariate Feature Selection

Mona Sharifnezhad, Mohsen Rahmani, Hosein Ghaffarian
Department of Computer Engineering, Faculty of Engineering,
Arak University, Arak, Iran

Abstract

Feature selection is a process for discovering relevant features and putting the irrelevant and redundant features aside. In recent years, feature domain studies have increased in machine learning and pattern recognition significantly, in a way that several techniques have been developed to solve the problem of feature reduction. Feature selection is an essential activity in data preprocessing. It describes features that significantly reduce the effects of noise and irrelevant and redundant data and provides desirable prediction results. Feature selection algorithms are divided into three groups: Filter methods, are typically divided into two subsets: univariate evaluation and multivariate evaluation. While in univariate evaluation, features are evaluated based on attributing matching weights with their relevance degree to the classes, multivariate evaluation approaches can manage the feature relevance to the classes together with their redundancy. Wrapper methods, in these methods, the feature set is selected based on the classifier in a way that all possible subsets of features are considered and evaluating all states, the best subset that has more accurate classification is selected. Embedded methods, in these methods, the search for an optimal feature set is located inside the classifier structure. In many current feature selection approaches, redundancy and relevance between features and class vector are considered

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۹/۵/۲ • تاریخ پذیرش: ۱۴۰۱/۲/۲۱ • تاریخ انتشار: ۱۴۰۱/۱۲/۲۹ • نوع مطالعه: پژوهشی

simultaneously. However, these multivariate filter algorithms calculate the suitability of features by only the intrinsic characteristics of the data. In this paper, we propose a new framework to offset the multivariate feature selection problem regarding the interaction with classifiers in feature selection. Our proposed framework calculates the relevance of each feature to class labels by embedded algorithms. Then, redundancy among features is examined through multivariate filter algorithms. In addition, in the proposed framework, instead of using all records at once, we use their horizontal distribution. This approach reduces the runtime of the process in datasets with many instances and in environments without centralized data. The results of the evaluation show that the suggested framework can improve classification accuracy compared with the methods just based on multivariate filters. In addition, the runtime speed is improved compared with the centralized methods.

Key words: Multivariate filter feature selection, Embedded feature selection, Classification, Distribution

۱- مقدمه

انتخاب ویژگی، فرایند کشف ویژگی‌های مرتبط و کنار گذاشتن ویژگی‌های غیرمرتبط و افزونه است [1,2]. در سال‌های گذشته دامنه ویژگی‌ها، از ده‌ها به صدها یا هزاران متغیر در زمینه‌هایی چون یادگیری ماشین یا شناسایی الگو با افزایش چشم‌گیری مواجه شده‌اند؛ به‌طوری‌که چندین روش برای حل مشکل کاهش ویژگی‌ها توسعه داده‌شده‌اند. انتخاب ویژگی به‌عنوان یک فعالیت مهم در پیش‌پردازش داده‌ها، بر روی ویژگی‌های ورودی متمرکز شده‌است که با کاهش تأثیرات نوفه، داده‌های غیرمرتبط و افزونه، ویژگی‌ها را به‌طور مؤثر توصیف می‌کند و نتایج پیش‌بینی خوبی را فراهم می‌کند [3,4,5].

به‌طور معمول روش‌های انتخاب ویژگی به‌صورت متمرکز به کار گرفته‌شده‌اند [6]. اما در سال‌های اخیر، به‌دلیل به کار گرفتن داده‌های بزرگ و توزیع شدن آن‌ها در مکان‌های مختلف، استفاده از انتخاب ویژگی توزیع‌شده ضروری است [4,6,7,8,9]. لذا تعداد زیادی از روش‌های توزیع‌شده به‌جای رویکردهای متمرکز توسعه داده‌شده‌اند. روش‌های اصلی برای بخش‌بندی و توزیع داده‌ها می‌توانند به‌صورت افقی یا عمودی باشند. در بخش‌بندی افقی، مجموعه‌داده‌ها به چندین بسته تقسیم می‌شوند که دارای ویژگی‌های مشابه با مجموعه‌داده‌های اصلی هستند، که هر کدام شامل یک زیرمجموعه از نمونه‌های اصلی هستند. در بخش‌بندی عمودی، مجموعه‌داده‌های اصلی به چندین بسته تقسیم می‌شود که دارای تعداد مشابهی از نمونه‌های اصلی هستند که هر کدام دارای زیرمجموعه‌ای از مجموعه اصلی ویژگی‌ها هستند [8]. روش‌های انتخاب ویژگی به سه دسته تقسیم می‌شوند:

۱. روش فیلتر: به‌طورمعمول رویکردهای فیلتر به دو زیرمجموعه مختلف تقسیم می‌شوند: ارزیابی تک‌متغیره[1] و ارزیابی چندمتغیره[2]. در ارزیابی تک‌متغیره، ویژگی‌ها برپایه نسبت‌دادن وزن‌هایی مطابق با درجه ارتباط آن‌ها با رده‌ها، مورد ارزیابی قرار می‌گیرند. درحالی‌که رویکردهای مبتنی بر ارزیابی چندمتغیره، می‌توانند افزونگی ویژگی‌ها را همراه با ارتباط آن‌ها با رده‌ها، مدیریت کنند. ویژگی‌هایی که به‌احتمال دارای رتبه‌های مشابه هستند، بررسی شده و از این طریق موارد افزونه حذف می‌شوند. بررسی افزونگی در بین ویژگی‌ها به زمان محاسباتی بیشتری نیاز دارد و دقت طبقه‌بندی را بهبود می‌دهد [10,11].

۲. روش پوشاننده[3]: در این روش، مجموعه‌ویژگی‌ها برپایهٔ طبقه‌بند، انتخاب می‌شوند. به‌طوری که تمامی زیرمجموعه‌های ممکن از ویژگی‌ها در نظر گرفته شده و با ارزیابی همه حالت‌ها بهترین زیرمجموعه که دارای دقت طبقه‌بندی بالاتری است انتخاب می‌شود [6,12,13].

۳. روش نهفته[4]: از مزایای هر دو روش قبلی با استفاده از معیارهای ارزیابی مختلف آن‌ها در انتخاب ویژگی سود می‌برد. در این روش، جستجو برای یک مجموعه ویژگی بهینه، درون ساختار طبقه‌بند قرار دارد [7,9]. به‌طور کلی تعدادی از زیرمجموعه‌های خوب از ویژگی‌ها با استفاده از روش فیلتر انتخاب می‌شوند. سپس روش پوشاننده، روی این زیرمجموعه‌های خوب به کار برده می‌شود تا یکی از بهترین‌ها به‌دست آید [13].

درنظرگرفتن تنها یک معیار در انتخاب ویژگی، یک استراتژی مناسب برای پیداکردن زیرمجموعه‌های خوب از ویژگی‌ها نیست. در روش‌های فیلتر مبتنی بر ارزیابی تک‌متغیره، انتخاب ویژگی برپایهٔ بیشینه ارتباط ویژگی با

[1] Univariate Evaluation
[2] Multivariate Evaluation
[3] Wrapper
[4] Embedded

رده است. به‌طوری که از سرعت بالایی برخوردارند و مستقل از طبقه‌بند هستند. اما در این روش‌ها وابستگی بین ویژگی‌ها درنظر گرفته نشده که این امر در کاهش کارایی طبقه‌بندی تأثیرگذار است. با استفاده از روش‌های فیلتر مبتنی بر ارزیابی چندمتغیره، می‌توان با حذف ویژگی‌های افزونه در جهت افزایش کارایی طبقه‌بندی گام موثرتری برداشت؛ ولی به‌دلیل نیاز به انجام پردازش بیشتر، سرعت کمتری دارد. در روش‌های پوشاننده، می‌توان از طریق طبقه‌بند نسبت به روش‌های فیلتر نتایج بهتری تولید کرد؛ ولی اجرای این روش‌ها پرهزینه است و در بعضی از مواقع که تعداد ویژگی‌ها زیاد هستند، منجر به شکست می‌شوند. همچنین وابستگی بین ویژگی‌ها درنظر گرفته نمی‌شود. روش‌های نهفته نیز با افزونگی ویژگی‌ها مواجه هستند، به‌طوری‌که ارتباط بین ویژگی‌ها نادیده گرفته‌می‌شود.

