بهبود رده‌بندی داده‌های نامتوازن با استفاده از معیارهای شباهت فازی و خوشه‌بندی کاهشی

سید احسان یثربی نایینی*[1] و مهلا حاتمی[2]
[1] گروه برق و کامپیوتر، دانشگاه تربت حیدریه
[2]دانشگاه شهید باهنر کرمان، کرمان، ایران

چکیده

یکی از قسمت‌های مهم در داده‌کاوی و کشف دانش از پایگاه داده، رده‌بندی است. در اغلب موارد داده‌هایی که برای آموزش رده‌بندها به کار می‌روند از توزیع مناسبی برخوردار نیستند. این توزیع نامناسب هنگامی رخ می‌دهد که یک رده تعداد نمونه‌های زیادی دارد؛ درحالی‌که به‌طور ذاتی نمونه‌های رده دیگر کم است. به‌طورکلی روش‌های حل این نوع مسائل به دو دسته نمونه‌گیری کاهشی و نمونه‌گیری افزایشی تقسیم می‌شود. در این مقاله یک روش نمونه‌گیری کاهشی با استفاده از ترکیب خوشه‌بندی و معیارهای شباهت فازی ارائه شده است و عملکرد آن‌ها ازنظر کارآمدی در رده‌بندی داده‌های نامتوازن مورد تحلیل و بررسی قرارگرفته‌اند. بدین منظور در ابتدا خوشه‌بندی کاهشی انجام شده و داده‌های ردهٔ اکثریت خوشه‌بندی، سپس با استفاده از معیارهای شباهت فازی نمونه‌های هر خوشه رتبه‌بندی و بر اساس این رتبه‌ها نمونه‌های مناسب انتخاب می‌شود؛ نمونه‌های انتخاب‌شده به همراه رده اقلیت مجموعه داده نهایی را تشکیل می‌دهند. در این پژوهش پیاده‌سازی در نرم‌افزار MATLAB، ارزیابی نتایج از طریق محاسبه معیار AUC و تحلیل نتایج با استفاده از آزمون‌های آماری استاندارد انجام شده است. نتایج مطالعه نشان‌دهنده عملکرد بهتر روش پیشنهادی، نسبت به سایر روش‌های شناخته شده است.

واژگان کلیدی: رده‌بندی داده‌های نامتوازن، معیارهای شباهت فازی، نمونه‌گیری، خوشه‌بندی کاهشی.

Improving Imbalanced Data Classification Accuracy by using Fuzzy Similarity Measure and Subtractive Clustering

Seyed Ehsan Yasrebi Naeini*[1] & Mahla Hatami[2]
[1]Computer and Electrical Dept., University of Torbat Heydarieh
[2]Shahid Bahonar University of Kerman, Kerman, Iran

Abstract

One of the biggest challenges in this field is classification problems which refers to the number of different samples in each class. If a data set includes two classes, imbalance distribution occurs when one class has a large number of samples while the other is represented by a small number of samples. In general, the methods of solving these problems are divided into two categories: under-sampling and over-sampling. In this research, it is focused on under-sampling and the advantages of this method will be analyzed by considering the efficiency of classifying imbalanced data and it's supposed to provide a method for sampling a majority data class by using subtractive clustering and fuzzy similarity measure. For this purpose, at first the subtractive clustering is conducted and the majority data class is clustered. Then, using fuzzy similarity measure, samples of each cluster will be ranked and appropriate samples are selected based on these rankings. The selected samples with the minority class create the final dataset. In this research, MATLAB software is used for implementation, the results are evaluated by using AUC criterion and analyzing the results has been performed by standard statistical tools. The experimental results show that the proposed method is superior to other methods of under-sampling.

Keywords: Imbalanced data; Fuzzy similarity measure, Under-sampling, Subtractive clustering

* **Corresponding author** * **نویسندهٔ عهده‌دار مکاتبات**

• تاریخ ارسال مقاله: ۱۳۹۸/۹/۱۱ • تاریخ پذیرش: ۱۳۹۹/۵/۲۸ • تاریخ انتشار: ۱۴۰۱/۷/۷ • نوع مطالعه: پژوهشی

[DOI: 10.52547/jsdp.19.2.27]

