# یک روش جدید انتخاب ویژگی یک‌طرفه در دسته‌بندی داده‌های متنی نامتوازن

**جعفر پورامینی [*1]، بهروز مینایی بیدگلی [2] و مهدی اسماعیلی [3]**

[1] گروه مهندسی فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه پیام نور، تهران، ایران

[2] دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران

[3] دانشکده مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد کاشان، کاشان، ایران

## چکیده

**توزیع نامتوازن داده‌ها باعث افت کارایی دسته‌بندها می‌شود. راه‌حل‌های پیشنهادشده برای حل این مشکل به چند دسته تقسیم می‌شوند، که روش‌های مبتنی بر نمونه‌گیری و روش‌های مبتنی بر الگوریتم از مهم‌ترین روش‌ها هستند. انتخاب ویژگی نیز به‌عنوان یکی از راه‌حل‌های افزایش کارایی دسته‌بندی داده‌های نامتوازن مورد توجه قرار گرفته است. در این مقاله یک روش جدید انتخاب ویژگی یک‌طرفه برای دسته‌بندی متون نامتوازن ارائه شده است. روش پیشنهادی با استفاده از توزیع ویژگی‌ها میزان نشان‌گربودن ویژگی را محاسبه می‌کند. به‌منظور مقایسه عملکرد روش پیشنهادی، روش‌های انتخاب ویژگی مختلفی پیاده‌سازی و برای ارزیابی روش پیشنهادی از درخت تصمیم C4.5 و نایوبیز استفاده شد. نتایج آزمایش‌ها بر روی پیکره‌های Reuters-21875 و WebKB برحسب معیار Micro F ، Macro F و G-mean نشان می‌دهد که روش پیشنهادی نسبت به روش‌های دیگر، کارایی دسته‌بندها را به اندازهٔ قابل توجهی بهبود بخشیده است.**

**واژگان کلیدی: انتخاب ویژگی، روش پالایه، داده‌های نامتوازن، دسته‌بندی متون**

# A Novel One Sided Feature Selection Method for Imbalanced Text Classification

**Jafar Pouramini[*1], Behrouz Minaei-Bidgoli[2] & Mahdi Esmaeili[3]**

[1]Department of Computer and Information Technology Engineering, Faculty of Engineering, Payame Noor University, Tehran, Iran

[2]Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran

[3]Faculty of Computer Engineering, Kashan Islamic Azad University, Kashan, Iran

## Abstract

**The imbalance data can be seen in various areas such as text classification, credit card fraud detection, risk management, web page classification, image classification, medical diagnosis/monitoring, and biological data analysis.**

**The classification algorithms have more tendencies to the large class and might even deal with the minority class data as the outlier data. The text data is one of the areas where the imbalance occurs. The amount of text information is rapidly increasing in the form of books, reports, and papers. The fast and precise processing of this amount of information requires efficient automatic methods. One of the key processing tools is the text classification. Also, one of the problems with text classification is the high dimensional data that lead to the impractical learning algorithms. The problem becomes larger when the text data are also imbalance. The imbalance data distribution reduces the performance of classifiers. The various solutions proposed for this problem are divided into several categories, where the sampling-based methods and algorithm-based methods are among the most important methods. Feature selection is also considered as one of the solutions to the imbalance problem. In this research, a new method of one-way feature selection is**

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال:۹۶/۰۸/۱۸ ● تاریخ پذیرش:۹۷/۱۰/۱۹ ● نوع مطالعه: پژوهشی

**presented for the imbalance data classification. The proposed method calculates the indicator rate of the feature using the feature distribution.**

**In the proposed method, the one-figure documents are divided in different parts, based on whether they contain a feature or not, and also if they belong to the positive-class or not. According to this classification, a new method is suggested for feature selection. In the proposed method, the following items are used.**

a) **If a feature is repeated in most positive-class documents, this feature is a good indicator for the positive-class; therefore, this feature should have a high score for this class. This point can be shown as a proportion of positive-class documents that contain this feature. Besides, if most of the documents containing this feature are belonged to the positive-class, a high score should be considered for this feature as the class indicator. This point can be shown by a proportion of documents containing feature that belong to the positive-class.**
b) **If most of the documents that do not contain a feature are not in the positive-class, a high score should be considered for this feature as the representative of this class. Moreover, if most of the documents that are not in the positive class do not contain this feature, a high score should be considered for this feature.**

**Using the proposed method, the score of features is specified. Finally, the features are sorted in descending order based on score, and the necessary number of required features is selected from the beginning of the feature list.**

**In order to evaluate the performance of the proposed method, different feature selection methods such as the Gini, DFS, MI and FAST were implemented. To assess the proposed method, the decision tree C4.5 and Naive Bayes were used. The results of tests on Reuters-21875 and WebKB figures per Micro F , Macro F and G-mean criteria show that the proposed method has considerably improved the efficiency of the classifiers than other methods.**

**Keywords: Feature selection, Imbalanced class, High dimensionality, Text classification**

## ۱- مقدّمه

داده‌های نامتوازن در حوزه‌های مختلفی مانند: دسته‌بندی متون[1]، کشف تخلفات کارت‌های اعتباری[2]، مدیریت ریسک[3]، دسته‌بندی صفحات وب، دسته‌بندی تصاویر، نظارت‌های دارویی[4]، تحلیل داده‌های بیولوژیکی[5] مشاهده می‌شوند.

عدم توازن به دو نوع ذاتی[6] و غیرذاتی[7] تقسیم می‌شود [1]. در عدم توازن ذاتی، طبیعت داده‌ها نامتوازن است. به‌عنوان نمونه، عدم توازن داده‌های حوزه‌هایی مانند کشف تقلب، تشخیص سرطان و پیش‌بینی زلزله و دسته‌بندی متون، از نوع ذاتی هستند؛ اما در برخی موارد، به‌دلیل محدودیت‌هایی مانند هزینه بالای جمع‌آوری نمونه‌ها، مشکلات قانونی و یا مسائل خصوصی امکان جمع‌آوری داده به‌اندازهٔ کافی نیست، به این حالت عدم توازن غیرذاتی یا بیرونی گفته می‌شود [2].

الگوریتم‌های دسته‌بندی تمایل بیشتری به دسته بزرگ دارند، حتی ممکن است با داده‌های دسته اقلیت به‌عنوان داده پرت[8] برخورد کنند[3]. وقتی تعداد نمونه‌های دسته اکثریت خیلی بیشتر از دسته اقلیت باشد و داده‌ها نیز دارای ابعاد بالا باشند، احتمال بیش‌برازش[9] بیشتر می‌شود [4, 5].

داده‌های متنی یکی از حوزه‌هایی است، که عدم توازن در آن بروز می‌کند. حجم اطلاعات متنی در قالب کتاب، گزارش، مقالات به‌سرعت در حال افزایش است. پردازش سریع و دقیق این حجم از اطلاعات نیازمند روش‌های کارآمد خودکار است. یکی از ابزارهای کلیدی پردازش، دسته‌بندی متون است. یکی از مشکلات دسته‌بندی متون، وجود ابعاد بالای داده‌ها است، که باعث می‌شود الگوریتم‌های یادگیری غیر عملی شوند. مشکل، زمانی بزرگ‌تر می‌شود که داده‌های متنی نامتوازن نیز باشند. به‌عنوان نمونه، زمانی که یک پیکره با موضوعات متنوع به دو دسته تقسیم می‌شوند، به‌طوری که یک دسته شامل اسناد یک موضوع خاص و دسته دیگر شامل اسناد همه موضوع‌های دیگر است. این حالت را یک در مقابل همه[10] می‌نامند [6]. به‌منظور حل مشکل داده‌های نامتوازن، روش‌های متعددی معرفی شده‌اند، که عبارتند از: روش‌های مبتنی بر نمونه‌گیری[11]، روش‌های مبتنی بر الگوریتم[12]، روش‌های انتخاب ویژگی[13] [6, 7].

در این مقاله یک روش انتخاب ویژگی جدیدی برای دسته‌بندی داده‌های متنی نامتوازن ارائه شده است. بخش‌های مختلف این نوشتار بدین شرح است: در بخش دوم پیشینه

[1] Text classification
[2] Credit card fraud detection
[3] Risk management
[4] Medical diagnosis/monitoring
[5] Biological data analysis
[6] Intrinsic
[7] Extrinsic
[8] Outlier
[9] Overfitting
[10] One_against_all
[11] Sampling methods
[12] Algorithmic Methods
[13] Feature selection

پژوهش و کارهای انجام شده مورد بررسی قرار می‌گیرد. در بخش سوم روش پیشنهادی معرفی شده است. در بخش چهارم تنظیمات آزمایش‌ها و نتایج آن ارائه شده است. در بخش پنجم نتایج بحث و بررسی شده‌اند. درنهایت در بخش ششم نتیجه‌گیری ارائه شده است.

## ۲- پیشینهٔ پژوهش

در این بخش پژوهش‌های انجام شده درخصوص داده‌های نامتوازن ارائه شده است.

### ۲-۱- روش‌های مبتنی بر نمونه‌گیری

روش‌های نمونه‌گیری توازن داده‌ها را از طریق افزایش نمونه‌های دسته اقلیت و یا کاهش نمونه‌های اکثریت تغییر می‌دهند. اما این روش‌ها مشکل اثرات جانبی[1] دارند [8]. روش‌های بیش‌نمونه‌گیری[2] که داده‌های دسته اقلیت را افزایش می‌دهند، موجب بیش‌برازش [9,4] و روش‌های کم‌نمونه‌گیری[3] که باعث کاهش نمونه‌های دسته اکثریت، می‌شوند، باعث ازدست‌رفتن داده‌های مفید می‌شوند [10]. روش SMOTE[4] برای کاهش اثرات نامتوازن‌بودن داده‌ها ارائه شده است [11]. این روش با اضافه‌کردن نمونه‌های جدید بین نمونه‌های موجود باعث افزایش نمونه‌های دسته اقلیت می‌شوند. این روش به‌طور مؤثر از بیش‌برازش دوری می‌کند. در [12] روش MWMOTE[5] ارائه شد. در این روش، نمونه‌هایی که به‌سختی توسط الگوریتم یاد گرفته می‌شوند، مشخص می‌شوند. برای این نمونه‌ها وزنی در نظر گرفته و این وزن برحسب فاصله اقلیدسی بین این نمونه و نزدیک‌ترین نمونه از دسته اکثریت مشخص می‌شود. در [13] مطالعات و آزمایش‌های متعددی بر روی دسته‌بندی متون نامتوازن با استفاده از دسته‌بند ماشین بردار پشتیبان انجام شد. در این پژوهش روش‌های نمونه‌گیری مانند کم‌نمونه‌گیری و بیش‌نمونه‌گیری مورد بررسی قرار گرفته و نتایج حاصل با روش دیگر برخورد با داده‌های نامتوازن مانند روش استفاده از هزینه برای خطای دسته‌بندی مقایسه شده است.

چن و همکاران برای بهبود دسته‌بندی متون نامتوازن، روش جدیدی به نام DCOM را برای تولید نمونه‌های جدید برای دسته اقلیت ارائه دادند [8]. در این پژوهش با استفاده از اطلاعات معنایی سراسری دسته اقلیت، یک مدل موضوعی احتمالی[6] برای دسته اقلیت ایجاد و سپس با استفاده از آن نمونه‌های جدید تولید می‌شود. مزیت روش ارائه‌شده کاهش احتمال بیش‌برازش است.

بورجو و همکاران با استفاده از مدل مخفی مارکوف[7]، روش جدید بیش‌نمونه‌گیری برای تولید اسناد جدید به نام COS-HMM[8] ارائه کردند [9]. مدل COS-HMM با استفاده از پیکره آموزش داده و سپس این مدل به عنوان موتور تولید اسناد ترکیبی جدید استفاده می‌شود. این مدل جدید در مقایسه با روش‌های ROS[9] و SMOTE کارایی بهتری را نشان داد.

لیو و همکاران با استفاده از روش نوینی برای وزن‌دهی به واژه‌ها، روشی را برای بهبود کارایی دسته‌بندهای نایوبیز و ماشین بردار پشتیبان برای دسته‌بندی متون نامتوازن ارائه دادند [30]. سان و همکاران، مطالعات و آزمایش‌های متعددی را بر روی دسته‌بندی متون نامتوازن با استفاده از دسته‌بند ماشین بردار پشتیبان انجام دادند [31]. در این پژوهش از دسته‌بند ماشین بردار پشتیبان استفاده شد. همچنین در این پژوهش اثر استفاده از برخی روش‌های کم‌نمونه‌گیری و بیش‌نمونه‌گیری مورد بررسی قرار گرفته است.