ما یک چارچوب با ترکیبی از انتخاب ویژگی فیلتر چندمتغیره همراه با انتخاب ویژگی نهفته پیشنهاد می‌دهیم. این ترکیب کمک می‌کند که انتخاب ویژگی مبتنی بر چند معیار در جهت افزایش بهبود کارایی طبقه‌بندی انجام شود: انتخاب ویژگی مبتنی بر بیشینه ارتباط ویژگی با رده، کمینه افزونگی و بیشینه دقت طبقه‌بند است. چارچوب پیشنهادشده شامل ۵ مرحله است: در مرحله نخست، بعد از تعیین داده‌های آموزشی و تست، داده‌های آموزشی به‌صورت افقی توزیع می‌شوند. به‌طوری که تمام زیرمجموعه‌ها دارای تعداد ویژگی‌های یکسانی هستند. در مرحله دوم، روی هر زیرمجموعه از ویژگی‌ها، با استفاده از یک روش انتخاب ویژگی نهفته، به‌ویژگی‌ها امتیاز داده‌می‌شود، به‌طوری که از طریق دقت طبقه‌بندی ارتباط هر ویژگی با رده تعیین و ویژگی‌هایی که رتبه کمتر دارند، حذف می‌شوند. بدین ترتیب در این مرحله هر دو معیار بیشینه ارتباط با رده و بیشینه دقت طبقه‌بندی بررسی می‌شود. همچنین از طریق انتخاب ویژگی فیلتر چندمتغیره روی هر زیرمجموعه از ویژگی‌ها، ارتباط بین ویژگی‌ها ارزیابی‌شده، به‌طوری که به هر ارتباط امتیاز داده‌می‌شود و ویژگی‌هایی که امتیاز بیشتری کسب کرده‌اند، حذف می‌شوند؛ بنابراین در این مرحله معیار کمینه افزونگی نیز بررسی می‌شود؛ سپس ویژگی‌های انتخاب‌شده از طریق انتخاب ویژگی نهفته و فیلتر چندمتغیره با هم ترکیب و سایر ویژگی‌ها حذف می‌شوند. در مرحله سوم، به ویژگی‌هایی که حذف شدند، رأی داده‌می‌شود تا در پایان، آن ویژگی‌هایی که بیشینه رأی را دریافت کرده‌اند، از مجموعه‌ویژگی‌های نهایی حذف شوند. در مرحله چهارم، سایر ویژگی‌های باقی‌مانده با هم ادغام شده تا مجموعه نهایی از ویژگی‌های انتخاب‌شده، مشخص شود. در مرحله آخر با استفاده از داده‌های آموزشی نهایی و داده‌های تست، دقت طبقه‌بندی ارزیابی می‌شود.

ارزیابی نتایج نشان‌می‌دهد که چارچوب پیشنهادی در بهبود دقت طبقه‌بندی مؤثر بوده است. همچنین زمان اجرا به‌دلیل توزیع‌شدگی داده‌ها نسبت به روش‌های متمرکز کاهش یافته‌است.

ادامه مقاله به‌صورت زیر تدوین شده‌است. روش‌های انتخاب ویژگی در بخش ۲، چارچوب پیشنهادی برای انتخاب ویژگی در بخش ۳، نتایج تجربی و تشریح مطالب در بخش ۴ آورده شده‌است. سرانجام بخش ۵ شامل نتیجه‌گیری مقاله و کارهای پیشنهادی آینده است.

۲. روش‌های انتخاب ویژگی

در این بخش، ابتدا چند روش انتخاب ویژگی به‌صورت متمرکز بررسی، سپس انواع انتخاب ویژگی توزیع‌شده مرور و همچنین بعضی از روش‌های ادغام که بعد از اِعمال انتخاب ویژگی روی بخش‌های توزیع‌شده استفاده می‌شوند، بررسی شده‌است.

۲-۱- انتخاب ویژگی متمرکز

بسیاری از روش‌های انتخاب ویژگی بر روی مجموعه‌داده‌های متمرکز بررسی‌شده‌اند. در [15]، یک الگوریتم دو مرحله‌ای برای انتخاب ویژگی معرفی شده‌است، به‌طوری که ویژگی‌ها با استفاده از دو روش مختلف فیلتر به‌نام نسبت فیشر[5] و روش ارتباط بیشینه و کمینه افزونگی[6]، به‌طور مجزا رتبه‌بندی و ارزیابی می‌شوند؛ سپس ویژگی‌های انتخاب‌شده مشترک بین هر دو روش بالا، به‌عنوان بهترین ویژگی‌ها درنظر گرفته‌می‌شود. در مرحله دوم، با استفاده از روش حذف بازگشتی ویژگی‌ها به‌وسیلهٔ بردار پشتیبان[7] ویژگی‌هایی به‌غیر از آن‌هایی که در مرحله نخست به‌دست‌آمده، بررسی می‌شوند. درنهایت، از اجتماع ویژگی‌های به‌دست‌آمده در مرحله نخست و دوم، به‌عنوان ویژگی‌های

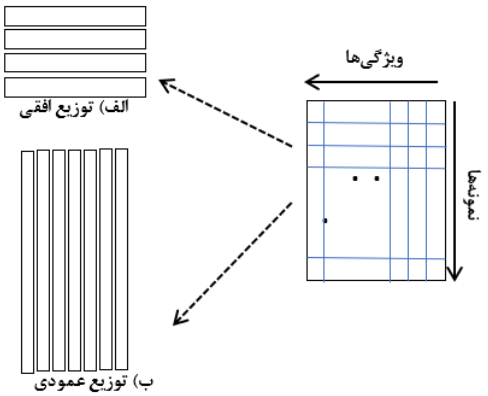
[5] Fisher Ratio
[6] maximum Relevance Minimum Redundancy (mRMR)
[7] Support Vector Machine Recursive Feature Elimination (SVM-RFE)

انتخاب‌شده، استفاده، سپس از طریق ماشین بردار پشتیبان، دقت و صحت آن‌ها بررسی می‌شود. در [16]، روشی به‌نام FWHM[8] در دو مرحله ارائه‌شده است: در مرحله نخست، مجموعه‌داده‌ها به‌وسیلهٔ شش روش متفاوت از فیلترهای تک‌متغیره امتیازدهی می‌شوند. علت استفاده از این تعداد فیلتر این است که روش‌های فیلتر هر کدام به‌تنهایی منطبق بر یک معیار هستند، برای کاهش وابستگی مربوط به هر معیار از آن‌ها استفاده شده‌است. از طریق آن‌ها می‌توان نتایج مختلفی از رتبه‌بندی یا امتیازدهی به‌دست‌آورد؛ درنهایت از میانگین آن‌ها برای رتبه‌بندی هر ویژگی استفاده می‌شود. در مرحلهٔ دوم، از استراتژی‌های جستجوی تصادفی مثل الگوریتم ژنتیک[9] به‌عنوان یک روش پوشاننده استفاده شده به‌طوری که میانگین رتبه‌بندی به‌دست‌آمده برای تولید جمعیت اولیه منظور شده‌است. ویژگی‌هایی که دارای رتبه‌ی بالاتری هستند، فرصت بیشتری برای انتخاب‌شدن دارند. در [17]، برای انتخاب ژن‌ها از روش SVM-RFE استفاده شده‌است. به‌طوری که دقت طبقه‌بندی در تشخیص ژن‌های سرطانی افزایش یافته‌است. در [18]، از سه روش جستجوی محلی LS[10]، SLS[11] و VNS[12] برای انتخاب ویژگی‌های مرتبط در زمینه Credit Scoring برای مسائل مالی و بانک‌داری استفاده شده، سپس نتایج هر کدام از روش‌ها به‌طور جداگانه با طبقه‌بند SVM ترکیب شده‌است. در [19]، با معرفی یک روش انتخاب ویژگی سلسله‌مراتبی در شبکه‌های عصبی تک‌لایه ویژگی‌های افزونه و نوفه‌دار هرس می‌شوند.

۲-۲- انتخاب ویژگی توزیع‌شده

ایده توزیع داده‌ها به دو صورت افقی و عمودی قابل انجام است. همان‌طور که در شکل (۱) نشان داده‌شده‌است، داده‌ها در توزیع افقی برپایهٔ نمونه‌ها و در توزیع عمودی، براساس ویژگی‌ها تقسیم می‌شوند. به‌طورمعمول اگر تعداد نمونه‌ها بیشتر باشند، از توزیع افقی استفاده می‌شود و اگر تعداد ویژگی‌ها بیشتر باشند، توزیع عمودی به کار برده می‌شود [4,6,7,8,9].