۱- مقدمه

رده‌بندی، عملی مهم در داده‌کاوی و کشف دانش از پایگاه داده است [1]. یکی از مسائل مهم در زمینه داده‌کاوی مسأله ردهٔ نامتوازن است [2]. مسألهٔ دادهٔ نامتوازن زمانی رخ می‌دهد که نمونه‌های یک یا چند ردهٔ ذاتاً نادرند و یا به‌سختی جمع‌آوری می‌شوند. مسألهٔ رده‌بندی نامتوازنِ دودویی مسأله‌ای است که در آن بین تعداد نمونه‌های دو ردهٔ تفاوت زیادی وجود دارد. ردهٔ اقلیت به رده‌ای گفته می‌شود که تعداد نمونه‌های آن کم‌تر است و درصورتی‌که نادرست رده‌بندی شود، منجر به هزینهٔ بیش‌تری می‌شود. رده‌ای که دارای تعداد نمونه‌های بیش‌تری است، ردهٔ اکثریت نامیده می‌شود [3]. اغلب الگوریتم‌های یادگیری ماشین فرض می‌کنند که تعداد نمونه‌های آموزشی در رده‌های متفاوت برابر هستند و بر این اساس، رده‌بندها را آموزش می‌دهند؛ بنابراین زمانی که این الگوریتم‌ها را به داده‌های نامتوازن اعمال می‌کنیم، رده‌بند آموزش داده‌شده اغلب از ردهٔ اکثریت منتج می‌شود که این موضوع به پیش‌بینی بسیار ضعیف از ردهٔ اقلیت منجر می‌شود، به این دلیل که آموزش ردهٔ اقلیت به‌درستی انجام نشده است. در اغلب موارد، کاربر تمایل بیش‌تری به پیش‌بینی نمونه‌های ردهٔ اقلیت دارد؛ بنابراین، کنترل و حل مسأله دادهٔ نامتوازن برای بهبود کارایی امری ضروری است [3].

مطالعات نشان می‌دهد که مجموعه‌داده‌های متوازن از کارایی بهتری نسبت به حالت نامتوازن برخوردارند [4]. روش‌های متعددی جهت حل مسأله داده‌های نامتوازن معرفی شده‌اند که در دو دسته زیر طبقه‌بندی می‌شوند:

رویکرد سطح داده: این رویکرد با نمونه‌گیری مجدد از فضای داده بدون تغییر در الگوریتم یادگیری باعث تغییر توزیع داده‌ها می‌شود و تلاش می‌کند در مرحله پیش‌پردازش تأثیرات ناشی از عدم توازن را برطرف کند [5-8].

رویکرد سطح الگوریتم: این رویکرد به الگوریتم‌های یادگیری رده‌بند کمک می‌کند تا فرآیند یادگیری را به سمت ردهٔ اقلیت سوق دهد [9 , 10]. این پژوهش با استفاده از خوشه‌بندی کاهشی و معیارهای شباهت فازی در مرحله پیش‌پردازش با رویکرد سطح داده به حل مسأله داده‌های نامتوازن دودویی پرداخته است.

در این مقاله ابتدا به‌اختصار راه‌کارهای برخورد با داده‌های نامتوازن و ساختار این روش‌ها در بخش ۲ معرفی‌شده است. سپس در بخش ۳ مفاهیم پایه‌ای روش پیشنهادی توضیح داده‌شده است. در بخش ۴ روش پیشنهادی و تشریح آن ارائه‌شده است. در ادامه و در بخش ۵ نتایج عددی حاصل از اعمال روش پیشنهادی بر روی مجموعه داده‌های نامتوازن بیان شده است. در انتها نتیجه‌گیری و پیشنهاد برای کارهای آینده در بخش ۶ بیان شده است.

۲- مروری بر ادبیات

اعمال یک گام پیش‌پردازش به‌منظور متوازن‌کردن توزیع رده، راه‌حلی مؤثر برای مسألهٔ مجموعه‌داده‌های نامتوازن می‌باشد. در این روش‌ها، نمونه‌های آموزشی به‌وسیله روش‌های متفاوت اصلاح می‌شوند و به رده‌بندها اجازه می‌دهند در فرم استاندارد خود اجرا شوند. اعمال نمونه‌گیری با هدف گسترش ردهٔ اقلیت یا کاهش ردهٔ اکثریت یا ترکیب هر دو به‌عنوان یک گام پیش‌پردازش به‌منظور متوازن‌کردن توزیع رده و تولید مجموعه‌داده‌ای متفاوت از مجموعه‌داده اصلی یکی از رایج‌ترین روش‌ها به‌حساب می‌آید [11]. در این گام ساختار مجموعه داده اصلی برای رسیدن به تعادل بهینه (به‌عنوان‌مثال ۵۰ ۵۰) تغییر می‌کند [12]. روش‌های نمونه‌گیری از داده، به دو دستهٔ اصلی تقسیم می‌شوند: روش‌های نمونه‌گیری کاهشی[1] و روش‌های نمونه‌گیری افزایشی[2].