### ۲-۲- روش‌های مبتنی بر الگوریتم

در روش‌های مبتنی بر الگوریتمی با ایجاد الگوریتم‌های جدید و یا تغییر الگوریتم‌های یادگیری موجود، الگوریتم‌هایی ایجاد می‌شوند که نامتوازن بودن داده‌های آموزشی را در نظر بگیرند. این الگوریتم‌ها به‌گونه‌ای طراحی می‌شوند که کارایی دسته اقلیت را افزایش دهند. روش‌های یادگیری یک‌دسته‌ای[10]، یادگیری حساس به هزینه[11] و ترکیبی[12] از جمله این روش‌ها هستند.

#### ۲-۲-۱- یادگیری یک‌دسته‌ای

در یادگیری یک‌دسته‌ای، یک دسته‌بند، برای شناسایی یک دسته خاص آموزش داده می‌شود. در این روش فقط داده‌های آموزشی دسته مورد نظر به دسته‌بند آموزش داده می‌شود. به همین دلیل کارایی پیش‌بینی دسته مورد نظر افزایش می‌یابد.

[1] Side Effect
[2] Oversampling
[3] Undersampling
[4] Synthetic Minority Over-sampling Technique
[5] Majority Weighted Minority Oversampling TEchnique

[6] Probabilistic Topic Models
[7] Hidden Markov Model
[8] Content-based Over-Sampling HMM
[9] Random Over Sampling
[10] One-class learning
[11] Cost sensitive learning
[12] Ensemble

این روش همواره جواب بهینه ارائه نمی‌دهد، زیرا فقط داده‌های یک دسته برای آموزش استفاده می‌شود و این نکته باعث می‌شود که مرزهای داده‌ها به‌خوبی مشخص نشود [14].

One-Class SVM در حوزه‌های مختلفی مانند کشف هرزنامه[1]، تشخیص ارقام دست‌نوشته، بازیابی اطلاعات[2]، تحلیل اطلاعات با موفقیت استفاده شده است[15].

### ۲-۲-۲- روش‌های ترکیبی

در روش‌های ترکیبی به‌جای استفاده از یک دسته‌بند، از مجموعه‌ای از دسته‌بندها استفاده می‌شود. در بیش‌تر موارد مجموعه دسته‌بندها کارایی بهتری نسبت به یک دسته‌بند دارند. روش‌های ترکیبی به دو دسته Boosting و Bagging تقسیم می‌شوند. در روش‌های Bagging دسته‌بندهای مختلفی با استفاده از نمونه‌گیری با جای‌گذاری[3] ساخته می‌شوند؛ سپس با استفاده از ترکیب نتایج به‌دست‌آمده از دسته‌بندها، نتیجه نهایی ایجاد می‌شود. روش‌های IIVotes, UnderOverBagging, Asymetric Bagging از جمله روش‌های Bagging ارائه‌شده برای داده‌های نامتوازن هستند [3].

در روش‌های Boosting (AdaBoost) از کل مجموعه داده به منظور آموزش استفاده می‌کند، اما بعد از هر بار آموزش، بیشتر بر روی داده‌های سخت تمرکز می‌کند تا به‌درستی دسته‌بندی شوند. در ابتدا تمام رکوردها وزن یکسانی می‌گیرند. وزن نمونه‌هایی که به‌اشتباه دسته‌بندی شده‌اند، افزایش خواهد یافت، درحالی‌که آن دسته از نمونه‌هایی که به‌درستی دسته‌بندی شده‌اند وزنشان کاهش خواهد یافت؛ سپس وزن دیگری به‌صورت مجزا به هر دسته‌بند با توجه به دقت کلی آن اختصاص داده می‌شود که بعداً در مرحلهٔ آزمون مورد استفاده قرار می‌گیرد. روش‌های SMOTEBoost, MSMOTEBoost, RUSBoost, DataBoost-IM از جمله روش‌هایی هستند که بر این اساس برای داده‌های نامتوازن ارائه شده‌اند[3].

*یانگ* و همکاران از نمونه‌گیری استفاده کردند. در این روش، با استفاده از SMOTE نمونه‌های اقلیت افزایش یافته و با استفاده از کم‌نمونه‌گیری نمونه‌های اکثریت کاهش می‌یابند [2]. از ترکیب نمونه‌های دسته اقلیت و همه نمونه‌های دسته اکثریت تعداد زیادی دسته‌بند ایجاد می‌شوند و ترکیب آن‌ها یک دسته‌بند به روش ترکیبی[4] ایجاد کرده و این ترکیب را به‌عنوان دسته‌بند اصلی استفاده کردند.

### ۳-۲-۲- روش‌های حساس به هزینه

روش‌های حساس به هزینه با تخصیص هزینه‌های[5] مختلف برای دسته‌ها، سعی می‌کنند، به‌جای افزایش کارایی کل دسته‌بند، کارایی دسته‌بندی دسته مورد نظر را افزایش دهند. کارایی این روش‌ها بستگی به نحوه تعیین ماتریس هزینه[6] دارد. نحوه تعیین ماتریس هزینه، یکی از مشکلات این روش‌ها است [16]. MetaCost [16] و LCSDM[7] [17] از جمله این روش‌ها هستند.

## ۳-۲- روش‌های انتخاب ویژگی

همواره روش‌های نمونه‌گیری و روش‌های الگوریتمی نمی‌توانند کارایی بالایی در داده‌های نامتوازن با ابعاد بالا داشته باشند [18]. انتخاب ویژگی به‌عنوان یکی از روش‌های حل مشکل عدم توازن نیز مورد توجه قرار گرفته است [5, 7]. پژوهش‌های متعددی کارایی روش‌های انتخاب ویژگی در داده‌های نامتوازن را نشان داده‌اند [19]. نتایج پژوهش‌هایی که درخصوص روش‌های نمونه‌گیری و روش انتخاب ویژگی و ترکیب هر دو روش برای بهبود دسته‌بندی داده‌های متنی نامتوازن با ابعاد بالا انجام شده، نشان داده است که تأثیر روش‌های انتخاب ویژگی نسبت به روش‌های نمونه‌گیری بیشتر است [20].

با توجه به ویژگی‌های داده‌های متنی، روش‌های انتخاب ویژگی معمول، نمی‌تواند کارایی لازم برای انتخاب ویژگی به‌منظور دسته‌بندی متون را داشته باشند. به همین دلیل روش‌های انتخاب ویژگی خاص داده‌های متنی ارائه شده است. روش‌های انتخاب ویژگی در داده‌های متنی بر اساس رویکرد، به دو دسته تقسیم می‌شوند [21]. این دو رویکرد عبارت‌اند از رویکرد نحوی و معنایی و رویکرد آماری:

**رویکرد نحوی و معنایی:** در این رویکرد از ترتیب واژه‌ها، معنای واژه‌ها، وابستگی واژه‌ها به یکدیگر، رابطه بین مفاهیم، نقش واژه‌ها در جمله استفاده می‌شود. در این رویکرد، پیش‌پردازش‌های حذف واژگان توقف و ریشه‌یابی انجام می‌گردد. در این رویکرد از منابع خارج از متون مانند هستان‌شناسی[8] نیز ممکن است استفاده شود [21-23].

[1] Spam detection
[2] Information Retrieval
[3] Bootstrapping
[4] Ensemble

[5] Cost
[6] Cost matrix
[7] Large cost-sensitive margin distribution machine
[8] Ontology

**رویکرد آماری:** در این رویکرد پارامترهایی مانند تعداد تکرار واژه‌ها در یک سند، تعداد اسناد حاوی واژه، تعداد تکرار واژه در یک دسته، توزیع واژه در سندهای مختلف استفاده می‌شود. در این رویکرد از برخی خصوصیات خاص متن مانند، ترتیب واژه‌ها، روابط واژه‌ها و وابستگی واژه‌ها صرف‌نظر می‌شود. در این رویکرد، به‌طور معمول پیش‌پردازش‌هایی مانند حذف واژگان توقف و ریشه‌یابی انجام می‌شود [24-26].

*کوک* و همکاران، روشی به نام SYMON برای انتخاب ویژگی به روش پوشانه برای داده‌های نامتوازن با استفاده از جستجوی هارمونی ارائه کردند [27].

تاکنون روش‌های انتخاب ویژگی زیادی به روش پالایه مانند، بهره اطلاعاتی[1]، اطلاعات متقابل[2]، شاخص جینی[3]، NDM[4] و Chi_square ارائه شده است [30,29]. با توجه به ویژگی‌های خاص داده‌های متنی، پژوهش‌های فراوانی برای ارائه روشی برای انتخاب ویژگی در داده‌های متنی انجام شده است. از جملهٔ این روش‌ها، می‌توان به شاخص جینی [25] و DFS[5] [24] اشاره کرد، که کارایی بهتری نسبت به دیگر روش‌ها در داده‌های متنی نشان داده‌اند. بررسی‌ها و آزمایش‌های مختلفی بر روی‌داده‌های متنی نامتوازن انجام شده است. این پژوهش‌ها نشان دادند که اهمیت انتخاب ویژگی در داده‌های متنی نامتوازن از الگوریتم یادگیری مورد استفاده نیز بیشتر است [30].

*یانلینگ* و همکاران داده‌های متنی نامتوازن را مورد بررسی قرار دادند [31] و عوامل مؤثر بر کارایی دسته‌بندی داده‌های متنی نامتوازن را بررسی کردند و عوامل مؤثر را به شرح زیر اعلام کردند:

**توزیع داده‌ها:** نسبت نمونه‌های دسته کوچک به بزرگ یکی از عوامل مهم در کارایی دسته‌بندی داده‌های نامتوازن است. این نسبت در کاربردهای مختلف متفاوت است.

**هم‌پوشانی دسته‌ها:** اگر همپوشانی بین دسته‌ها زیاد باشد، توزیع نامتوازن، کارایی دسته‌بندها را بیشتر کاهش خواهد داد. اما با افزایش همپوشانی دسته‌ها، حساسیت دسته‌بندهای خطی به توزیع داده‌ها افزایش می‌یابد.

**حجم داده‌های آموزشی:** اگر نسبت نمونه‌های دسته کوچک و بزرگ ثابت فرض شود. حجم داده‌ها می‌تواند کارایی دسته‌بندی را تحت تأثیر قرار دهد، به‌طوری که هرچه تعداد نمونه‌ها کمتر باشند، کارایی کاهش بیشتری خواهد یافت.