با تعیین نوع توزیع، مجموعه‌داده‌ها به زیرمجموعه‌هایی کوچک‌تر تقسیم می‌شوند. سپس الگوریتم انتخاب ویژگی مورد نظر روی هر زیرمجموعه به‌طور مجزا می‌تواند اعمال شود. در مرحله بعد، برای ترکیب نتایج، استفاده از یک روش ادغام ضروری است.



هاهای توزیع داده(شکل ۱). تکنیک
(Figure-1): Techniques for Data Distribution

به‌طوری که باید ویژگی‌های انتخاب‌شده در هر زیرمجموعه ادغام شوند و یک مجموعه‌ی واحد را تشکیل دهند.

روش‌های مختلفی برای ادغام نتایج از طریق تعیین یک حدآستانه، ارائه‌شده است. در [7,14]، برای محاسبه حدآستانه از خطای طبقه‌بندی و درصد ویژگی‌های باقیمانده استفاده شده‌است. به‌طوری که هر دو مقدار باید تا حد امکان به حداقل برسند. به‌ازای هر ویژگی که ممکن است در هر زیرمجموعه برچسب حذف را دریافت کرده باشد، مقدار آن‌ها تعیین می‌شود. کمینه مقدار به‌عنوان حدآستانه درنظر گرفته‌می‌شود. ویژگی‌هایی که تعداد برچسب‌های حذف‌شان بیشتر از مقدار حدآستانه باشند، حذف می‌شوند.

در [3,4]، دقت طبقه‌بندی نخستین زیرمجموعه ویژگی انتخاب‌شده را محاسبه کرده و آن را به‌عنوان مبنا[13] در نظر گرفته‌اند؛ سپس دقت طبقه‌بندی سایر زیرمجموعه ویژگی‌های باقیمانده به‌طور مجزا با نخستین زیرمجموعه محاسبه شده‌است. درصورتی‌که مقدار مبنا را بهبود دهند، آن‌ها نیز بخشی از انتخاب نهایی خواهند شد. در این روش، نخستین زیرمجموعهٔ ویژگی انتخاب‌شده همیشه به‌عنوان بخشی از انتخاب نهایی خواهد بود.

در [4,8,9]، با استفاده از روش شاخص پیچیدگی[14]، حدآستانه تعیین شده‌است. ویژگی‌هایی که به‌عنوان نامزد خوب در نظر گرفته می‌شوند، باعث کاهش

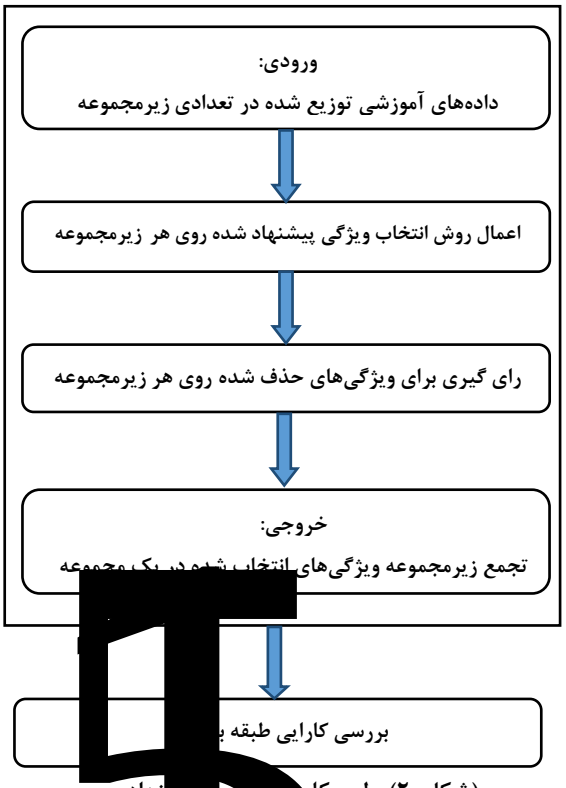
[8] Filter-Wrapper Hybrid Method
[9] Genetic Algorithm
[10] Local Search Method
[11] Stochastic Local Search Method
[12] Variable Neighborhood Search Method
[13] baseline
[14] Complexity Measure

پیچیدگی شده و باید حفظ شوند. درحالی‌که ویژگی‌هایی که نامزد بدی محسوب می‌شوند، سبب افزایش پیچیدگی می‌شوند و باید از بین بروند. این روش مستقل از طبقه‌بند[15] است.

در نهایت بعد از اتمام فرایند ادغام زیرمجموعه‌ها توسط یکی از روش‌های بالا، دقت مدل ساخته‌شده با استفاده از طبقه‌بند بر روی مجموعه‌داده‌ها محاسبه می‌شود.

۳- چارچوب پیشنهادی

در بسیاری از روش‌های انتخاب ویژگی ارائه‌شده، بررسی ویژگی‌های مرتبط و حذف ویژگی‌های افزونه مبتنی بر الگوریتم‌های فیلتر بوده است. به‌طوری که برای انتخاب ویژگی‌ها، مستقل از طبقه‌بند عمل می‌کنند که این امر می‌تواند باعث کاهش کارایی در انتخاب ویژگی‌ها شود. همچنین در بعضی از رویکردها، به‌دلیل استفاده از روش‌های متمرکز در انتخاب ویژگی‌ها، برای مجموعه‌داده‌هایی که دارای نمونه‌های زیادی هستند، زمان‌بر بوده و مناسب نیستند.



(شکل-۲): طرح کلی چارچوب پیشنهادی
(Figure-2): Schematic of Suggested Method

در این بخش ما یک چارچوب انتخاب ویژگی چندمتغیره به‌صورت توزیع‌شده مبتنی بر رتبه‌بندی[16] ویژگی‌ها از طریق معیارهای مختلف، پیشنهاد داده‌ایم. این چارچوب پیشنهادی دارای ۵ مرحله به شرح زیر باطشد:

مرحله نخست شامل توزیع افقی مجموعه‌داده‌های آموزشی در چندین زیرمجموعه است. مرحله دوم، در دو بخش پیشنهاد شده‌است. ابتدا با استفاده از انتخاب ویژگی نهفته روی هر زیرمجموعه، ویژگی‌ها با بیشینه ارتباط بررسی، سپس از طریق انتخاب ویژگی فیلتر چندمتغیره[17] روی هر زیرمجموعه، ویژگی‌هایی با کمینه افزونگی محاسبه می‌شوند. در این مرحله ویژگی‌ها با بیشینه ارتباط و کمینه افزونگی ترکیب شده و سایر ویژگی‌ها حذف می‌شوند. در مرحله سوم به ویژگی‌های حذف‌شده در هر زیرمجموعه، رأی داده‌می‌شود، سپس زیرمجموعه ویژگی‌های انتخاب‌شده در یک زیرمجموعه نهایی ترکیب می‌شوند؛ درنهایت کارایی طبقه‌بند[18]، روی زیرمجموعه نهایی از ویژگی‌های انتخاب‌شده و مجموعه‌داده‌های تست بررسی می‌شود.

طرح کلی چارچوب پیشنهادی در (شکل۲) و جزئیات آن در (شکل۳)، نمایش داده‌شده‌است. پارامترهای استفاده‌شده در جدول ۱، شرح داده‌شده‌است.