۲-۱- نمونه‌گیری کاهشی

روش‌های کاهشی با حذف نمونه‌ها از فضای ردهٔ اکثریت به‌صورت (غیر هیوریستیک) یا به‌وسیله روش‌های مؤثر هیوریستیک مانند حذف نمونه‌های مرزی باعث ایجاد توازن در مجموعه‌داده می‌شود. ازجمله روش‌های شناخته‌شده در زمینه نمونه‌گیری کاهشی می‌توان به موارد زیر اشاره کرد:

- RUS (Random Under-Sampling): یک روش غیر هیوریستیک است. ساده‌ترین روش که نمونه‌های ردهٔ اکثریت را به‌صورت تصادفی حذف می‌کند این روش ممکن است، داده‌هایی را که مفید و برای فرآیند استنتاج مهم هستند حذف کند [13].
- CNN (Condensed nearest neighbor rule): برای پیداکردن یک زیرمجموعهٔ سازگار از نمونه‌ها استفاده می‌شود و این زیرمجموعهٔ سازگار به‌عنوان مجموعه‌دادهٔ اصلاح‌شده استفاده می‌شود [14].

[1] Under Sampling
[2] Over Sampling

[DOI: 10.52547/jsdp.19.2.27]

- TL (Tomek Link): روشی هیوریستیک است و هدف آن پیداکردن Tomek Link ها است. اگر به‌عنوان یک روش نمونه‌گیری کاهشی به کار رود. به کار روند، تنها نمونه‌هایی که متعلق به ردهٔ اکثریت هستند حذف می‌شوند [15].
- NCL (Neighborhood cleaning rule): برای هر نمونهٔ x در مجموعهٔ آموزشی، سه تا از نزدیک‌ترین همسایه‌های آن را انتخاب می‌کنیم. اگر x متعلق به ردهٔ اکثریت باشد و رده‌بندی ارائه‌شده توسط سه تا از نزدیک‌ترین همسایه آن در تضاد باکلاس اصلی x باشد، در این صورت نمونه x حذف می‌شود. اگر نمونه x متعلق به ردهٔ اقلیت باشد و دو تا از نزدیک‌ترین همسایه‌های آن ردهٔ x را به اشتباه پیش‌بینی کنند در این صورت نزدیک‌ترین همسایه‌هایی که متعلق به ردهٔ اکثریت هستند حذف خواهند شد [16].
- SBC (under-Sampling Based Clustering): ایدهٔ اصلی این روش این است که خوشه‌های متفاوت، مشخصه‌های متفاوتی دارند؛ بنابراین اگر در یک رده‌بندی تعداد نمونه‌های ردهٔ اقلیت از اکثریت بیشتر باشد، مشخصه‌های ردهٔ اکثریت نادیده گرفته می‌شود و بالعکس [17].
- OSS (One -Sided Selection): تکنیکی هیوریستیک است که از ترکیب Tomek Link با CNN به‌دست می‌آید [18].

منتقدین روش‌های نمونه‌گیری کاهشی اعتقاد دارند که حذف نمونه‌های ردهٔ اکثریت می‌تواند منجر به از بین رفتن برخی از نمونه‌های مهم می‌شود [11,19].

۲-۲- نمونه‌گیری افزایشی

روش‌های افزایشی با تولید نمونه‌های جدید و یا تکرار نمونه‌های موجود از داده اقلیت منجر به تولید نمونه‌های مصنوعی و برقراری توازن در مجموعه‌داده می‌شوند. تکرار نمونه‌ها به‌صورت تصادفی (غیر هیوریستیک) و یا به‌صورت (هیوریستیک) با افزایش نمونه‌هایی که در مرز ردهٔ اقلیت و اکثریت هستند، انجام می‌شود. تکنیک‌ها SMOTE [7]، Borderline-SMOTE [20]، ADASYN [21]، AHC [22]، ADOMS [23] و SPIDER [24] ازجمله این روش‌های نمونه‌گیری افزایشی هستند.