**وجود زیردسته‌ها:** وجود زیردسته‌ها باعث پیچیدگی آموزش دسته‌بندها می‌شوند.

*ژنگ* و همکاران چارچوبی را برای انتخاب ویژگی در دسته‌بندی متون در حالت نامتوازن ارائه دادند [26]. در این چارچوب ویژگی ها به دو دسته ویژگی مثبت[6] و ویژگی منفی[7] تقسیم می‌شوند. وجود ویژگی مثبت در یک سند نشان‌دهندهٔ تعلق آن سند به دسته مورد نظر و وجود ویژگی منفی در یک سند نشان‌دهندهٔ عدم تعلق آن سند به دسته مورد نظر است. روش‌های انتخاب ویژگی به دو دسته یک‌طرفه[8] و دوطرفه[9] تقسیم می‌شوند [26]. روش‌های یک‌طرفه فقط ویژگی مثبت را انتخاب و روش‌های دوطرفه، ترکیبی از ویژگی مثبت و منفی را انتخاب می‌کنند. در [26] از روش چندمرحله‌ای برای انتخاب ویژگی استفاده شد. ابتدا ویژگی‌های مثبت و سپس ویژگی‌های منفی انتخاب شد. در این پژوهش فرض می‌شود که L تعداد ویژگی‌های مورد نیاز است که توسط کاربر تعیین می‌شود. ابتدا L1 ویژگی t با بالاترین مقدار f(t,c) انتخاب می‌شوند. تابع f(t,c) میزان نمایان‌گربودن ویژگی t برای دسته c را نشان می‌دهد. بیشینه مقدار تابع f(t,c) زمانی است که وقوع t نشان‌دهندهٔ دسته c است؛ سپس L2=L-L1 ویژگی که کمترین مقدار $f(t,\bar{c})$ را دارند، انتخاب می‌شوند. نسبت L1/L2 پارامتر مهمی در این راه‌کار است. بر اساس نتایج این پژوهش، انتخاب ویژگی، کارایی دسته‌بندی داده‌های متنی نامتوازن را به‌طور قابل توجهی افزایش می‌دهد. نتایج این پژوهش به‌خوبی اهمیت انتخاب ویژگی را برای دسته‌بندی داده‌های متنی نامتوازن نشان داد. همچنین نحوه ترکیب ویژگی‌های مثبت و منفی و تأثیر آن در دسته‌بندی داده‌های متنی نامتوازن مورد ارزیابی قرار گرفت [6]. در این پژوهش تأثیر ترکیب صریح[10] و ضمنی[11] ویژگی‌های مثبت و منفی مقایسه شد. برای کنترل صریح نسبت تعداد ویژگی‌های مثبت و منفی از روش‌های یک‌طرفه استفاده شد. نرخ ترکیب ۰/۱ تا ۰/۹تغییر داده شد. نرخ ترکیب از رابطه (۱) محاسبه می‌شود:

$$SR \equiv \frac{n_+}{N} = \frac{n_+}{n_+ + n_-} \qquad (۱)$$

[1] Information gain
[2] Mutual information
[3] Gini index
[4] Normalized different measure
[5] Distinguishing feature selector

[6] Positive Feature
[7] Negative Feature
[8] One-Sided
[9] Two-Sided
[10] Explicitly
[11] Implicitly

N تعداد ویژگی‌ها، n+ تعداد ویژگی های مثبت و n- تعداد ویژگی‌های منفی است. نتایج این پژوهش نشان داد، که بهترین کارایی در ترکیب صریح ویژگی‌های مثبت و منفی به‌دست می‌آید. هر چند رسیدن به بالاترین کارایی، مستلزم تعیین نسبت مناسب تعداد ویژگی‌های مثبت و منفی بر اساس تعداد ویژگی‌ها مورد نیاز است. در [33,32] روش‌هایی با ترکیب روش‌های پالایه و پوشانه ارائه شده است.

عملکرد روش پیشنهادی با روش‌های مختلفی که در پژوهش‌ها عملکرد بهتری داشته‌اند، مقایسه می‌شود. این روش‌ها در زیر به‌اختصار تشریح می‌شوند. از تعاریف زیر برای بیان روش محاسبه روش‌های مختلف استفاده می‌شود:

$P(c)$: احتمال این‌که سند x به دستهٔ c متعلق است.

$P(\bar{c})$: احتمال این‌که سند x به دستهٔ c متعلق نیست.

$P(c|t)$: احتمال این‌که سند x به دسته c متعلق باشد به شرط اینکه سند x شامل واژه t است.

$P(t|c)$: احتمال این‌که سند x شامل واژه tباشد، به شرط این‌که سند x متعلق دسته c باشد.

$P(\bar{c}|t)$: احتمال این‌که سند x به دسته c متعلق نباشد به شرط این‌که سند x شامل واژه t است.

به همین ترتیب بقیه موارد هم مانند $P(c|t)$ تعریف می‌شوند.

ارزیابی ویژگی با تغییر حد آستانه (FAST[1])، روشی برای انتخاب ویژگی در داده‌های نامتوازن است [18]. این روش با تغییر حدآستانه[2]، برای یک دسته‌بند با یک ویژگی، مقدار AUC[3] را اندازه‌گیری می‌کند. در [18] نشان داده شده است که این روش از RELIEF و ضریب همبستگی[4] بهتر عمل می‌کند.

شاخص جینی برای اندازه‌گیری میزان ناخالصی در ساخت درخت تصمیم استفاده می‌شود. رابطه (۲) نحوه محاسبه شاخص جینی را نشان می‌دهد [34].

$$Gini(t) = 1 - \sum_{i=1}^{m} P_i^2 \qquad (۲)$$

m، تعداد دسته‌ها و $P_i$ احتمال تعلق هر نمونه به دسته $c_i$ را نشان می‌دهد. اگر نمونه‌ها در تمام دسته‌ها به‌طور مساوی توزیع شوند، بیشینهٔ این رابطه به دست می‌آید که برابر با 1-(1/m) است. در مرجع [25] نسخه تغییر یافته‌ای از شاخص جینی برای دسته‌بندی متون به‌صورت رابطه (۳) ارائه شده است.

$$Gini(t) = \sum_{i=1}^{m} P(t|c_i)^2 P(c_i|t)^2 \qquad (۳)$$

بهترین حالت زمانی رخ می‌دهد که یک واژه، فقط در یک دسته وجود داشته و در دسته‌های دیگر نباشند. در بهترین حالت این رابطه عدد یک را به دست می‌دهد. در این مقاله از رابطه (۳) استفاده شده است. روش جینی جدید نسبت به روش‌های بهره اطلاعاتی و Chi-square در دسته‌بندی متون عملکرد بهتری دارد [25].

انتخاب‌کننده ویژگی‌های متمایز (DFS[5]) یک روش انتخاب ویژگی خاص داده‌های متنی است. رابطه (۴) نحوه محاسبه امتیاز واژه t را نشان می‌دهد. آزمایش‌های مختلف نشان داده است که، DFS نسبت به Chi-square و بهره اطلاعاتی عملکرد بهتری دارد [24].

$$DFS(t) = \sum_{i=1}^{M} \frac{P(c_i|t)}{P(\bar{t}|c_i)+P(t|\bar{c}_i)+1} \qquad (۴)$$

روش اطلاعات متقابل[6] از رابطه (۵) برای محاسبه امتیاز واژه t در دسته C استفاده می‌کند [35].

$$I(t,c) = log\frac{P(c|t)}{P(t)} \qquad (۵)$$

## ۳- روش پیشنهادی

یکی از مشکلات روش‌های انتخاب ویژگی موجود در دسته‌بندی داده‌های متنی نامتوازن، ترکیب ویژگی‌های مثبت و منفی است. روش‌های موجود انتخاب ویژگی، وزن یکسانی برای این دو نوع ویژگی که نشان‌دهنده دو دسته مختلف هستند، در نظر می‌گیرند. درحالی‌که در دسته‌بندی داده‌های نامتوازن تأکید بر کارایی دسته اقلیت است [36]. به‌عنوان نمونه OCFS[7] مقدار وزن بزرگ‌تر را به جای دسته کوچک‌تر به دسته عمومی‌تر تخصیص می‌دهد. به همین دلیل کارایی این روش برای دسته‌های عمومی بیشتر است تا دسته‌های کوچک. بنابراین برای داده‌های نامتوازن مناسب نیست [37]. به همین دلیل روش‌های موجود، در حالت نامتوازن کارایی لازم را ندارند. در این بخش روش جدیدی برای انتخاب ویژگی در داده‌های متنی نامتوازن ارائه می‌شود. تمرکز این مقاله داده‌های نامتوازن دوردای است.

در این روش، هدف، یافتن ویژگی‌های مثبت است. داده‌های نامتوازن به دو دستهٔ منفی (اکثریت) و مثبت (اقلیت)

[1] Feature Assessment by Sliding Thresholds
[2] Threshold
[3] Area Under Curve
[4] Correlation Coefficient
[5] Distinguishing Feature Selector
[6] Mutual Information
[7] Optimal Orthogonal Centroid Feature Selection

تقسیم می‌شوند. روش پیشنهادی یک روش یک‌طرفه است که ویژگی مثبت را می‌یابد. در روش‌های موجود انتخاب ویژگی متون برخی از وضعیت‌ها برای محاسبه میزان متمایزبودن یک واژه مانند t استفاده نشده است. به‌عنوان نمونه DFS و MI از $P(t|c)$ برای محاسبه امتیاز ویژگی استفاده نمی‌کند. در روش جینی ویژگی‌هایی که تعداد تکرار پایین دارند، امتیاز پایینی می‌گیرند.

در روش پیشنهادی سعی شده است از پارامترهای بیشتری نسبت به روش‌های مشابه برای اندازه‌گیری ویژگی‌های مثبت استفاده شود.

اسناد یک پیکره بر اساس این‌که حاوی یک ویژگی مانند $t_i$ باشند یا نباشند و همچنین متعلق به رده مثبت باشند یا خیر به بخش‌های مختلف، به‌صورتی‌که در جدول (۱) نشان داده شده است، تقسیم می‌شوند.

**(جدول-۱): تقسیم‌بندی اسناد بر اساس یک ویژگی**
**(Table-1): Documents based on one feature**

| | تعداد اسناد حاوی ویژگی $t_i$ | تعداد اسناد فاقد ویژگی $t_i$ | مجموع |
|---|---|---|---|
| رده مثبت (positive) | $a_{t_i}$ | $c_{\bar{t}_i}$ | $n_p$ |
| رده منفی (negative) | $b_{t_i}$ | $d_{\bar{t}_i}$ | $n_n$ |
| مجموع | $n_{t_i}$ | $n_{\bar{t}_i}$ | |

برای تشریح روش پیشنهادی از پارامترهای جدول (۱) استفاده می شود. در بخش زیر مواردی که در نحوهٔ محاسبه امتیاز یک ویژگی در نظر گرفته شده، بیان شده است.

الف- اگر یک ویژگی در بیشتر اسناد رده مثبت تکرار شود، این ویژگی یک نشان‌گر خوب برای رده مثبت است، بنابراین این ویژگی برای این رده امتیاز بالایی باید داشته باشد. این نکته را می‌توان با نسبتی از اسناد رده مثبت که حاوی این ویژگی هستند، نشان داد؛ بنابراین نسبت $a_{t_i}/n_p$ باید مقدار بالایی داشته باشد. همچنین اگر بیشتر اسنادی که حاوی این ویژگی هستند، متعلق به رده مثبت باشند، امتیاز بالایی باید برای این ویژگی به‌عنوان نشان‌گر رده در نظر گرفت. این نکته را می‌توان با نسبتی از اسناد حاوی ویژگی که متعلق به رده مثبت هستند، نشان داد. بنابراین نسبت $a_{t_i}/n_{t_i}$ باید زیاد باشد. با استفاده از دو پارامتر یادشده، نشان‌گربودن ویژگی را با رده با استفاده از رابطه (۶) محاسبه می‌شود:

$$v1_{t_i} = \frac{a_{t_i}}{(a_{t_i}+c_{\bar{t}_i})} + \frac{a_{t_i}}{(a_{t_i}+b_{t_i})} = \frac{a_{t_i}}{n_p} + \frac{a_{t_i}}{n_{t_i}} \quad (۶)$$

ب- اگر بیشتر اسنادی که حاوی یک ویژگی نیستند در رده مثبت نباشند، امتیاز بالایی باید برای این ویژگی به‌عنوان نماینده رده مثبت در نظر گرفت. این نکته یعنی نسبت $d_{\bar{t}_i}/n_{n_{\bar{t}_i}}$ باید بالا باشد. همچنین اگر بیشتر اسنادی که در رده مثبت نیستند، حاوی این ویژگی هم نباشند، باید امتیاز بالایی برای این ویژگی در نظر گرفت؛ این نکته یعنی نسبت $d_{\bar{t}_i}/n_n$ باید مقدار بالایی داشته باشد:

$$v2_{t_i} = \frac{d_{\bar{t}_i}}{(c_{\bar{t}_i}+d_{\bar{t}_i})} + \frac{d_{\bar{t}_i}}{(b_{t_i}+d_{\bar{t}_i})} = \frac{d_{\bar{t}_i}}{n_{\bar{t}_i}} + \frac{d_{\bar{t}_i}}{n_n} \quad (۷)$$

با توجه به این‌که در رابطه (۶) صورت هر یک از کسرها همواره از مخرج همان کسرها کوچک‌تر است، بنابراین هر کسر بین صفر و یک و درنتیجه هر یک از روابط (۶ و ۷) بین صفر و دو است. با استفاده از هنجارسازی بیشینه-کمینه هر یک از روابط (۶ و ۷) به مقادیر بین صفر و یک تبدیل می‌شود.