(جدول-۱): پارامترهای روش پیشنهادی
(Table-1): Parameters of Suggested Method

پارامتر	کاربرد
D	مجموعه‌داده‌های آموزشی
D_i	زیرمجموعه‌ای از D
F	مجموعه‌ی غیرتهی و محدود از ویژگی‌ها
C	مجموعه‌ی غیرتهی و محدود از رده‌ها
X_F	روش انتخاب ویژگی نهفته
W_{fi}	وزن محاسبه شده برای وابستگی ویژگی f_i به مجموعه رده C از طریق روش انتخاب ویژگی نهفته
W_{Fx}	بردار وزن ویژگیها در روش انتخاب ویژگی نهفته
$W_{(fi, fj)}$	وزن محاسبه شده برای ارتباط بین ویژگی f_i با ویژگی f_j از طریق روش انتخاب ویژگی فیلتر چندمتغیره
W_{FY}	بردار وزن برای ارتباط بین ویژگیها در روش انتخاب ویژگی فیلتر چندمتغیره
Δ	تابع ترکیب W_{Fx} و W_{FY}
Λ	ویژگی‌های حذف‌شده
FFS	ویژگی‌های انتخاب‌شده نهایی

برپایهٔ شکل(۳)، در ابتدا، مجموعه‌داده‌های آموزشی (D) به‌صورت افقی به چند زیرمجموعه مجزا (D_i) و بدون جای‌گذاری تقسیم می‌شوند. D_i به‌شکل زیر تعریف می‌شوند:

[15] Classifier
[16] Ranking
[17] Multivariate Filter
[18] Performance of Classifier

$D= \{(S, F) \mid S \in samples, F \in input\ features\}$

$D_i= \{(s_j, F) \mid s_j \in S \quad j \in n\}$

$S = \{ s_1, s_2, ..., s_n\}$, $F = \{f_1, f_2, ..., f_m\}$

که در این تعاریف S و F به‌ترتیب مجموعه‌های غیرتهی و محدود از نمونه‌ها و ویژگی‌ها هستند. با محاسبه وابستگی ویژگی f_i به مجموعه رده C با استفاده از روش‌های انتخاب ویژگی نهفته، می‌توان از طریق فاکتور دقت طبقه‌بند آن را به‌صورت زیر درنظر گرفت:

X_F = Embedded method with tuning parameter W_{fi} for F

$W_{Fx} = \{w_{fi}\}$ for $f_i \forall F$

از طریق بردار وزن (W_F)، وزن مجموعه ویژگی‌ها برپایهٔ دقت طبقه‌بند مطابق با رویکرد نهفته X_F محاسبه می‌شود. سپس ویژگی‌ها با استفاده از W_F، رتبه‌بندی می‌شوند.

Y_F = Multivariate Filter method with tuning parameter $W_{(fi, fj)}$ for F

$W_{FY} = \{w_{(fi, fj)}\}$ for $f_i, f_j \in F, i \neq j$

$\lambda = \arg\min \{\delta (W_{Fx}, W_{FY})\}$

$FFS (D_i) = F- \{\lambda\}$

(شکل- ۳): جزئیات روش پیشنهادی

(Figure-3): Details of Suggested Me

در مرحله بعد، برای بررسی ویژگی‌های افزونه، از یکی از روش‌های انتخاب ویژگی فیلتر چندمتغیره بر روی هر زیرمجموعه استفاده می‌شود. به‌طوری که به ارتباط هر ویژگی با سایر ویژگی‌ها امتیاز داده‌می‌شود (W_{FY}). امتیازدهی برای ارتباط بین ویژگی‌ها می‌تواند به‌وسیلهٔ هر نوع انتخاب ویژگی مبتنی بر فیلتر انجام شود. سپس دو رابطه ارتباط و افزونگی با هم ترکیب می‌شوند (δ).

δ، تابعی است که برپایهٔ بیشینه‌کردن ارتباط بین ویژگی‌ها با رده و کمینه‌کردن افزونگی بین ویژگی‌ها عمل می‌کند. معیارهای ترکیب رابطه ارتباط و افزونگی می‌تواند به روش‌های مختلفی باشد. در [15]، از معیار تفاضل اطلاعات متقابل[19] بین بیشینه ارتباط و کمینه افزونگی استفاده شده‌است. در [11,16]، علاوه‌بر معیارMID، از نسبت اطلاعات متقابل[20] نیز استفاده شده‌است؛ سپس ویژگی‌هایی که کمترین رتبه را دارند، برپایهٔ یک حدآستانه معین، از مجموعه‌ویژگی‌های نهایی حذف می‌شوند (λ).

در مرحله بعد به ویژگی‌هایی که در مرحله قبل برای حذف‌شدن انتخاب‌شده‌اند، رأی داده‌می‌شود. این رأی‌گیری کمک می‌کند تا ویژگی‌هایی که بیشتر از یک حدآستانه معین رأی آوردند، حذف شوند؛ سپس ویژگی‌های انتخاب‌شده در هر زیرمجموعه، با استفاده از یک تابع ادغام در یک زیرمجموعه نهایی جمع می‌شوند. راه‌کار ادغام به‌کاررفته در این مقاله، مشابه [4,8,9] مبتنی بر استفاده از روش شاخص پیچیدگی[21] است. در پایان با استفاده از مجموعه‌داده‌های نهایی آموزش و مجموعه‌داده‌های تست، کارایی و دقت طبقه‌بندی بررسی می‌شود. مشخصات جزئیات بیشتر در الگوریتم ۱ نشان داده شده‌است.

Algorithm 1 Pseudo-code for Proposed Approach

Input:
- Load Dataset D (m×s) with m Samples and s Features
  a. Train Dataset = $D(m_1 \times s)$
  b. Test Dataset = $D(m_2 \times s)$
  c. Vector of votes for Removing Features = V, length of Number of Features, Initialize the V to 0

Output:
- FS: Subset of Final Selected Features

Compute:
1. Horizontal Partition on Train Data
   D_1 to D_n: Subset of Train Datasets
2. Proposed Framework of Feature selection on each Subset of Train Data
   –For each D_i Subset of Train Data
   a. Maximum Relevancy by Embedded Feature Selection (W_{Fx})
   b. Minimum Redundancy by Multivariate Filter Feature Selection (W_{FY})
   c. δ (W_{Fx}, W_{FY}): Combine of Relevancy and Redundancy
   d. Remove Threshold Percent of Features with Lower Score (λ)
   e. Increment one Vote in Vector V for each Remove Feature
3. Aggregated Subsets of Selected Features
   a. Th: A threshold of votes to remove a Feature
   b. FFS: Subset of Final Selected Features
4. Obtaining Accuracy by Classifier

۴- نتایج تجربی

در این بخش، ابتدا درباره تنظیمات اجرا توضیح می‌دهیم؛ سپس نتایج چارچوب پیشنهادی را از مشخصات این مجموعه‌دادهها در (جدول۲) آورده شده‌است که شامل تعداد نمونه‌ها، تعداد ویژگی‌ها و تعداد رده‌ها است. نمونه‌ها در مجموعه‌داده‌ها به دو مجموعه آموزش و آزمون تقسیم می‌شوند. به‌طورمعمول این تقسیم‌بندی به‌صورت $\frac{2}{3}$ مجموعه‌داده‌ها برای آموزش و $\frac{1}{3}$ برای آزمون درنظر گرفته‌می‌شود.

همه نتایج روی یک PC با مشخصات Intel (R) Core (TM) i7-2670QM CPU@ 2.2GHz 2.2.GHz و 8 GB RAM، با استفاده از نرم‌افزار متلب (نسخه ۲۰۱۳) در محیط ویندوز ۷ اجرا شده‌است.

۴-۱- تنظیمات اجرا

در این بخش برای ارزیابی چارچوب پیشنهادی از شش مجموعه‌داده‌ها [20] استفاده کردیم. چندمتغیره استفاده‌کرد. در این‌جا ما از یکی از روش‌های معروف در انتخاب ویژگی نهفته به نام SVM-RFE و نیز برای انتخاب ویژگی فیلتر چندمتغیره از روش mRMR استفاده کردیم. همچنین چارچوب پیشنهادی را با چند روش معروف از انتخاب ویژگی فیلتر مانند بهره اطلاعاتی[22] و ReliefF و نیز روش‌های غیرترکیبی انتخاب ویژگی SVM-RFE و mRMR مقایسه کردیم.

[19] Mutual Information Difference criterion (MID)
[20] Mutual Information Quotient criterion (MIQ)
[21] Complexity Measure
[22] Information Gain (IG)

- **حذف بازگشتی ویژگی‌ها به‌وسیلهٔ بردار پشتیبان**

انتخاب ویژگی SVM-RFE توسط گایون[۲۳] و همکاران به‌عنوان روشی برای انتخاب ویژگی معرفی شده‌است [15]. مطابق با این روش ویژگی‌ها به‌صورت بازگشتی حذف می‌شوند، به‌طوری که در هر مرحله به‌وسیله بردار پشتیبان وزن ویژگی‌ها محاسبه و به هر یک از آن‌ها یک امتیاز داده‌می‌شود و ویژگی که دارای کمترین امتیاز باشد، حذف خواهد شد [21]. این فرآیند تکرار خواهد شد تا زمانی که تمام ویژگی‌ها حذف شوند؛ درنهایت فهرستی از رتبه‌بندی ویژگی‌ها وجود خواهد داشت که انتخاب ویژگی می‌تواند با انتخاب یک گروه از ویژگی‌های برتر به‌دست آید.