در این مطالعه یک روش جدید با ترکیب خوشه‌بندی کاهشی و معیارهای شباهت فازی برای حذف نمونه‌های ردهٔ اکثریت معرفی می‌شود.

۳- پیش‌زمینه

هدف نهایی یک سامانهٔ یادگیری ماشین، رسیدن به بالاترین نرخ رده‌بندی ممکن برای مسأله موردنظر است. ازآنجایی‌که هیچ الگوریتم رده‌بندی وجود ندارد که به‌تنهایی به‌طور کامل برای تمام مسائل مناسب باشد، در این مطالعه از درخت تصمیم C4.5 [25] استفاده شده است.

۳-۱- متریک

به‌منظور ارزیابی رده‌بندهای باینری ارائه‌شده روش‌های متعددی بر اساس ماتریس آشفتگی مورداستفاده قرار می‌گیرد [26]. در این مطالعه برای بررسی عملکرد روش پیشنهادی از شاخص AUC استفاده شده است. این شاخص سطح زیرمنحنی گرافیکی نمودار ROC است که نرخ تشخیصات درست در مقابل نادرست را نشان می‌دهد. شاخص AUC از رابطه (۴) به‌دست می‌آید [26].

$$AUC = \frac{1+TP-FP}{2} \quad (۴)$$

پارامترهای True Positive(TP)، False Positive(FP)، به‌ترتیب عبارت‌اند از "نمونه‌های مثبت که به‌عنوان مثبت رده‌بندی شده‌اند" و "نمونه‌های منفی که به‌عنوان مثبت رده‌بندی شده‌اند".

۳-۲- خوشه‌بندی کاهشی (Subtractive clustering)

الگوریتم خوشه‌بندی فازی یکی از مهم‌ترین روش‌های به کار رفته در شناسایی الگوی بدون سرپرست است [27, 28]. به‌طورمعمول الگوریتم‌های خوشه‌بندی نیاز به دانستن تعداد خوشه‌ها از قبل هستند [29]. الگوریتم Fuzzy C-means نمونه شناخته‌شده‌ای از الگوریتم‌ها در این زمینه است [30]. یاگر و فیلو یک الگوریتم ساده و مؤثر برای برآورد تعداد خوشه‌ها و مراکز اولیه آن‌ها به نام کوهستان ارائه کردند [31]. جیو [32] یک فرم تغییریافته از الگوریتم کوهستان را به نام خوشه‌بندی کاهشی ارائه داد. در هنگامی‌که اطلاعات دقیقی از تعداد

[DOI: 10.52547/jsdp.19.2.27]

خوشه‌ها وجود ندارد می‌توان از خوشه‌بندی کاهشی استفاده کرد [33, 34].

هر نقطه به‌عنوان یک پتانسیل برای مرکز خوشه در نظر گرفته می‌شود. پتانسیل هر نقطه به‌صورت زیر محاسبه می‌شود:

$$P_i = \sum_{j=1}^{n} e^{-\frac{4\|x_i - x_j\|^2}{r_a^2}} \quad (۱)$$

که در این رابطه نماد $\|.\|$ نشان‌دهنده فاصله اقلیدسی و r_a مقداری مثبت است. مقدار ثابت r_a به‌عنوان شعاع همسایگی برای مرکز خوشه در نظر گرفته می‌شود. بنابراین پتانسیل تخصیص‌یافته به هر نقطه تابعی از فاصله با نقاط دیگر است. بعد از محاسبه پتانسیل برای هر نقطه، نقطه‌ای با بیشترین پتانسیل به‌عنوان مرکز خوشه نخست در نظر گرفته می‌شود. فرض کنید x_1^* مرکز خوشه نخست و P_1^* پتانسیل آن تعیین‌شده باشد، سپس پتانسیل هر نقطه x_i به‌صورت زیر اصلاح می‌شود:

$$P_i = P_i - P_1^* e^{\frac{4\|x_i - x_1\|^2}{r_b^2}} \quad (۲)$$

که r_b عددی مثبت است که همسایه‌ای را با کاهش قابل‌توجهی در همسایگی تعیین می‌کند. هنگامی‌که پتانسیل تمام نقاط اصلاح شد، نقطه‌ای با بیشترین پتانسیل به‌عنوان مرکز خوشه دوم در نظر گرفته می‌شود. این روند تا به‌دست‌آوردن تعداد کافی از خوشه‌ها ادامه می‌یابد [29, 35].