برای اینکه یک ویژگی نشان‌گر مناسبی برای یک رده باشد، باید مقدار روابط (۶ و ۷) محاسبه‌شده برای آن مقدار بالایی داشته باشد. بنابراین میزان نشان‌گربودن یک ویژگی برای یک رده را می‌توان با استفاده از رابطه (۸) محاسبه کرد:

$$value(t_i) = MinMax(v1_{t_i}) * MinMax(v2_{t_i}) \quad (۸)$$

کمینه امتیاز محاسبه‌شده برای یک ویژگی صفر و بیشینه یک است. با استفاده از $value(t_i)$، امتیاز ویژگی‌ها مشخص می‌شوند. درنهایت ویژگی‌ها به‌صورت نزولی بر حسب value مرتب می‌شوند و تعداد لازم از ویژگی‌های مورد نیاز از ابتدای فهرست ویژگی‌ها انتخاب می‌شوند.

## ۴- ارزیابی روش پیشنهادی

جهت بررسی و ارزیابی روش پیشنهادی از دو مجموعه‌داده استفاده می‌شود، یک مجموعه‌داده کوچک که برای شرح عملکرد روش پیشنهادی ایجاد شده است تا عملکرد آن را نشان دهد و مجموعه‌داده واقعی و استاندارد تا کارایی روش پیشنهادی را در مقایسه با روش‌های مشابه نمایش دهد.

جهت بررسی عملکرد روش پیشنهادی از جدول (۲) به‌عنوان داده نمونه استفاده می‌شود. جدول (۲)، ده سند را

نشان می‌دهد که به‌ترتیب با Doc1,Doc2,. . .Doc10 نشان داده شده است. ستون class دسته هر سند را مشخص می‌کند. اسناد دسته طبقهٔ یک نشان‌دهنده دسته اقلیت و طبقهٔ دو نشان‌دهندهٔ دسته اکثریت است. تعداد اسناد دسته اقلیت دو و تعداد اسناد دسته اکثریت هشت است، بنابراین نرخ عدم توازن ۰/۲ است. تعداد واژه‌ها نیز چهار در نظر گرفته شده است که به‌ترتیب با Term1, Term2, Term3, Term4 نشان داده شده است. وجود عدد یک در سطر i و ستون j نشان‌دهنده وجود واژه j در سند i است و صفر نشان‌دهنده عدم وجود واژه در سند است. بررسی جدول (۲) نشان می‌دهد که term1 نشان‌گر خوبی برای دسته class1 است و term2 نشانگر خوبی برای این دسته نیست. جدول (۳) امتیازهای محاسبه‌شده توسط روش پیشنهادی برای هر ویژگی را نشان می‌دهد. جدول (۳) نشان می‌دهد واژه term1 بالاترین امتیاز و term2 کمترین امتیاز را دارد.

**(جدول-۲): مجموعه داده نمونه**

**(Table-2): Sample data set**

| Document | Class | Term1 | Term2 | Term3 | Term4 |
|---|---|---|---|---|---|
| Doc1 | Class1 | 1 | 1 | 1 | 0 |
| Doc2 | Class1 | 1 | 1 | 0 | 1 |
| Doc3 | Class2 | 0 | 1 | 1 | 1 |
| Doc4 | Class2 | 0 | 1 | 1 | 1 |
| Doc5 | Class2 | 0 | 1 | 1 | 1 |
| Doc6 | Class2 | 0 | 1 | 1 | 0 |
| Doc7 | Class2 | 0 | 1 | 1 | 0 |
| Doc8 | Class2 | 0 | 1 | 1 | 0 |
| Doc9 | Class2 | 0 | 1 | 1 | 0 |
| Doc10 | Class2 | 0 | 1 | 0 | 0 |

**(جدول-۳): امتیازات محاسبه شده هر ویژگی**

**(Table-3): Scores calculated for each feature**

| Term1 | Term2 | Term3 | Term4 |
|---|---|---|---|
| 1 | 0 | 0.096 | 0.273 |

به‌عنوان مثال دوم از عملکرد روش پیشنهادی، جدول داده نمونه چهار در نظر گرفته می‌شود. term1 تنها در class1 قرار دارد و در همه اسناد آن نیز تکرار شده است. term2 هم در طبقهٔ یک و هم طبقهٔ دو وجود دارد و بنابراین باید امتیاز پایین‌تری نسبت به term1 داشته باشد. term3 نسبت به term2 در class2 بیشتر تکرار شده است و بنابراین باید امتیاز پایین‌تری نسبت به term2 داشته باشد. همچنین term4 مانند term3 در class2 تکرار شده است، اما در همه اسناد class1 نیست؛ درنتیجه باید امتیاز کمتری نسبت به term3 بگیرد. با توجه به اطلاعات جدول (۴) ویژگی term1 بهترین ویژگی برای رده class1 است و به‌ترتیب ویژگی‌های term2 وterm3 و درنهایت term4 بدترین ویژگی است. جدول (۵) امتیازهای ویژگی‌ها را نشان می‌دهد. ترتیب امتیازات محاسبه‌شده نیز این موضوع را نشان می‌دهد.

**(جدول-۴): مجموعه‌دادهٔ نمونه**

**(Table-4): Sample data set**

| Document | Class | Term1 | Term2 | Term3 | Term4 |
|---|---|---|---|---|---|
| Doc1 | Class1 | 1 | 1 | 1 | 1 |
| Doc2 | Class1 | 1 | 1 | 1 | 0 |
| Doc3 | Class2 | 0 | 1 | 1 | 1 |
| Doc4 | Class2 | 0 | 0 | 1 | 1 |
| Doc5 | Class2 | 0 | 0 | 0 | 0 |
| Doc6 | Class2 | 0 | 0 | 0 | 0 |
| Doc7 | Class2 | 0 | 0 | 0 | 0 |
| Doc8 | Class2 | 0 | 0 | 0 | 0 |
| Doc9 | Class2 | 0 | 0 | 0 | 0 |
| Doc10 | Class2 | 0 | 0 | 0 | 0 |

**(جدول-۵): امتیازات محاسبه‌شده هر ویژگی**

**(Table-5): Scores calculated for each feature**

| Term1 | Term2 | Term3 | Term4 |
|---|---|---|---|
| 1 | 0.781 | 0.656 | 0.334 |

روش DFS نزدیک‌ترین روش به روش پیشنهادی است. برای مقایسه عملکرد روش DFS و روش پیشنهادی مجموعه‌دادهٔ جدول (۶) ارائه شده است. جدول (۷) امتیازهای محاسبه‌شده توسط هر دو روش را نشان می‌دهد.

**(جدول-۶): مجموعه‌دادهٔ نمونه**

**(Table-6): Sample data set**

| ردیف | رده | ویژگی | |
|---|---|---|---|
| | | Term1 | Term2 |
| 1 | Class1 | 1 | 1 |
| 2 | Class1 | 0 | 1 |
| 3 | Class2 | 1 | 1 |
| 4 | Class2 | 0 | 1 |
| 5 | Class2 | 0 | 1 |
| 6 | Class2 | 0 | 1 |
| 7 | Class2 | 0 | 0 |
| 8 | Class2 | 0 | 0 |
| 9 | Class2 | 0 | 0 |
| 10 | Class2 | 0 | 0 |

**(جدول-۷): امتیازات محاسبه‌شده هر ویژگی**

**(Table-7): Scores calculated for each feature**

| روش | ویژگی | |
|---|---|---|
| | Term1 | Term2 |
| روش پیشنهادی | 0.437 | 0.5 |
| DFS | 0.51828 | 0.48889 |

جدول (۷) نشان می‌دهد که روش DFS ویژگی term1 را انتخاب می‌کند درحالی‌که روش پیشنهادی ویژگی term2 را انتخاب کرده است. با بررسی داده‌های جدول مشخص است که term2 نسبت به term1 نشانگر بهتری است.

علاوه‌بر انجام آزمایش‌هایی با داده‌های نمونهٔ تولیدشده، آزمایش‌هایی با داده‌های واقعی نیز انجام شد و نتایج آن گزارش شده است. برای ارزیابی روش پیشنهادی آزمایش‌های مختلفی انجام شد. در این بخش پیکره‌های مورد استفاده و پیش‌پردازش‌های انجام‌شده روی آن‌ها تشریح می‌شود. فرآیند انجام آزمایش‌ها و همچنین معیارها و نتایج آزمایش‌ها نیز معرفی می‌شود.

## ۴-۱- داده‌های مورد استفاده

برای انجام آزمایش‌ها از پیکره‌های استاندارد و با ویژگی‌های متفاوت استفاده شده است. این پیکره‌ها بارها توسط پژوهش‌گران مورد استفاده قرارگرفته‌اند. نخستین پیکره مورد استفاده، پیکره Reuters-21875 است که شامل ۵۲ دسته مختلف و نامتوازن هستند و اسناد می‌توانند در چند دسته قرار گیرند. دومین پیکره مورد استفاده WebKB است، که شامل چهار دسته از اسناد و هر سند متعلق به یک دسته است. پس از استخراج متون لازم از پیکره، عملیات پیش‌پردازش، شامل استخراج واژه‌ها، حذف واژه‌های توقف، و ریشه‌یابی به روش پورتر[1] انجام شد.

با توجه به اینکه موضوع پژوهش درخصوص داده‌های نامتوازن است. دسته‌هایی انتخاب شد که نسبت به یکدیگر نامتوازن هستند. برای بررسی جامع‌تر، آزمایش‌هایی به‌صورت یک در مقابل همه نیز انجام شد. به این صورت که دسته اقلیت دسته‌ای با تعداد کم و دسته اکثریت ترکیبی از دسته‌های دیگر است. با توجه به اینکه در پیکره Reuters-21875 می‌تواند متعلق به چند دسته نیز باشد، بنابراین اسنادی در دسته اکثریت قرار گرفت که در دسته اقلیت نباشند. در این حالت تعداد اسناد از دسته‌های دیگر به‌گونه‌ای انتخاب شد که تعداد آن ده برابر تعداد اسناد دسته اقلیت باشد به‌عبارت دیگر نرخ عدم توازن یک به ده باشد. نرخ عدم توازن از رابطه (۹) محاسبه شد. $n_{minor}$ تعداد اسناد دسته اقلیت و $n_{all}$ تعداد کل اسناد است. جدول (۸) دسته‌های مورد استفاده پیکره Reuters-21875 را نشان می‌دهد.

$$rate = \frac{n_{minor}}{n_{all}} \qquad (۹)$$

(جدول- ۸): دسته‌های استفاده شده از پیکره Reuters-21875

(Table-8): Characteristics of text data sets used in experiments.