- **انتخاب ویژگی ارتباط بیشینه و کمینه افزونگی**

انتخاب ویژگی mRMR یک الگوریتم انتخاب ویژگی چندمتغیره مبتنی بر فیلتر است که در میان ویژگی‌های انتخاب‌شده به‌طور هم‌زمان ارتباط را بیشینه و افزونگی را کمینه می‌کند. این روش از اطلاعات متقابل[۲۴] برای تجزیه و تحلیل مرتبط‌بودن و افزونگی استفاده می‌کند؛ به‌طوری که بیشینه ارتباط از طریق محاسبه بیشینه مقدار میانگین از تمام اطلاعات متقابل بین تک‌تک ویژگی‌های موجود و بردار رده به‌دست می‌آید. همچنین کمینه افزونگی از طریق محاسبه کمینه مقدار میانگین از تمام اطلاعات متقابل بین بردار ویژگی‌ها بررسی می‌شود؛ سپس روابط بالا را می‌توان با یکی از دو روش تفاضل اطلاعات متقابل یا روش نسبت اطلاعات متقابل پیاده‌سازی کرد [10,11,15,22].

- **انتخاب ویژگی بهره اطلاعاتی**

IG یکی از رایج‌ترین روش‌های تک‌متغیره مبتنی بر فیلتر برای ارزیابی ویژگی‌ها است. این روش براساس بهره اطلاعاتی، به ارتباط هر ویژگی با رده رتبه و امتیاز می‌دهد؛ سپس با استفاده از یک حدآستانه، تعداد معینی از ویژگی‌ها که رتبه بالاتری را به‌دست آوردند را انتخاب می‌کند [23].

- **انتخاب ویژگی ReliefF**

این روش، توسعه‌یافته الگوریتم Relief است که قابلیت کارکردن با مجموعه‌داده‌های چندرده، نوفه‌دار و ناقص (غیرکامل) را فراهم می‌کند [24]. به‌جای یافتن n نمونه نزدیک‌ترین برخورد و نزدیک‌ترین شکست،eliefF از هر رده n نمونه را انتخاب می‌کند و سهم هر غیر هم‌ردهی در وزن‌دهی برپایهٔ احتمالات پیشین آن‌ها محاسبه می‌شود. نوفه در یک رده و یا مقدار یک ویژگی، تأثیر قابل‌توجه‌ای در انتخاب نزدیک‌ترین برخورد و نزدیک‌ترین شکست دارد. برای این‌که انتخاب نزدیک‌ترین برخورد و نزدیک‌ترین شکست با دقت و اطمینان بالاتری صورت گیرد الگوریتم ReliefF، از n نزدیک‌ترین برخورد و n نزدیک‌ترین شکست استفاده می‌کند و در تخمین کیفیت یک ویژگی میانگین سهم هرکدام را در نظر می‌گیرد. زمانی که نمونه‌ها در بعضی از ویژگی‌ها فاقد مقدار باشند، الگوریتم ReliefF تابع مربوطه را تغییر می‌دهد و آن را مدیریت می‌کند [25].

۴-۲- نتایج صحت طبقه‌بندی

در این بخش، دقت طبقه‌بندی به‌دست‌آمده به‌وسیلهٔ KNN[25]، NB[26] و SVM[27] که برای ارزیابی بسیاری از الگوریتم‌های انتخاب ویژگی استفاده می‌شوند [3,4,6,7-9,15,17] را بر روی چارچوب پیشنهادی و چهار رویکرد مختلف انتخاب ویژگی IG، ReliefF، SVM-RFE و mRMR مقایسه کردیم. جدول‌های ۳،۴و۵ نتایج تجربی حاصل از دقت طبقه‌بندی را به دو شکل: توزیع‌شده (مشخص‌شده با پیشوندD در ابتدای نام روش‌ها در جداول) و متمرکز (مشخص‌شده با پیشوند C در ابتدای نام روش‌ها در جداول) بر روی مجموعه‌داده‌های یادشده، نشان می‌دهند. مواردی که به‌عنوان بهترین نتایج در هر مجموعه‌داده به‌دست‌آمده، به‌صورت پررنگ نشان داده‌شده‌اند. نتایج به‌دست‌آمده متغیر هستند و به مجموعه‌داده‌ها و طبقه‌بند وابسته هستند. در ستون آخر از هر جدول، نتایج حاصل از دقت طبقه‌بندی بدون استفاده از الگوریتم انتخاب ویژگی گزارش شده‌است. در سطر آخر، میانگین دقت طبقه‌بند مورد نظر محاسبه شده‌است. برای مجموعه‌داده‌های Ozon، Madelon، Isolet و Connect4 با طبقه‌بند KNN بهترین نتیجه برای چارچوب پیشنهادی به‌صورت توزیع گزارش شده‌است. مجموعه‌داده Ozon، بهترین مقدار صحت را

[23] Guyon et. al
[24] Mutual Information

[25] k-Nearest Neighbor

[26] Naive Bayes

[27] Support Vector Machine

برای حالت متمرکز همراه با انتخاب ویژگی ReliefF داشته است. مجموعه‌داده Madelon، هم در حالت متمرکز و هم در حالت توزیع‌شده در چارچوب پیشنهادی با طبقه‌بند KNN بهترین نتیجه را به خود اختصاص داده است. در طبقه‌بند SVM، برای تمام مجموعه‌داده‌ها به‌جز Spect و Ozon، بهترین نتیجه مربوط به چارچوب پیشنهادی است. مجموعه‌داده Ozon، بهترین مقدار صحت را برای حالت متمرکز در تمام انتخاب ویژگی‌ها به‌غیر از ReliefF داشته است. مجموعه‌داده Madelon، هم در حالت توزیع‌شده و هم در حالت متمرکز در تمام انتخاب ویژگی‌ها دارای نتایج مشابهی است. اما در طبقه‌بند NB بهترین میانگین صحت مربوط به نسخه توزیع‌شده الگوریتم IG است.

چارچوب پیشنهادی در حالت متمرکز، بهترین نتیجه را در میانگین صحت دارد. همچنین برای مجموعه‌داده‌های Ozon، Isolet و Connect4 با طبقه‌بند NB بهترین نتیجه مربوط به چهارچوب پیشنهادی به‌صورت توزیع‌شده است. البته در مجموعه‌داده Connect4 در تمام انتخاب ویژگی‌ها در حالت توزیع‌شده دارای نتایج مشابهی است. همچنین بهترین نتیجه در حالت توزیع‌شده در این طبقه بند مربوط به مجموعه‌داده‌های Ozonو Spambase است که دارای مقادیر مشابهی هستند. از طرفی با توجه به نتایج حاصل از ستون آخر در هر جدول، تنها در کمتر از $\frac{1}{3}$ موارد آزمایش‌شده، دقت طبقه‌بندی بدون استفاده از الگوریتم انتخاب ویژگی، بهتر شده‌است. در بین مجموعه‌داده‌های آزمایش‌شده، مجموعه‌دادهٔ Spambase بیشترین تعداد بهبود را در هر سه طبقه‌بند بدون استفاده از الگوریتم انتخاب ویژگی، نسبت به سایرین داشته است. بیشترین تعداد بهبود دقت طبقه‌بندی با استفاده از الگوریتم انتخاب ویژگی در مقایسه با بدون استفاده از آن، به‌ترتیب مربوط به مجموعه‌داده‌های Madelon، Isolet و Connect4 است. برای رویکرد پیشنهادی نسبت به سایر روش‌ها در کمتر از $\frac{1}{20}$ موارد آزمایش‌شده، دقت رده‌بندی آن در مقایسه با حالت بدون استفاده از الگوریتم انتخاب ویژگی، کاهش یافته‌است.SVM در مقایسه با سایر رده‌بندهای مورد بررسی، دارای تعداد نتایج بهبودیافته بیشتری در استفاده از الگوریتم انتخاب ویژگی نسبت به بدون استفاده از آن، بوده است. نمودار میانگین دقت طبقه‌بندی برای هر سه طبقه‌بند KNN، SVM و NB ، به‌ترتیب در شکل‌های ۴، ۵ و ۶ نشان داده‌شده‌است. میانگین دقت رده‌بندی برای رده‌بندهای KNN و NB در حالت توزیع‌شده نسبت به متمرکز در تمام روش‌های انتخاب ویژگی آزمایش‌شده، بهبود یافته‌است. میانگین دقت طبقه‌بندی برای چارچوب پیشنهادی نسبت به سایر روش‌های انتخاب ویژگی در حالت توزیع‌شده با KNN و SVM بیشتر گزارش شده‌است. اما میانگین دقت رده‌بندیNB برای چارچوب پیشنهادی نسبت به سایر روش‌های انتخاب ویژگی توزیع‌شده کاهش یافته‌است. به‌طوری که در سایر روش‌های انتخاب ویژگی، حالت توزیع‌شده نسبت به متمرکز بهبود یافته‌است.