۳-۳- معیارهای شباهت فازی

معیارهای شباهت فازی را می‌توان به‌طورکلی به دو دسته زیر تقسیم‌بندی کرد. ۱) معیار مبتنی بر متریک، ۲) معیار مبتنی بر تئوری- مجموعه [36].

معیارهای شباهت مبتنی بر فاصله برای مجموعه‌های فازی: یکی از آشکارترین روش‌ها برای محاسبه شباهت بین دو مجموعه فازی، به‌دست‌آوردن فاصله هست. بر این اساس که هرچه دو مفهوم یا حقیقت به یکدیگر نزدیک‌تر باشند، از فاصله کمتری نسبت به هم برخوردار هستند. محاسبه این شباهت در دو مرحله انجام می‌گیرد. در مرحله نخست فاصله بین مجموعه فازی با استفاده از معیارهای فاصله مختلف نظیر فاصله همینگ یا فاصله اقلیدسی به‌دست آمده و در مرحله دوم یکی از روابط بین شباهت و فاصله به‌کاربرده شده تا درجه شباهت بین دو مفهوم به‌دست آید [36].

روش‌های زیادی برای بیان رابطه بین دو مفهوم شباهت و فاصله در قالب تابع وجود دارد که در زیر به چند مورد اشاره شده است. در تمام موارد اشاره‌شده، SM معیار شباهت و DM فاصله اقلیدسی بین دو مجموعه فازی A و B هستند. یکی از توابع که توسط Koczy ارائه شده به‌صورت زیر است [37].

$$SM(A,B) = \frac{1}{1+DM(A,B)} \quad (۳)$$

تابع دیگر که توسط Williams و Steele بیان‌شده در زیر نشان داده شده است [38].

$$SM(A,B) = e^{-\alpha.DM(A,B)} \quad (۴)$$

در این رابطه منظور از α اندازه شیب می‌باشد.

روش دیگر تخمین شباهت بین دو مجموعه فازی روشی است که توسط Sanitini پیشنهادشده و در رابطه (۵) این تخمین نشان داده‌شده است [39].

$$SM(A,B) = 1 - DM_r(A,B), \quad r=1, 2, \ldots, \infty \quad (۵)$$

معیارهای مبتنی بر تئوری مجموعه: اگر $\cup$ و $\cap$ به‌صورت max و min، $t-norm$ مدل بندی شود و $\square$ به‌صورت زیر فرض شود.

$$A \square B(x) = max\left[min\left(A(x), 1 - B(x)\right), min\left(B(x), 1 - A(x)\right)\right] \quad (۶)$$

یکی دیگر از معیارهای شباهت فازی مبتنی بر تئوری مجموعه مدل Rastle است که به‌صورت زیر نمایش داده می‌شود [40].

$$SM(A,B) = 1 - |A \square B| \quad (۷)$$

یکی دیگر از معیارهای شباهت فازی مبتنی بر تئوری مجموعه که به مدل Gregson معروف است به‌صورت زیر ارائه شده است [40].

$$SM(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (۸)$$

مدل Enta یکی دیگر از معیارهای شباهت مبتنی بر تئوری مجموعه بوده که به‌صورت زیر بیان می‌شود [40].

$$SM(A,B) = \sup_{x \in X} A \cap B(x) \quad (۹)$$

در این رابطه X مجموعه اعداد فازی است.

۴- روش پیشنهادی

در روش پیشنهادی ابتدا تمامی نمونه‌های اکثریت به تعدادی خوشه تقسیم می‌شوند. سپس با استفاده از معیارهای شباهت فازی به هریک از نمونه‌ها رتبه‌ای

[DOI: 10.52547/jsdp.19.2.27]

اختصاص داده‌شده است. هدف از خوشه‌بندی انتخاب تعداد مناسبی از نمونه‌های اکثریت است. پس از اعمال خوشه‌بندی با استفاده از معیارهای شباهت فازی به هریک از نمونه‌ها با توجه به میزان تعلق آن‌ها به هر خوشه، رتبهٔ اختصاص داده می‌شود از بین نمونه‌های رتبه‌بندی شده نمونه‌هایی با کمترین شباهت (نمونه‌هایی با بیش‌ترین فاصله) انتخاب‌شده‌اند.