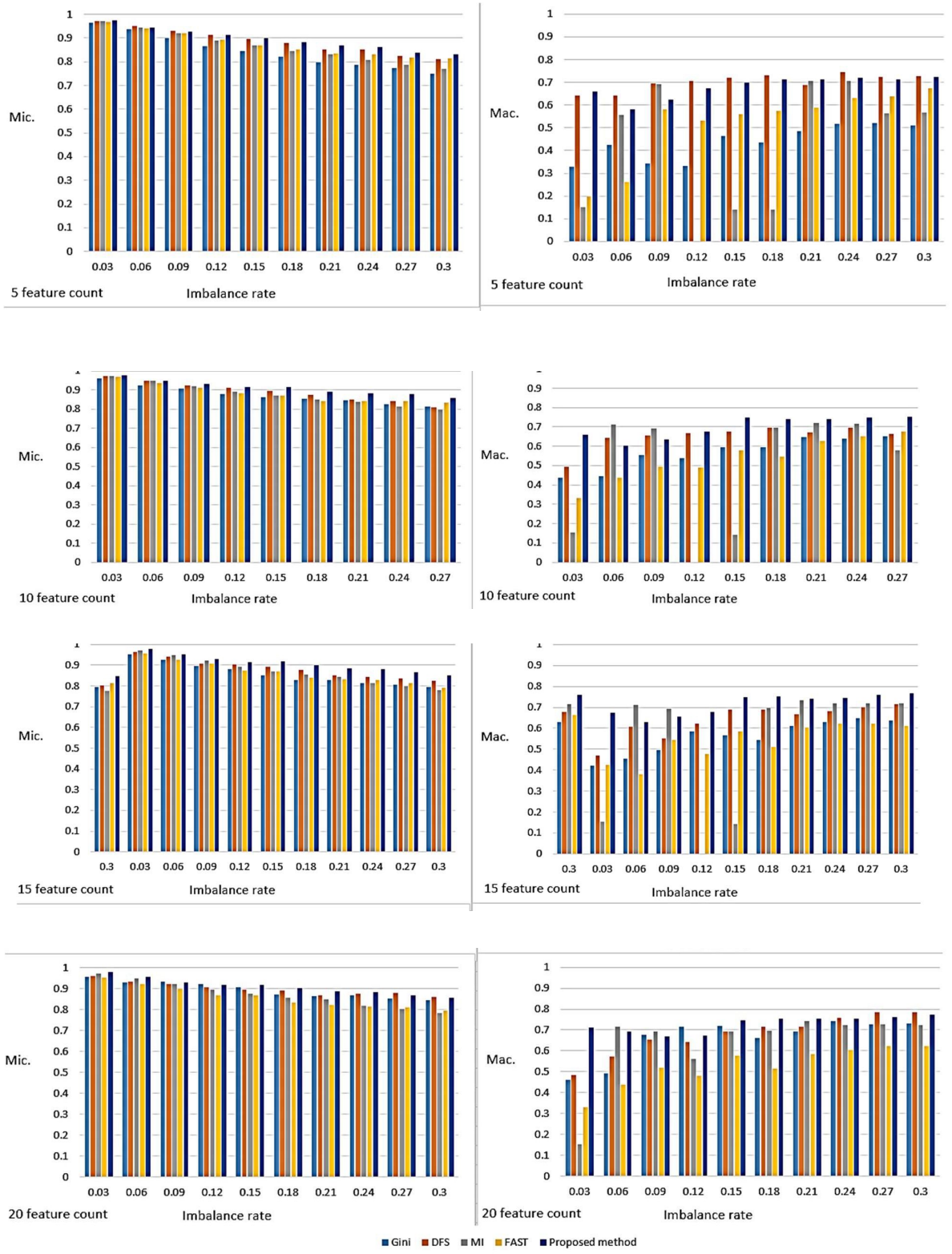
| پیکره | دسته اقلیت | تعداد اسناد دسته اقلیت | دسته اکثریت | تعداد اسناد دسته اکثریت | نرخ عدم توازن | تعداد ویژگی‌ها |
|---|---|---|---|---|---|---|
| **WebKB** | Project | 504 | Student | 1641 | 0.3 | 7211 |
| **Reuters-21875** | Ship | 290 | Acq | 2383 | 0.12 | 15350 |
| **Reuters-21875** | Corn | 252 | ترکیبی تصادفی از دسته‌های دیگر که (All) نیستند Cornشامل | 2520 | 0.1 | 14465 |

داده‌ها به پنج بخش تقسیم شدند، به این صورت که چهار بخش برای آموزش و یک بخش برای آزمون استفاده شد. در مرحلهٔ نخست از آزمایش، بخش نخست داده‌ها برای آزمون و چهار بخش باقیمانده برای آموزش استفاده شد و مرحله دوم، بخش دوم داده‌ها به‌عنوان داده آزمون و چهار بخش باقیمانده به‌عنوان داده آموزش استفاده شد. این مراحل پنج بار به همین ترتیب انجام شد. میانگین پنج مرحله به‌عنوان کارایی یک آزمایش در نظر گرفته شد. به‌منظور ارزیابی جامع‌تر نرخ عدم توازن، تعداد اسناد انتخاب شده دسته اقلیت از ده تا صد درصد افزایش داده شد، تا نرخ‌های عدم توازن متفاوتی ایجاد شود، و تأثیر روش‌های انتخاب ویژگی در نرخ‌های مختلف عدم توازن بررسی شوند. کارایی به‌ازای هر یک از نرخ‌ها و دسته‌بندها ارزیابی شد. به‌عنوان نمونه برای Corn-All نرخ عدم توازن از ۰/۰۱ تا ۰/۱ در آزمایش‌ها تغییر کرده است. تعداد ویژگی‌های انتخاب‌شده برای دسته‌بندی ۵، ۱۰، ۱۵ و ۲۰ در نظر گرفته شد. در دسته‌بندی داده‌های نامتوازن وضعیت دسته اقلیت مهم است[36]. بنابراین برای ارزیابی روش‌های مختلف از معیار F دسته اقلیت استفاده و از دسته‌بند C4.5 و نایوبیز به‌عنوان دسته‌بند استفاده شد.

## ۴-۲- نتایج آزمایش‌ها

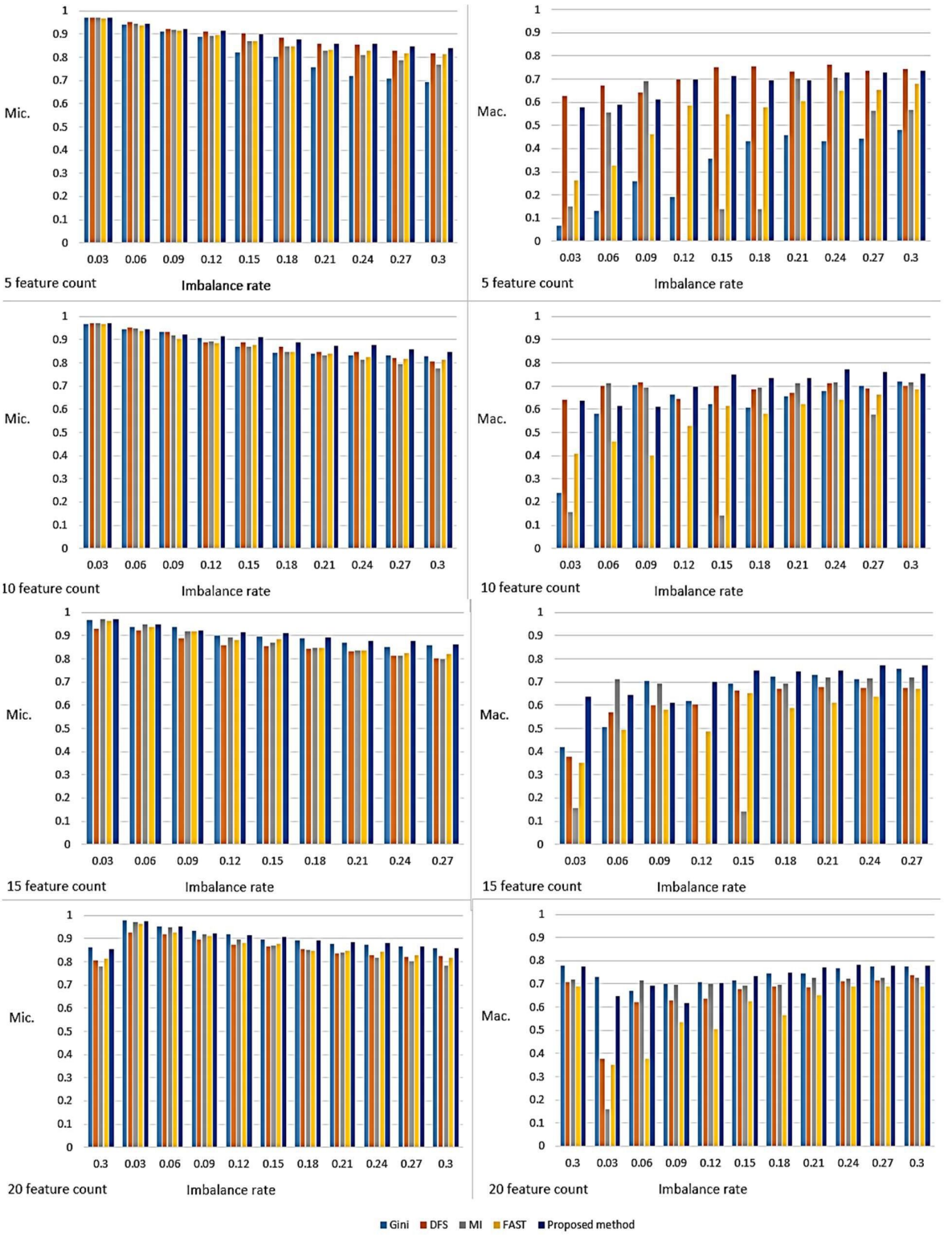
نتایج آزمایش‌های انجام‌شده با استفاده از داده‌های جدول (۳) در نمودارهای شکل‌های (۱ تا ۶) نشان داده شده است. محور افقی نشان‌دهنده نرخ عدم توازن و محور عمودی نشان‌دهندهٔ سنجه Micro F و یا Macro F است. هر نمودار نشان‌دهنده مقایسه روش پیشنهادی با روش‌های دیگر با استفاده از یکی از دسته‌بندها برای یک پیکره مشخص است. هر نمودار عملکردها را برای تعداد مشخصی از ویژگی‌ها نشان می‌دهد.

[1] Porter



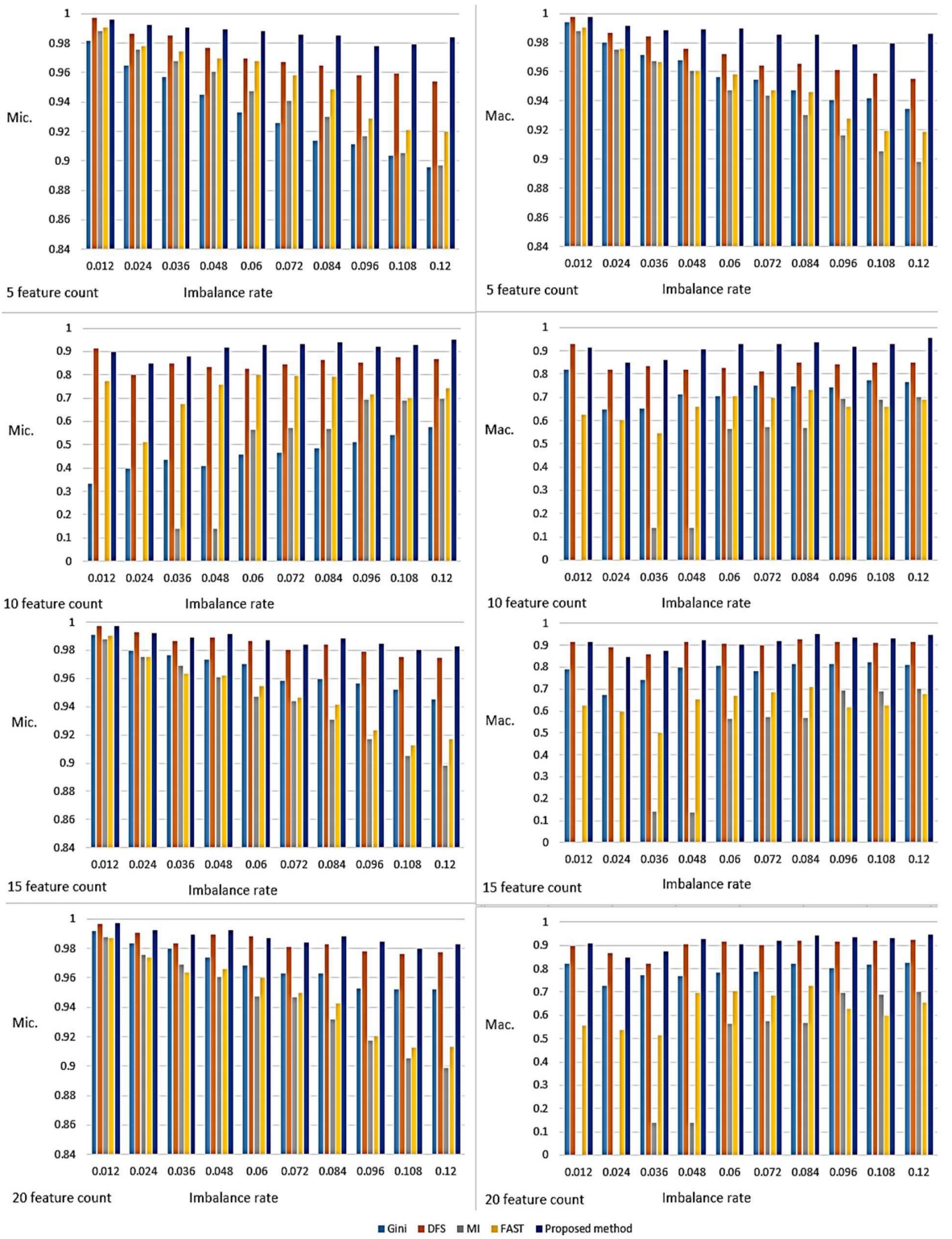

(شکل-۱): مقایسه کارایی دسته‌بند C4.5 دسته‌های Student و Project با استفاده از روش‌های مختلف انتخاب ویژگی بر حسب Micro F و Macro F