(جدول-۲): مشخصات مجموعه‌داده‌ها

(Table-2): Characteristics of used Dataset

Dataset	Features	Samples		Classes
		Training	Test	
Ozone	72	1691	845	2
Madelon	500	1600	800	2
Spambase	57	3067	1534	2
Connect4	42	45038	22519	3
Spect	22	178	89	2
Isolet	617	5198	2599	26

(جدول-۳): نتایج تست دقت طبقه‌بندی با KNN
(Table-3): Results of KNN Classifier Accuracy

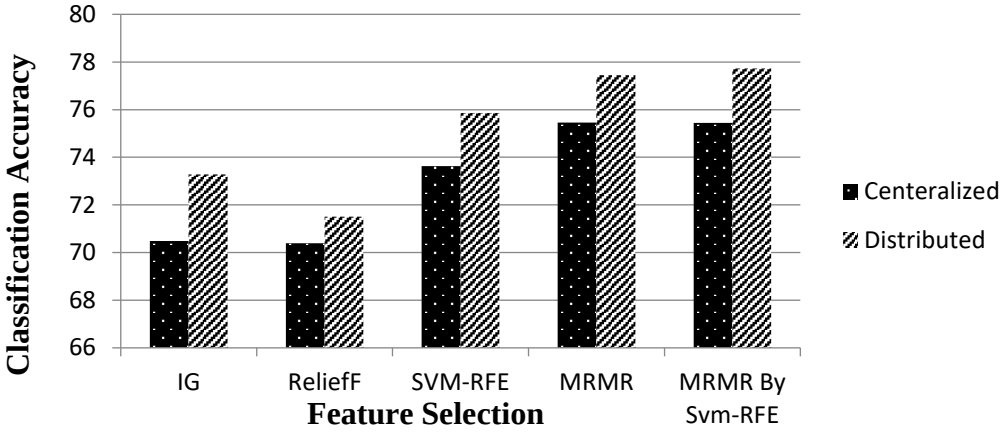
Dataset	C-IG D-IG	C-ReliefF D-ReliefF	C-SVM_RFE D-SVM_RFE	C-mRMR D-mRMR	SVM-RFE_By _C-mRMR SVM-RFE_By_D-mRMR	With out Feature Selection
Ozone	98.70 98.82	98.22 93.79	98.70 90.18	98.70 98.70	98.70 98.58	98.12
Madelon	49.62 49.62	49.62 49.62	49.62 49.62	49.62 49.62	49.62 49.62	48.62
Spambase	95.56 94.98	77.69 88.00	94.65 91.86	92.37 96.41	90.54 99.61	91.65
Spect	68.54 64.04	66.29 64.04	64.04 70.79	65.17 65.17	64.04 62.92	62.92
Isolet	92.90 92.99	95.17 95.19	94.25 95.30	91.02 93.35	92.68 95.60	90.42
Connect4	66.42 66.42	60.42 60.42	71.22 74.72	72.26 73.12	73.12 75.42	65.58
Aerage	78.62 77.81	74.57 75.18	78.75 78.74	78.19 79.39	78.12 80.29	----------

(جدول-۴): نتایج تست دقت طبقه‌بندی با SVM
(Table-4): Results of SVM Classifier Accuracy

Dataset	C-IG D-IG	C-ReliefF D-ReliefF	C-SVM_RFE D-SVM_RFE	C-MRMR D-MRMR	SVM-RFE_By_C-MRMR SVM-RFE_ By_D-MRMR	With out Feature Selection
Ozone	98.75 98.70	98.70 98.58	98.58 98.24	98.58 98.58	98.58 98.80	98.35
Madelon	49.37 49.13	48.25 50.87	49.50 50.75	50.15 52.13	51.26 52.75	48.25
Spambase	77.62 78.80	75.60 79.71	72.73 75.53	74.49 76.02	73.84 77.69	77.10
Spect	73.03 73.03	71.91 76.40	65.17 71.91	74.15 76.50	69.66 72.79	73.03
Isolet	73.64 81.14	90.62 90.72	88.26 89.14	90.06 91.35	91.45 92.82	85.23
Connect4	50.53 58.92	37.26 32.74	67.54 69.51	66.35 70.12	67.87 71.22	62.06
Average	70.49 73.28	70.39 71.50	73.63 75.86	75.63 77.45	75.44 77.84	----------

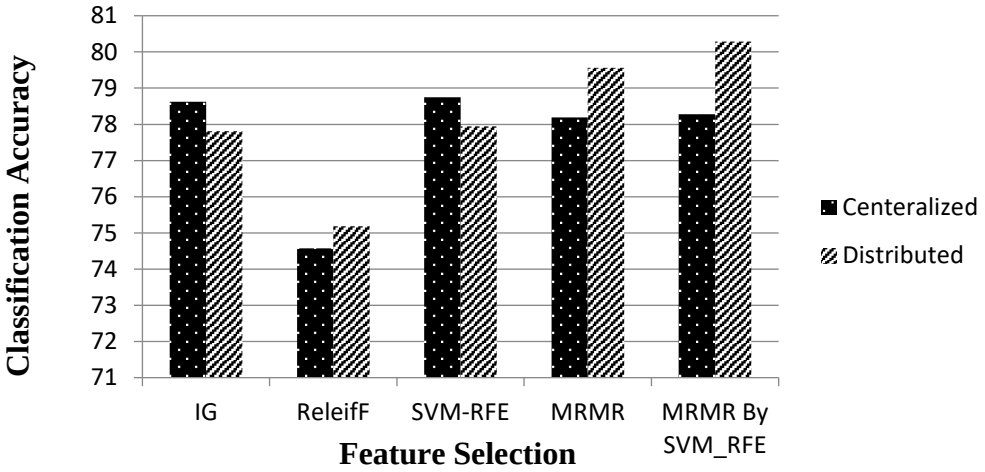
(جدول-۵): نتایج تست دقت طبقه‌بندی با NB
(Table-5): Results of NB Classifier Accuracy

Dataset	C-IG D-IG	C-ReliefF D-ReliefF	C-SVM_RFE D-SVM_RFE	C-mRMR D-mRMR	SVM-RFE_By_C-mRMR SVM-RFE_By_D-mRMR	With out Feature Selection
Ozone	70.41 73.61	71.84 71.71	71.24 79.53	73.67 73.50	72.43 72.20	71.95
Madelon	47.25 47.87	46.87 50.00	48.85 49.5	46.50 47.0	47.42 48.88	46.87
Spambase	76.53 89.70	92.85 91.25	80.12 89.68	78.88 93.22	90.35 93.46	93.20
Spect	73.03 74.16	76.50 75.28	71.91 74.16	74.15 71.91	70.79 74.16	74.16
Isolet	79.98 78.68	73.10 78.18	80.46 82.42	81.25 82.42	82.24 86.45	80.34
Connect4	60.58 60.42	60.30 60.42	60.42 60.42	60.10 60.42	60.42 60.42	58.62
Average	67.96 70.74	70.24 71.14	68.84 72.62	69.09 71.41	70.61 72.59	----------



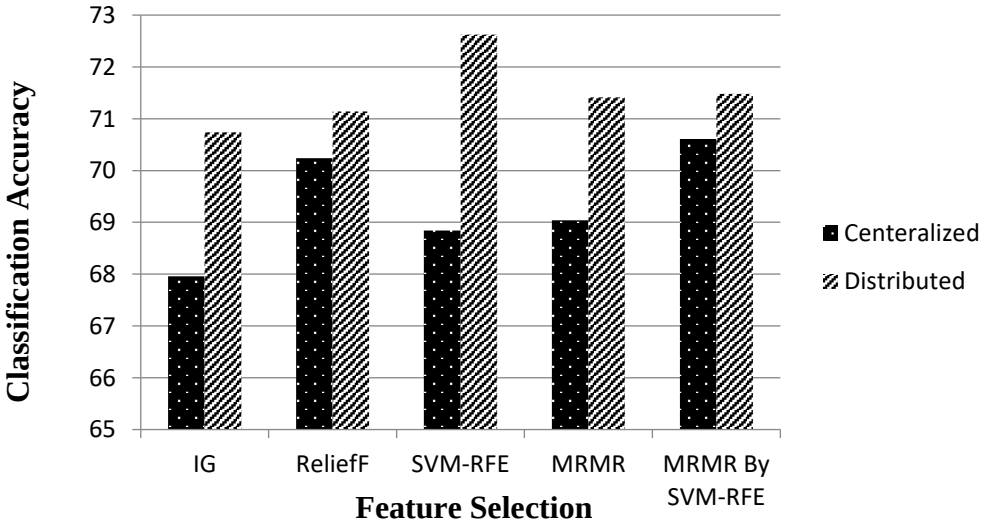
(شکل-۴): میانگین دقت طبقه‌بندی KNN روی رویکردهای مختلف انتخاب ویژگی

(Figure-4): Mean of KNN classifier accuracy on different feature selection methods



(شکل-۵): میانگین دقت طبقه‌بندی SVM روی رویکردهای مختلف انتخاب ویژگی

(Figure-5): Mean of SVM classifier accuracy on different feature selection methods



(شکل-۶) میانگین دقت طبقه‌بندی NB روی رویکردهای مختلف انتخاب ویژگی

(Figure-6): Mean of NB classifier accuracy on different feature selection methods

(جدول- ۶): زمان اجرای الگوریتم‌های انتخاب ویژگی تست‌شده بر حسب ثانیه
(Table-6): Runtime (seconds) for all tested feature selection algorithms

		Ozone	Madelon	Spect	Spambase	Connect4	Isolet	Average
mRMR	C	40.450	3631.541	0.646	18.917	1565.512	5700.237	1826.217
	D	31.561	98.21	0.288	13.052	62.820	164.64	61.762
SVM-RFE	C	4.728	6.960	0.680	8.962	12.250	9.146	7.121
	D	0.245	0.712	0.224	0.645	0.828	0.947	0.600
IG	C	2.116	22.008	0.316	3.077	7.847	30.742	11.018
	D	1.429	16.037	0.096	2.737	2.183	20.05	7.089
RELEIFF	C	10.284	61.802	0.711	15.920	1502.104	603.497	365.720
	D	0.805	15.310	0.024	2.712	8.195	123.17	25.036
_By_mRMR	C	88.703	3925.284	1.015	26.374	1743.731	6028.157	1968.877
SVM-RFE	D	41.863	155.683	0.771	7.348	37.387	256.091	83.190

(جدول-۷): تعداد ویژگی‌های انتخاب‌شده توسط رویکردهای مختلف انتخاب ویژگی
(Table-7): Number of selected features by tested feature selection algorithm

		Ozone	Madelon	Spect	Spambase	Connect4	Isolet
Full Set		72	500	22	57	42	617
mRMR	C	55	185	21	42	30	226
	D	31	52	16	29	32	124
SVM-RFE	C	41	175	16	42	30	226
	D	16	40	7	30	33	102
IG	C	41	175	16	42	30	226
	D	37	155	10	37	30	168
ReliefF	C	40	174	16	42	30	226
	D	17	68	10	21	34	175
SVM-RFE_By_mRMR	C	59	175	16	42	30	226
	D	37	50	18	34	32	131

۴-۳-- نتایج زمان اجرا

زمان اجرای الگوریتم‌های انتخاب ویژگی به‌صورت توزیع‌شده (سطرهای با برچسب Dدر جدول) و متمرکز (سطرهای با برچسب C در جدول) در (جدول ۶)، بر روی شش مجموعه‌داده نشان داده‌شده‌است. در رویکرد توزیع‌شده با توجه به این‌که همه زیرمجموعه‌ها می‌توانند در یک زمان پردازش شوند؛ لذا زمان نمایش داده‌شده در این جدول، بیشینه زمان مورد نیاز برای الگوریتم انتخاب ویژگی روی هر زیرمجموعه ایجادشده در مرحله تقسیم‌بندی است. در این آزمایش، همه زیرمجموعه‌ها در یک ماشین پردازش‌شده‌اند، اما الگوریتم پیشنهاد شده می‌تواند بر روی چندین پردازنده اجرا شود [3,6,8,9].

در مقایسه با روش متمرکز، زمان اجرای مورد نیاز در روش‌های توزیع‌شده روی تمام مجموعه‌داده‌ها کاهش یافته‌است. کمترین زمان اجرا در بین مجموعه‌داده‌های آزمایش‌شده به‌صورت توزیع‌شده، متعلق به مجموعه‌دادهٔ Spect با انتخاب ویژگی ReliefF است. این مجموعه‌داده نسبت به سایرین دارای کمترین تعداد ویژگی و نمونه است. در بین سایر مجموعه‌داده‌ها، کمترین زمان اجرای توزیع‌شده مربوط به انتخاب ویژگی SVM-RFE است. ستون آخر از جدول ۶. میانگین زمان اجرا را در هر دو حالت توزیع‌شده و متمرکز نشان‌می‌دهد. با توجه به داده‌های این جدول، بیشترین اختلاف بین زمان اجرای متمرکز و توزیع‌شده مربوط به چارچوب پیشنهادی است؛ به‌طوری که به میزان قابل توجه‌ای کاهش یافته‌است. (از ۱۹۶۸/۸۷۷ ثانیه به ۸۳/۱۹۰ کاهش یافته‌است).

۴-۴- نتایج تعداد ویژگی‌ها

تعداد ویژگی‌های انتخاب‌شده هم در حالت توزیع‌شده (D) و هم در حالت متمرکز (C)، با استفاده از پنج روش مختلف انتخاب ویژگی بر روی مجموعه‌داده‌های یادشده در (جدول ۷)، نشان داده‌شده‌است. در تمام حالت‌ها از حدآستانه ۲۵ درصد برای حذف ویژگی‌ها استفاده شده‌است. همان‌طور که دیده‌می‌شود، تعداد ویژگی‌های

انتخاب‌شده به‌وسیلهٔ رویکردهای متمرکز نسبت به توزیع‌شده در بیشتر مجموعه‌داده‌ها، بیشتر است. در مواردی که دقت رده‌بندی در حالت متمرکز نسبت به توزیع‌شده بیشتر است، می‌توان به این نتیجه رسید که از طریق توزیع داده‌ها تنزل قابل توجه‌ای در دقت طبقه‌بندی رخ نداده‌است. در مجموعه‌داده Ozon در بیشتر موارد دقت رده‌بندی به‌وسیلهٔ رویکردهای متمرکز بیشتر از توزیع‌شده به‌دست‌آمده است. در بین مجموعه‌داده‌های مورد ارزیابی، بالاترین دقت در حالت متمرکز نسبت به توزیع‌شده مربوط به مجموعه‌داده Spect، است. این در حالی است که نتایج به‌دست‌آمده به‌وسیلهٔرویکردهای توزیع‌شده تنها دو یا سه درصد پایین‌تر هستند. برای سایر مجموعه‌داده‌ها، بهترین نتایج برای رویکردهای توزیع‌شده گزارش شده‌است. در مجموعه‌داده Connect4 در اکثر موارد دقت طبقه‌بندی به‌وسیلهٔ رویکردهای توزیع‌شده بهبود یافته‌است که این ممکن است، به‌علت تعداد کم ویژگی‌های انتخاب‌شده توسط رویکردهای متمرکز باشد. بالاترین بهبود مربوط به رویکرد پیشنهادی (MRMRBYSVM-RFE) با مجموعه‌داده Spambase همراه با طبقه‌بند SVM است که با وجود کاهش تعداد ویژگی‌های انتخاب‌شده در حالت توزیع‌شده، دقت آن کاهش نیافته است.