**(Figure-1): Comparison of the performances of C4.5 classifier in Student and Project classes using various methods of feature selection based on Micro F and Macro F.**



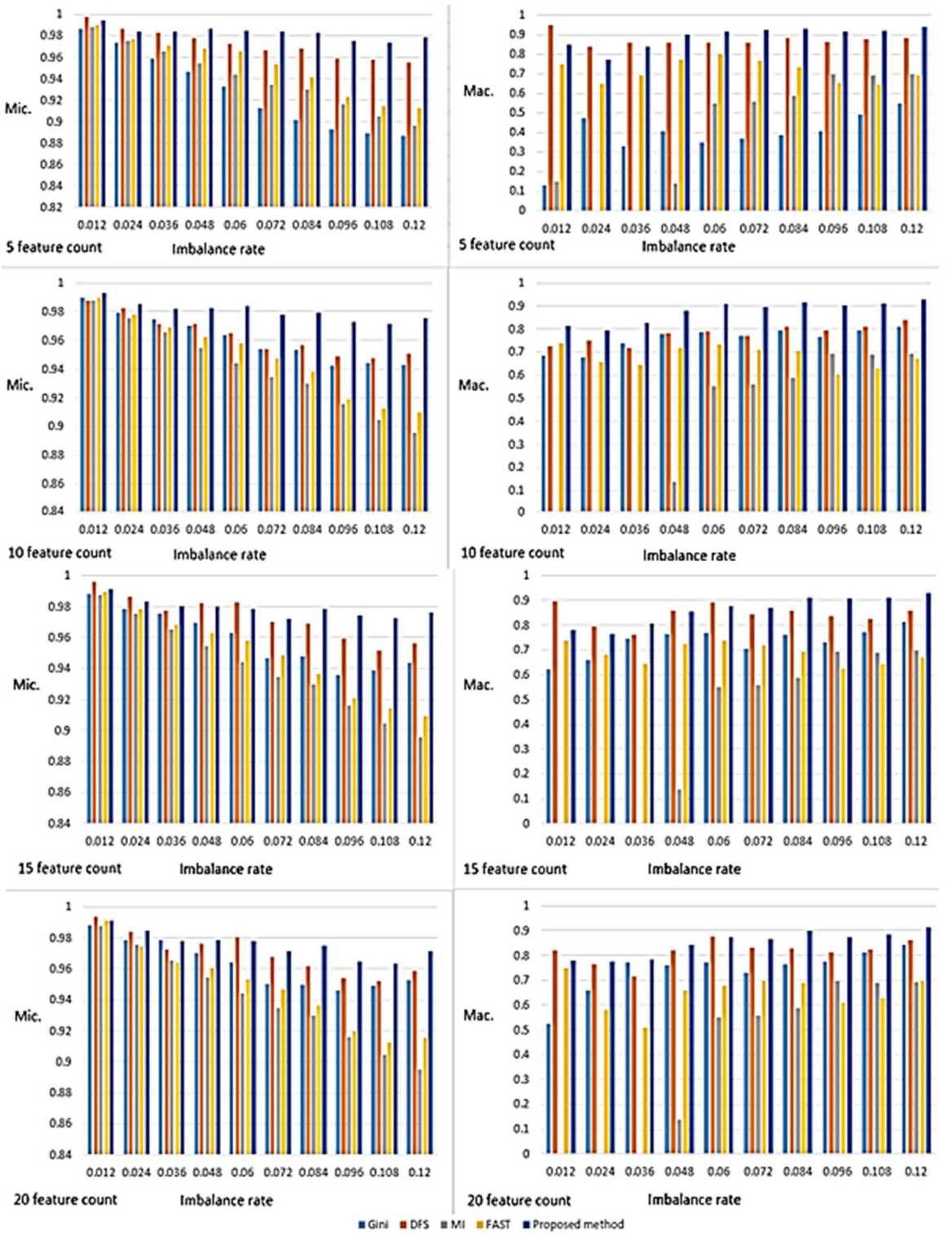

(شکل-۲): مقایسه کارایی دسته‌بند نایو بیزی دسته‌های Student و Project با استفاده از روش‌های مختلف انتخاب ویژگی بر حسب Micro F و Macro F

**(Figure-2): Comparison of the performances of Naive Bayes classifier in Student and Project classes using various methods of feature selection based on Micro F and Macro F**



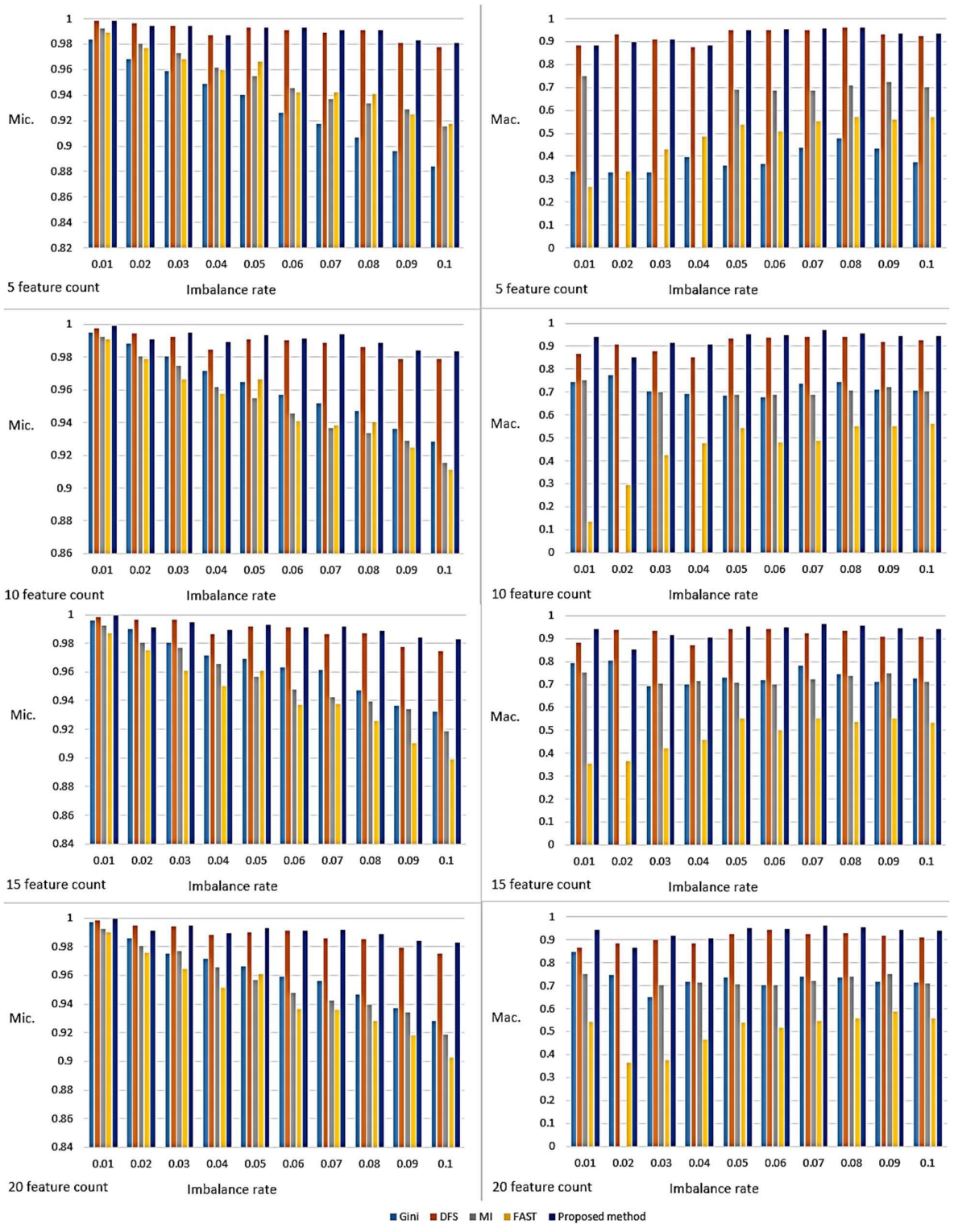

(شکل-۳): مقایسه کارایی دسته‌بند C4.5دسته‌های Ship و Acq با استفاده از روش‌های مختلف انتخاب ویژگی برحسب Micro F و Macro F

**(Figure-3): Comparison of the performances of C4.5 classifier in Acq and Ship classes using various methods of feature selection based on Micro F and Macro F.**



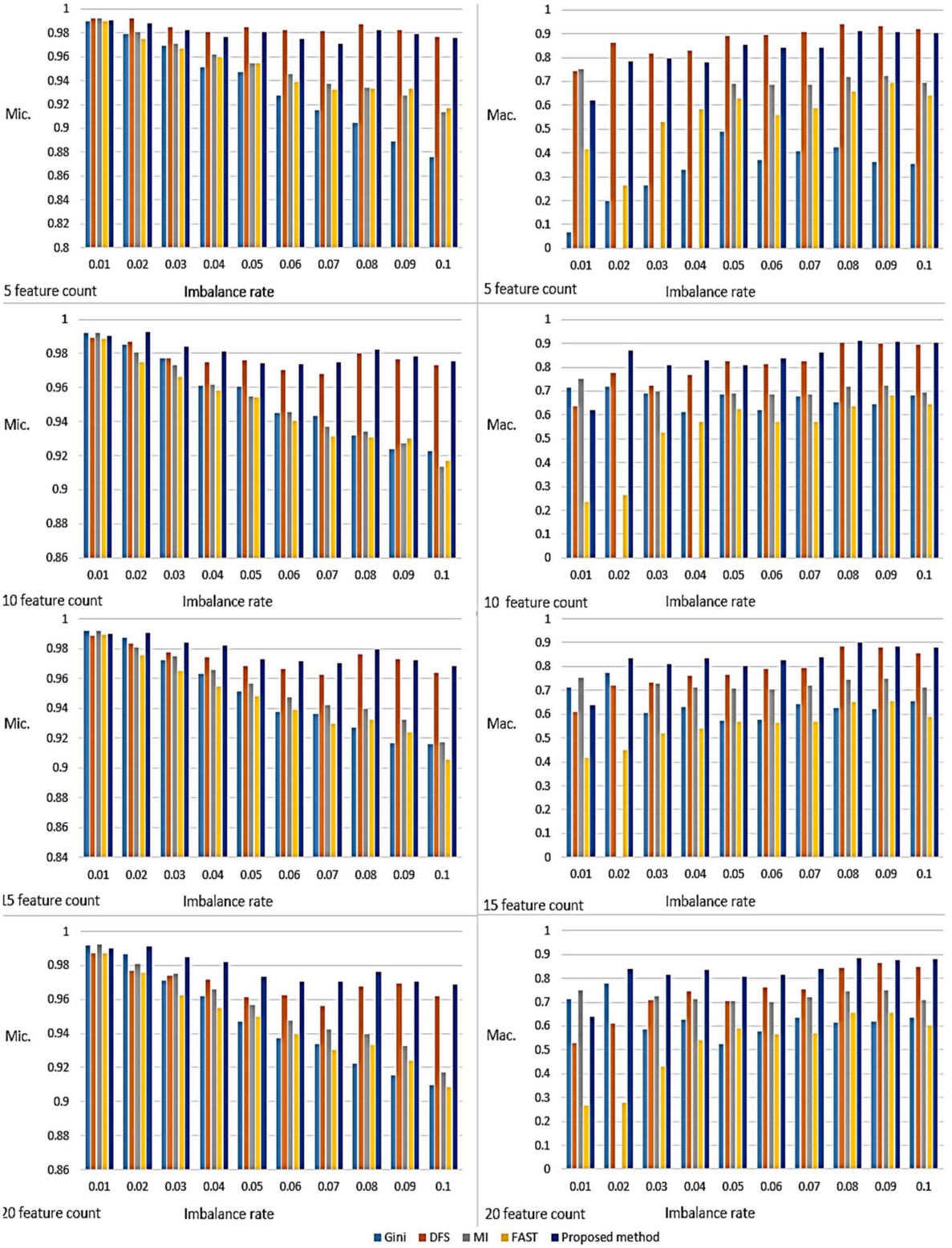

(شکل-۴): مقایسه کارایی دسته‌بند نایو بیز دسته‌های Ship و Acq با استفاده از روش‌های مختلف انتخاب ویژگی برحسب Micro F و Macro F

**(Figure-4): Comparison of the performances of Naïve Bayes classifier in Acq and Ship classes using various methods of feature selection based on Micro F and Macro F**




(شکل-۵): مقایسه کارایی دسته‌بند درخت تصمیم دسته‌های Corn و All با استفاده از روش‌های مختلف انتخاب ویژگی بر حسب Micro F و Macro F

(Figure-5): Comparison of the performances of C4.5 classifier in Corn and All classes using various methods of feature selection Micro and Macro F.




( شکل-۶): مقایسه کارایی دسته‌بند نایوبیز دسته‌های Corn و All با استفاده از روش‌های مختلف انتخاب ویژگی بر حسب Micro F و Macro F

**(Figure-6): Comparison of the performances of Naïve Bayes classifier in Corn and All classes using various methods of feature selection based on Micro F and Macro F**

(جدول-۹): نتیجه آزمون معنی‌داری روش پیشنهادی و روش‌های دیگر بر حسب Micro F

**(Table-9): Results of significance tests of the differences between PFS and FAST, MI, DFS, and Gini methods based on Micro F**

| Labels | Classifier | | Mean | Std. Deviation | t | Sig. |
|---|---|---|---|---|---|---|
| Project-student | C4.5 | Proposed method - Gini | .2215284 | .1420147 | 9.866 | .000 |
| | | Proposed method-DFS | .0392580 | .0405141 | 6.128 | .000 |
| | | Proposed method-MI | .5075978 | .2640894 | 12.156 | .000 |
| | | Proposed method – FAST | .2521250 | .0632796 | 25.199 | .000 |
| | Naïve Bayes | Proposed method - Gini | .0250788 | .0414836 | 3.823 | .000 |
| | | Proposed method - DFS | .0277771 | .0218393 | 8.044 | .000 |
| | | Proposed method – MI | .0338046 | .0257198 | 8.313 | .000 |
| | | Proposed method – FAST | .0279313 | .0135865 | 13.002 | .000 |
| Ship-Acq | C4.5 | Proposed method - Gini | .0308551 | .0201307 | 9.694 | .000 |
| | | Proposed method - DFS | .0092107 | .0088141 | 6.609 | .000 |
| | | Proposed method - MI | .0439852 | .0247475 | 11.241 | .000 |
| | | Proposed method – FAST | .0363538 | .0198723 | 11.570 | .000 |
| | Naïve Bayes | Proposed method – GINI | .0266215 | .0245525 | 6.858 | .000 |
| | | Proposed method DFS | .0095677 | .0094692 | 6.390 | .000 |
| | | Proposed method - MI | .0385077 | .0240526 | 10.125 | .000 |
| | | Proposed method – FAST | .0303108 | .0208659 | 9.187 | .000 |
| Corn-All | C4.5 | Proposed method - GINI | .0352383 | .0239180 | 9.318 | .000 |
| | | proposed method - DFS | .0020081 | .0029516 | 4.303 | .000 |
| | | Proposed method - MI | .0368421 | .0193858 | 12.020 | .000 |
| | | Proposed method - FAST | .0419387 | .0210802 | 12.583 | .000 |
| | Naïve Bayes | Proposed method - GINI | .0324684 | .0253699 | 8.094 | .000 |
| | | Proposed method - DFS | .0032087 | .0058829 | 3.450 | .001 |
| | | Proposed method - MI | .0258070 | .0181550 | 8.990 | .000 |
| | | Proposed method - FAST | .0312245 | .0173018 | 11.414 | .000 |

(جدول ۱۰-): نتیجه آزمون معنی‌داری روش پیشنهادی و روش‌های دیگر بر حسب Macro F

**(Table10): Results of significance tests of the differences between PFS and FAST, MI, DFS, and Gini methods based on Macro F**

| | | mac | | | | |
|---|---|---|---|---|---|---|
| Labels | Classifier | | Mean | Std. Deviation | t | Sig. |
| Project-student | C4.5 | Proposed method - Gini | .1505517 | .0893478 | 10.657 | .000 |
| | | Proposed method-DFS | .0393399 | .0643889 | 3.864 | .000 |
| | | Proposed method-MI | .1820306 | .2616866 | 4.399 | .000 |
| | | Proposed method – FAST | .1730649 | .0868316 | 12.606 | .000 |
| | Naïve Bayes | Proposed method - Gini | .1214185 | .1660269 | 4.625 | .000 |
| | | Proposed method - DFS | .0393753 | .0773087 | 3.221 | .003 |
| | | Proposed method - MI | .1769136 | .2600285 | 4.303 | .000 |
| | | Proposed method – FAST | .1461393 | .0719871 | 12.839 | .000 |
| Ship-Acq | C4.5 | Proposed method - Gini | .2215284 | .1420147 | 9.866 | .000 |
| | | Proposed method - DFS | .0392580 | .0405141 | 6.128 | .000 |
| | | Proposed method - MI | .5075978 | .2640894 | 12.156 | .000 |
| | | Proposed method – FAST | .2521250 | .0632796 | 25.199 | .000 |
| | Naïve Bayes | Proposed method – GINI | .2137205 | .1807172 | 7.480 | .000 |
| | | Proposed method – DFS | .0423581 | .0561289 | 4.773 | .000 |
| | | Proposed method - MI | .4747638 | .2529405 | 11.871 | .000 |
| | | Proposed method - FAST | .1861018 | .0689564 | 17.069 | .000 |
| Corn-All | C4.5 | Proposed method - GINI | .2881490 | .1574671 | 11.573 | .000 |
| | | proposed method - DFS | .0153647 | .0303661 | 3.200 | .003 |
| | | Proposed method - MI | .3414916 | .2528745 | 8.541 | .000 |
| | | Proposed method - FAST | .4513617 | .0854588 | 33.404 | .000 |
| | Naïve Bayes | Proposed method - GINI | .2597635 | .1692409 | 9.707 | .000 |
| | | Proposed method - DFS | .0279309 | .0645041 | 2.739 | .009 |
| | | Proposed method - MI | .2355308 | .2853513 | 5.220 | .000 |
| | | Proposed method - FAST | .2876514 | .0929295 | 19.577 | .000 |

به‌منظور بررسی معنی‌داربودن اختلاف نتایج به‌دست‌آمده از روش پیشنهادی و روش‌های دیگر از آزمون T Student ($p<0.05$) استفاده و فرض‌های زیر در نظر گرفته شد:

H0: اختلاف معنی‌داری بین نتایج روش پیشنهادی با روش جینی وجود ندارد.

H1: اختلاف معنی‌داری بین نتایج روش پیشنهادی و روش جینی وجود دارد.

با توجه به اینکه برای تعداد ویژگی‌های ۵، ۱۰، ۱۵ و بیست و برای ده نرخ مختلف آزمایش‌ها ثبت شدند، بنابراین چهل آزمایش به ازای هر دسته‌بند انجام شده است. این آزمون معناداربودن اختلاف، بین کارایی روش پیشنهادی و روش‌های دیگر نیز انجام شد. جدول (۹) برای برحسب Micro F و جدول (۱۰) برحسب Macro F نتایج به دست آمده از آزمون معنی‌داربودن را نشان می‌دهند.

با توجه به $p<0.05$ اختلاف عملکرد روش پیشنهادی و روش‌های دیگر تأیید شد.

یکی از معیارهایی که علاوه بر Micro F و Macro F برای ارزیابی دسته‌بندی مورد استفاده قرار می‌گیرد معیار G-mean است. این معیار با استفاده از رابطه (۱۰) محاسبه می‌شود.

$$G-mean=\sqrt{\frac{TP}{TP+FN}*\frac{TN}{TN+FP}} \qquad (۱۰)$$

عملکرد روش پیشنهادی در مقایسه با روش‌های دیگر برحسب معیار G-mean برای رده اقلیت در جداول(۱۱و۱۲) برای تعداد ویژگی‌های مختلف نشان داده شده است. برای هر تعداد ویژگی، نرخ عدم توازن از ده تا صددرصد تعداد اسناد رده اقلیت افزایش یافت و میانگین این ده حالت به‌عنوان نتیجه در جدول نشان داده شده است. نتایج نشان می‌دهند روش پیشنهادی عملکرد قابل قبولی داشته است.

نتایج به‌دست‌آمده از مقایسه عملکرد کارایی دسته‌بندها با استفاده از انتخاب ویژگی پیشنهادی و روش‌های دیگر انتخاب ویژگی که در نمودارهای تصاویر یک تا شش ارائه شد و همچنین نتایج حاصل از مقایسه برحسب G-mean نشان می‌دهد روش پیشنهادی کارایی دسته‌بندها را در مقایسه را روش‌های دیگر بهبود بخشیده است.

**( جدول-۱۱): مقایسه کارایی دسته‌بند C4.5 با روش‌های مختلف انتخاب ویژگی بر حسب G-mean**

**(Table-11): Comparison of the performances of C4.5 classifier using various methods of feature selection base on G-mean**

| Corpus | Class | Feature Count | Feature Selection Method | | | | |
|---|---|---|---|---|---|---|---|
| | | | Gini | DFS | MI | FAST | Proposed Method |
| WebKB | Project-Student | 5 | 0.269272 | 0.46446 | 0.11536 | 0.335906 | 0.538399 |
| | | 10 | 0.582396 | 0.514957 | 0.177042 | 0.456869 | 0.570265 |
| | | 15 | 0.578942 | 0.608538 | 0.18715 | 0.494007 | 0.577292 |
| | | 20 | 0.681157 | 0.694927 | 0.235472 | 0.488444 | 0.58801 |
| Reuters-21875 | Acq-Ship | 5 | 0.368152 | 0.774766 | 0.133931 | 0.498781 | 0.89463 |
| | | 10 | 0.757251 | 0.821866 | 0.135474 | 0.546826 | 0.895778 |
| | | 15 | 0.818887 | 0.901245 | 0.13551 | 0.539825 | 0.898167 |
| | | 20 | 0.823773 | 0.901395 | 0.137513 | 0.552348 | 0.89714 |
| | Corn-All | 5 | 0.198376 | 0.913599 | 0.185822 | 0.32889 | 0.913952 |
| | | 10 | 0.764238 | 0.91641 | 0.212548 | 0.353309 | 0.907477 |
| | | 15 | 0.77263 | 0.923646 | 0.302424 | 0.427082 | 0.905797 |
| | | 20 | 0.767943 | 0.918231 | 0.286274 | 0.453414 | 0.905324 |
| **میانگین کل** | | | **0.615251** | **0.779503** | **0.187043** | **0.456308** | **0.791019** |

**(جدول-۱۲): مقایسه کارایی دسته‌بند نایوبیز با روش‌های مختلف انتخاب ویژگی بر حسب G-mean**

**(Table-12): Comparison of the performances of Naïve Bayes classifier using various methods of feature selection base on G-mean**

| Corpus | Class | Feature Count | Feature Selection Method | | | | |
|---|---|---|---|---|---|---|---|
| | | | Gini | DFS | MI | FAST | Proposed Method |
| WebKB | Project-Student | 5 | 0.19624 | 0.455025 | 0.111981 | 0.33207 | 0.54292 |
| | | 10 | 0.513311 | 0.586378 | 0.168698 | 0.388171 | 0.575526 |
| | | 15 | 0.581756 | 0.639219 | 0.176356 | 0.429831 | 0.58557 |
| | | 20 | 0.685474 | 0.705742 | 0.234614 | 0.44955 | 0.599686 |
| Reuters-21875 | Acq-Ship | 5 | 0.217512 | 0.752024 | 0.134519 | 0.499436 | 0.858348 |
| | | 10 | 0.672796 | 0.845196 | 0.12565 | 0.477677 | 0.831231 |
| | | 15 | 0.70273 | 0.876603 | 0.12565 | 0.471988 | 0.822774 |
| | | 20 | 0.717439 | 0.851699 | 0.125752 | 0.502614 | 0.800845 |
| | Corn-All | 5 | 0.111019 | 0.810554 | 0.191299 | 0.4144 | 0.799606 |
| | | 10 | 0.459247 | 0.772823 | 0.212816 | 0.425024 | 0.798065 |
| | | 15 | 0.423376 | 0.74714 | 0.298143 | 0.406354 | 0.779827 |
| | | 20 | 0.402932 | 0.76326 | 0.31305 | 0.430677 | 0.800194 |
| **میانگین کل** | | | **0.473653** | **0.733805** | **0.184877** | **0.435649** | **0.732883** |

## ۵- بحث

نمودارهای نتایج آزمایش‌های انجام‌شده عملکرد روش‌های مختلف انتخاب ویژگی و روش پیشنهادی را در مقابل داده‌های متنی نامتوازن نشان می‌دهد. نتایج، برتری روش پیشنهادی را نسبت به روش‌های دیگر نشان می‌دهد.