۵- نتیجه‌گیری

در این مقاله یک چارچوب جدید برای انتخاب ویژگی مبتنی بر الگوریتم‌های نهفته و فیلتر چندمتغیره پیشنهاد شده‌است. در این چارچوب پیشنهادی، فرایند انتخاب ویژگی به‌صورت توزیع افقی برای مجموعه‌داده‌هایی که تعداد نمونه هایشان بیشتر است، ارائه‌شده است. مراحل اصلی این چارچوب عبارتند از: توزیع افقی مجموعه‌داده‌ها، اجرای فرایند انتخاب ویژگی پیشنهاد شده روی هر زیرمجموعه و ترکیب نتایج نهایی از انتخاب ویژگی پیشنهادی در یک زیرمجموعه. در فرایند انتخاب ویژگی پیشنهادی، استفاده از الگوریتم‌های فیلتر چندمتغیره به‌وسیله الگوریتم‌های نهفته معرفی شده‌است. به‌طوری که انتخاب ویژگی‌های مرتبط مبتنی بر طبقه‌بند از طریق الگوریتم‌های نهفته بررسی شده و تشخیص و کاهش ویژگی‌های افزونه از طریق الگوریتم‌های فیلتر چندمتغیره محاسبه شده‌است. نتایج نشان‌می‌دهد که چارچوب پیشنهادشده در بیشتر مجموعه‌داده‌های آزمایش‌شده، به‌دلیل استفاده از دو نوع روش متنوع برای انتخاب ویژگی به‌منظور بررسی ویژگی‌های مرتبط و افزونه توانسته دقت طبقه‌بندی را بهبود دهد. از طرفی زمان اجرای فرایند انتخاب ویژگی پیشنهادی به‌صورت توزیع‌شده نسبت به‌حالت متمرکز کاهش یافته‌است. همچنین کاهش تعداد ویژگی‌ها باعث کاهش دقت طبقه‌بندی نشده و در اکثر موارد نسبت به‌حالت متمرکز بهبود داشته است.

روش پیشنهادشده، می‌تواند با هر الگوریتم انتخاب ویژگی از نوع نهفته و از نوع فیلتر چندمتغیره استفاده شود. به‌عنوان کار و پژوهش آینده، می‌توان توزیع و تقسیم‌بندی عمودی مجموعه‌داده‌ها را به‌جای افقی بررسی کرد، همچنین می‌توان استراتژی‌های جدید دیگری را برای ترکیب سایر روش‌های انتخاب ویژگی، توسعه داد.

۶- منابع 6- Refrence

[1] I. Guyon, A. Elisseeff, (2003)," An introduction to variable and feature selection", Journal of Machine Learning Research, Vol.3, pp.1157–1182.

[2] I.Guyon, S.Gunn, M.Nikravesh and L.A.Zadeh, (2006), "Feature Extraction: Foundations and Applications", vol. 207, Springer, ISBN-10: 9783540354871.

[3] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos (2013), "A Distributed Wrapper Approach for Feature Selection", ESANN proceedings, Computational Intelligence and Machine Learning, ISBN 978-2-87419-081-0.

[4] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos(2015), "A Distributed Feature Selecion Approach Based on a Complexity Measure", Advances in Computational Intelligence, pp. 15-28.

[5] G. Chandrashekar and F. Sahin (2014), "A survey on feature selection methods", journal of Computers and Electrical Engineering vol. 40, pp.16–28.

[6] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos (2015), "Distributed feature selection: An application to microarray data classification", Applied Soft Computing, vol. 30, pp. 136-150.

[7] V. Bolón-Canedo, N. Sánchez-Maroño, and J. Cerviño-Rabuñal(2013), "Scaling up feature selection: a distributed filter approach", Advances in Artificial Intelligence, pp. 121-130.

[8] L. Morán-Fernández, V. Bolón-Canedo, and A. Alonso-Betanzos (2016), "Centralized vs. distributed feature selection methods based on data complexity measures", Journal of Knowledge-Based Systems, vol. 117 , pp.27–45.

[9] L. Mor´an-Fern´andez, V. Bol´on-Canedo, and A. Alonso-Betanzos(2015), “A Time Efficient Approach for Distributed Feature Selection Partitioning by Features”, Lecture Notes in Computer Science book series (LNCS), vol. 9422, pp.245–254.
[10] L. Yu, H. Liu, (2004),” Efficient feature selection via analysis of relevance and redundancy”, J. Mach. Learn. Res. 5 , 1205–1224.
[11] C. Ding, H. Peng, (2005) “Minimum redundancy feature selection from microarray gene expression data”, Journal of Bioinformatics and computational Biology, Vol.03, No.02, pp.185–205.
[12] R. Kohavi, GH. John (1997), “Wrappers for feature subset selection”, Artificial Intelligence, Vol. 97, Issues 1-2, pp.273–324.
[13] J.Li, K.Cheng, S.Wang, F. Morstatter, and R. P. Trevino(2018),” Feature Selection: A Data Perspective”, Journal of ACM Computing Surveys (CSUR), Vol. 50 ,Issue 6.
[14] A.De Haro Garc´ıa, (2011), “Scaling data mining algorithms. Application to instance and feature selection”, Ph.D. Thesis, University of Granada.
[15] H. Djellali, N. Ghoualmi Zine and N. Azizi (2016), “Two Stages Feature Selection Based on Filter Ranking Methods and SVMRFE on Medical Applications”, Modelling and Implementation of Complex Systems. Lecture Notes in Networks and Systems, Springer, Cham,, vol. 1, pp. 281-293.
[16] H. Min and W. Fangfang (2010), ”Filter-Wrapper Hybrid Method on Feature Selection ”, Second WRI Global Congress on Intelligent Systems (GCIS), pp.98-101.
[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik (2002), “Gene selection for cancer classification using support vector machines”, Journal of Machine Learning, vol. 46, Issue 1-3, pp. 389–422.
[18] D. Boughaci and A.A Alkhawaldeh (2018), ”Three local search-based methods for feature selection in credit scoring”, Vietnam Journal of Computer Science, May 2018, Vol. 5, Issue 2, pp. 107–121.
[19] Q.Wang , J. Wan, F. Nie , B. Liu , C.Yan , and X. Li (2019), “Hierarchical Feature Selection for Random Projection”, IEEE Transactions on Neural Networks and Learning Systems, Vol. 30 , Issue 5, pp. 1581 – 1586.
[20] http://archive.ics.uci.edu/ml/datasets/
[21] I.Guyon, J.Weston, S.Barnhill and V.Vapnik,(2002), "Gene selection for cancer classificationusing support vector machines, " Jornal of Machine Learning, vol.46, pp.389–422.
[22] H. Peng, F. Long, and C. Ding, (2005), “Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and minredundancy, ”IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug..
[23] M.A. Hall, L.A. Smith, (1998), ”Practical feature subset selection for machine learning”, Comput. Sci.98,181–191
[24] I. Kononenko,(1994),” Estimating attributes: analysis and extensions of RELIEF”, Machine Learning: ECML-94, vol. 784, pp 171-182
[25] M. Robnik-Šikonja and I. Kononenko, (2003), "Theoretical and empirical analysis of ReliefF and RReliefF", Machine learning, vol. 53,Issue:1-2, pp. 23-69.

منا شریف‌نژاد، فارغ‌التحصیل دوره کارشناسی‌ارشد رشته مهندسی کامپیوتر- نرم‌افزار است. در حال حاضر دانشجوی دوره دکترای در دانشگاه اراک است. ایشان هم‌اکنون عضو هیئت علمی دانشگاه آزاد اسلامی اراک است. حوزه تخصصی ایشان هوش مصنوعی است.
نشانی رایانامه ایشان عبارت است از:
m.sharifnezhad98@gmail.com

محسن رحمانی، مدرک دکترای خود را در سال ۱۳۸۷ از دانشگاه علم و صنعت ایران دریافت کردند. ایشان هم اکنون عضو هیئت علمی دانشگاه اراک هستند. تخصص ایشان پردازش سیگنال‌های صوتی و هوش مصنوعی است.
نشانی رایانامه ایشان عبارت است از:
m-rahmani@araku.ac.ir

حسین غفاریان، مدرک دکترای خود را از دانشگاه علم و صنعت ایران دریافت کردند. ایشان هم‌اکنون عضو هیئت علمی دانشگاه اراک هستند. تخصص ایشان شبکه‌های رایانه‌ای، کیفیت خدمات، کنترل‌کننده‌های فازی، داده‌کاوی، الگوریتم‌های تکاملی، سامانه‌های حمل و نقل هوشمند است.
نشانی رایانامه ایشان عبارت است از:
h-ghaffarian@araku.ac.ir