بررسی نتایج آزمایش‌ها نشان می‌دهد که، انتخاب ویژگی درصورتی‌که روش مناسبی انتخاب شود، می‌تواند مشکل افت کارایی ناشی از عدم توازن داده‌های متنی را تا حدود زیادی حل کند. همچنین بررسی نمودارها نشان می‌دهد که هر چه شدت عدم توازن کاهش یافته است، کارایی به‌صورت کلی افزایش یافته است.

روش پیشنهادی پارامترهای بیشتری را نسبت به روش‌های دیگر در انتخاب ویژگی لحاظ کرده است، به همین دلیل از روش‌های دیگر موفق‌تر بوده است. نزدیک‌ترین روش به روش پیشنهادی روش DFS است که معیارهای مشابهی را نیز استفاده کرده و حتی در برخی موارد از روش پیشنهادی موفق‌تر عمل کرده ولی در حالت کلی روش پیشنهادی بهتر عمل کرده است.

## ۶- نتیجه‌گیری و کارهای آینده

در این پژوهش روش جدیدی برای انتخاب ویژگی دسته‌بندی داده‌های متنی نامتوازن با ابعاد بالا و همچنین روش جدیدی برای دسته‌بندی داده‌های متنی نامتوازن ارائه شد. انتخاب ویژگی پیشنهادی با استفاده از دسته‌بندهای درخت تصمیم C4.5 و نایوبیز ارزیابی و برای ارزیابی روش جدید از پیکره reuters-21875 و WebKB استفاده شد. نتایج ارزیابی روش پیشنهادی با استفاده از سنجه‌های Micro F و Macro F مقایسه و تعداد ویژگی‌های انتخاب‌شده ۵، ۱۰، ۱۵و ۲۰ در نظر گرفته شد. ارزیابی‌ها برای نرخ‌های مختلف نرخ عدم توازن تغییر داده شد. مقایسه نتایج به‌دست‌آمده از آزمایش‌های مختلف در نرخ‌های عدم توازن مختلف و داده‌های مختلف نشان می‌دهد روش پیشنهادی نسبت به روش‌های دیگر برتری دارد.

در روش انتخاب ویژگی پالایه پیشنهادی، وزن v1 و v2 در محاسبه امتیاز ویژگی، یکسان در نظر گرفته شدند؛ اما می‌توان برای هر پارامتر وزنی در نظر گرفت. بررسی تأثیر و نحوه محاسبه این وزن‌ها می‌تواند به‌عنوان موضوع پژوهش‌های آینده در نظر گرفته شود.

استفاده کامل‌تر و جامع‌تر از روابط هستان‌شناسی و خصوصیات زبانی دیگر در انتخاب ویژگی داده متنی نامتوازن به‌طور کامل بررسی نشده‌اند؛ بنابراین پژوهش در این زمینه به‌عنوان کارهای آینده ضرورت دارد. استفاده از ترکیب روش‌های مختلف انتخاب ویژگی می‌تواند به‌عنوان راه‌کار بهبود انتخاب ویژگی استفاده شود.

## ۷- مراجع 7- References


[1] He, H. ,and E.A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21(9),p p. 1263-1284, 2009.

[2] P.Yang, et al. , "Ensemble-based wrapper methods for feature," *springer,Advances in Knowledge Discovery and Data Mining*, vol. 7818, pp. 544-555,2013.

[3] M.Galar, et al., "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics*, Part C (Applications and Reviews), vol. 42(4), pp. 463-484,2012.

[4] N.V. Chawla, N. Japkowicz, and A. Kotcz, *Editorial: special issue on learning from imbalanced data sets.* SIGKDD Explor. Newsl., 2004. ch,6(1), pp. 1-6.

[5] J.V.Hulse, T.M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," *in Proceedings of the 24th international conference on Machine learning*, ACM: Corvalis, Oregon, USA, 2007. pp. 935-942.

[6] H. Ogura, , H. Amano, and M. Kondo, "Comparison of metrics for feature selection in imbalanced text classification," *Expert Systems with Applications*, vol. 38(5), pp. 4978-4989. 2011.

[7] S.Maldonadoa, R. Weberb, and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines," *National Research Council of Canada*, Ottawa, Canada Information Sciences, pp. 228–246, 2014.

[8] E.Chen, et al., "Exploiting probabilistic topic models to improve text categorization under class imbalance," *Information Processing & Management*, vol. 47(2), pp. 202-214, 2011.

[9] E.L. Iglesias, A. Seara Vieira, and L. Borrajo, "An HMM-based over-sampling technique to improve text classification," *Expert Systems with Applications*, vol. 40(18), pp. 7184-7192, 2013.

[10] R. Barandela , et al., "The imbalanced training sample problem: Under or over sampling?" in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, Springer, 2004.

[11] N.V.Chawla, et al., "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, pp. 321-357, 2002.

[12] S.Barua, et al., "MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol.26(2), pp. 405-425, 2014.

[13] A.Sun, E.-P. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: *A comparative stud*y," *Decision Support Systems*, vol, 48(1), pp. 191-201, 2009.

[14] C.Sanchez-Hernandez, , D.S. Boyd, and G.M. Foody, "One-class classification for mapping a specific land-cover class: SVDD classification of fenland," *IEEE Transactions on Geoscience and Remote Sensing*, vol.45(4), pp. 1061-1073, 2007.

[15] S.S. Khan, and M.G. Madden, "A survey of recent trends in one class classification," in Irish conference on Artificial Intelligence and Cognitive Science, Springer, 2009.

[16] K.M. Ting, "A comparative study of cost-sensitive boosting algorithms," in *Proceedings of the 17th International Conference on Machine Learning*, Citeseer, 2000.

[17]Cheng, F., et al., *Large cost-sensitive margin distribution machine for imbalanced data classification.* Neurocomputing, 2017. 224, pp. 45-57.

[18] X.-w. Chen, and M. Wasikowski, "FAST: a roc-based feature selection metric for small samples and imbalanced data classification problems," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM: Las Vegas, Nevada, USA, 2008, pp. 124-132.

[19] Y. Xu, "A Comparative Study on Feature Selection in Unbalance Text Classification," i*n Proceedings of the 2012 Fourth International Symposium on Information Science and Engineering*, IEEE Computer Society, 2012, p p. 44-47.

[20] T. Lei, and L. Huan, "Bias analysis in text classification for highly skewed data," in *Data Mining, Fifth IEEE International Conference on*. 2005.

[21] S. Chua, and N. Kulathuramaiyer, "Feature selection semantic based," *Springer Netherlands,Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering*, pp. 471-476, 2008.

[22] A. Khan, B. Baharudin, and K. Khan, "Efficient Feature Selection and Domain Relevance Term Weighting Method for Document Classification," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 2, pp. 398-403, 2010.

[۲۳]رضائی وحیده، محمدپور مجید، پروین حمید، نجاتیان صمد. ارائه روشی برای استخراج واژگان کلیدی و وزن‌دهی واژگان برای بهبود طبقه‌بندی متون فارسی. پردازش علائم و داده‌ها. ۱۳۹۶; ۱۴ (۴) :۷۸-۵۵

[23] r. V, et al., *An Approach for Extraction of Keywords and Weighting Words for Improvement Farsi Documents Classification.* JSDP, vol. 14(4), 2018,pp. 55-78.

[24] A.K.Uysal, and S. Gunal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based Systems*, vol.36, p p. 226-235, 2012.

[25] W.Shang, et al., "A novel feature selection algorithm for text categorization," *Expert Systems with Applications*, vol. 33(1), pp. 1-5, 2007.

[26] Z. Zheng, and R.S. X Wu, *Feature Selection for Text Categorization on Imbalanced Data*, ACM SIGKDD Explorations Newsletter, 2004 - dl.acm.org, 2004.

[27] A.Moayedikia, et al., "Feature selection for high dimensional imbalanced class data using harmony search," *Engineering Applications of Artificial Intelligence*, vol. 57, pp. 38-49, 2017.

[28] A.Rehman, K. Javed, and H.A. Babri,"Feature selection based on a normalized difference measure for text classification," *Information Processing & Management*, vol. 53(2), pp. 473-489, 2017.

[29] S.Kansheng, et al., "Efficient text classification method based on improved term reduction and term weighting," *The Journal of China Universities of Posts and Telecommunications,* vol.18, pp. 131–135, 2011.

[30] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of machine learning research*, vol-.3(Mar), pp. 1289-1305, 2003.

[31] G.S. Yanling, and Y. Zhu, "Data imbalance problem in text classification," *IEEE ,Third International Symposium on Information Processing*, 2010.

[32] P.Bermejo, et al., "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking," *Knowledge-Based Systems*, vol. 25(1), pp. 35-44, 2012.

[33] Z. Zhu, Y.-S. Ong, and M. Dash, "Wrapper–filter feature selection algorithm using a memetic

framework," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37(1), pp. 70-76, 2007.

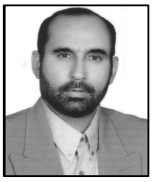
[34] L.Breiman, Friedman ,and O. J. H., R. A., et al., *Classification and regression trees. Montery* CA: Wadsworth International Group, 1984.

[35] S.Li, et al., "A framework of feature selection methods for text categorization", in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics*. Vol.2. 2009.

[36] M. Alibeigi, S. Hashemi, and A. Hamzeh, "DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets," *Data & Knowledge Engineering*, pp.81-82, pp. 67-10,2012.

[37] H. Jing, et al., "A General Framework of Feature Selection for Text Categorization, in Machine Learning and Data Mining in Pattern Recognition," *6th International Conference, MLDM 2009*, Leipzig, Germany, July 23-25, *2009. Proceedings*, P. Perner, Editor. 2009, Springer Berlin Heidelberg: Berlin, Heidelberg. pp. 647-662.

**جعفر پورامینی** مقاطع کارشناسی و کارشناسی ارشد خود را در رشته مهندسی نرم‌افزار دانشگاه صنعتی امیرکبیر به پایان رسانده است. همچنین مدرک دکترای مهندسی فناوری اطلاعات را از دانشگاه قم دریافت کرده است. حوزه پژوهشی و تخصصی وی داده‌کاوی و متن‌کاوی است.
نشانی رایانامه ایشان عبارت است از:
**j_pouramini@pnu.ac.ir**



**بهروز مینایی بیدگلی** دکترای خود را در رشته علوم و مهندسی کامپیوتر از دانشگاه ایالتی میشیگان آمریکا در سال ۱۳۸۴ گرفتند. تخصص ایشان هوش مصنوعی و داده‌کاوی است. ایشان هم اکنون به‌عنوان عضو هیئت علمی دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت به تدریس دروس هوش مصنوعی و نرم‌افزار مشغول هستند. وی سرپرستی گروه متن‌کاوی متون عربی و فارسی را در پژوهشکده داده‌کاوی نور نیز به‌عهده دارد.
نشانی رایانامه ایشان عبارت است از:
**b_minaei@iust.ac.ir**

**مهدی اسماعیلی** کارشناسی مهندسی نرم‌افزار را از دانشگاه اصفهان دریافت و کارشناسی ارشد را در رشته علوم کامپیوتر اخذ و دکترای خود را از دانشگاه دبرسن در رشته مهندسی کامپیوتر دریافت کردند. وی اکنون عضو هیات علمی دانشگاه آزاد واحد کاشان هستند. تخصص ایشان حوزه‌های مرتبط با داده‌کاوی است.
نشانی رایانامه ایشان عبارت است از:
**m.esmaeili@iaukashan.ac.ir**