گفتنی است ازآنجاکه اطلاع دقیقی از تعداد خوشه‌ها وجود ندارد از الگوریتم خوشه‌بندی کاهشی استفاده شده است.

در روش پیشنهادی ابتدا تمامی نمونه‌های اکثریت از مجموعه‌داده‌های اصلی جدا شده و سپس الگوریتم خوشه‌بندی کاهشی روی این نمونه‌ها اجرا می‌شود. بدین‌وسیله تعداد خوشه‌ها و مراکز آن‌ها به‌دست می‌آیند. پس از به‌دست‌آوردن تعداد خوشه‌ها و مراکز آن‌ها با استفاده از هر یک از معیارهای شباهت فازی مختلف بیان، شباهت هر نمونه به مراکز خوشه سنجیده و با استفاده از فرمول زیر میزان تعلق نمونه‌های اکثریت به هر خوشه محاسبه می‌شود:

$$d_{ik} = 1 \Big/ \sum_{j=1}^{c} \left(\frac{S_{ik}}{S_{jk}} \right) \qquad (۱۰)$$

در رابطه (۱۰)، c تعداد خوشه‌ها است و S_{ik} بیان‌گر میزان شباهت فازی بین نمونهٔ c_i که (مرکز خوشهٔ iام) و x_k است. ماتریس d میزان تعلق نمونه‌های اکثریت به هر خوشه را نشان می‌دهد. در این ماتریس از معیارهای شباهت فازی مختلف به‌منظور تعیین درجه تعلق استفاده شده است. نمونهٔ x_i در خوشه‌ای قرار می‌گیرد که به آن بیش‌ترین میزان تعلق را دارد. پس از به‌دست‌آوردن اعضای هر خوشه، رتبهٔ هر نمونه دوباره با استفاده از معیارهای شباهت فازی محاسبه و با توجه به‌اندازهٔ خوشه، تعدادی نمونه با بالاترین رتبه از هر خوشه انتخاب شده است. عملیات را بر روی تمامی خوشه‌ها اعمال می‌کنیم و رتبهٔ بر اساس معیارهای شباهت فازی برای تمامی نمونه‌های درون هر خوشه به‌صورت مجزا محاسبه شده است. رتبهٔ نمونه x_i در خوشهٔ c_l ($c = 1, \dots, l$) را با استفاده از رابطهٔ زیر محاسبه می‌کنیم:

$$rank_i = \sum_{j=1}^{n_l} 1 \Big/ S(x_i, x_j) \Big/ (R(x_i) \times n_l) \qquad (۱۱)$$

$rank_i$ رتبهٔ نمونهٔ x_i را نشان می‌دهد و x_i و x_j نمونه‌هایی در خوشهٔ c_l هستند. n_l بیان‌گر تعداد نمونه‌های موجود در این خوشه است، $S(x_i, x_j)$ بیان گر میزان شباهت بین نمونه x_i و x_j است و $R(x_i)$ با استفاده از رابطهٔ زیر به‌دست می‌آید:

$$R(x_i) = np \Big/ \sum_{p=1}^{np} 1 \Big/ S(x_i, x_p) \qquad (۱۲)$$

np تعداد نمونه‌های اقلیت در مجموعه آموزشی است. درنهایت از هر خوشه با توجه به‌اندازهٔ آن، تعدادی نمونه با رتبهٔ بالاتر انتخاب می‌شود:

$$n_i = 1 \Big/ IR \times NC_i \qquad ,i = 1, \dots, c \qquad (۱۳)$$

که n_i و NC_i به‌ترتیب تعداد نمونه‌های اکثریتی که باید از خوشهٔ iام انتخاب شود و تعداد کل نمونه‌های اکثریت موجود در این خوشه هستند. IR بیان‌گر نرخ عدم توازن می‌باشد. رابطه (۱۱) بیان می‌کند که نمونه‌های موجود در هر خوشه که شباهت کمتری با نمونه‌های اقلیت دارند و همچنین فاصلهٔ بیش‌تری با دیگر نمونه‌های موجود در خوشه دارند، رتبه بالاتری را به خود اختصاص می‌دهند. روند اجرای این روش در شکل (۱) نمایش داده شده است.

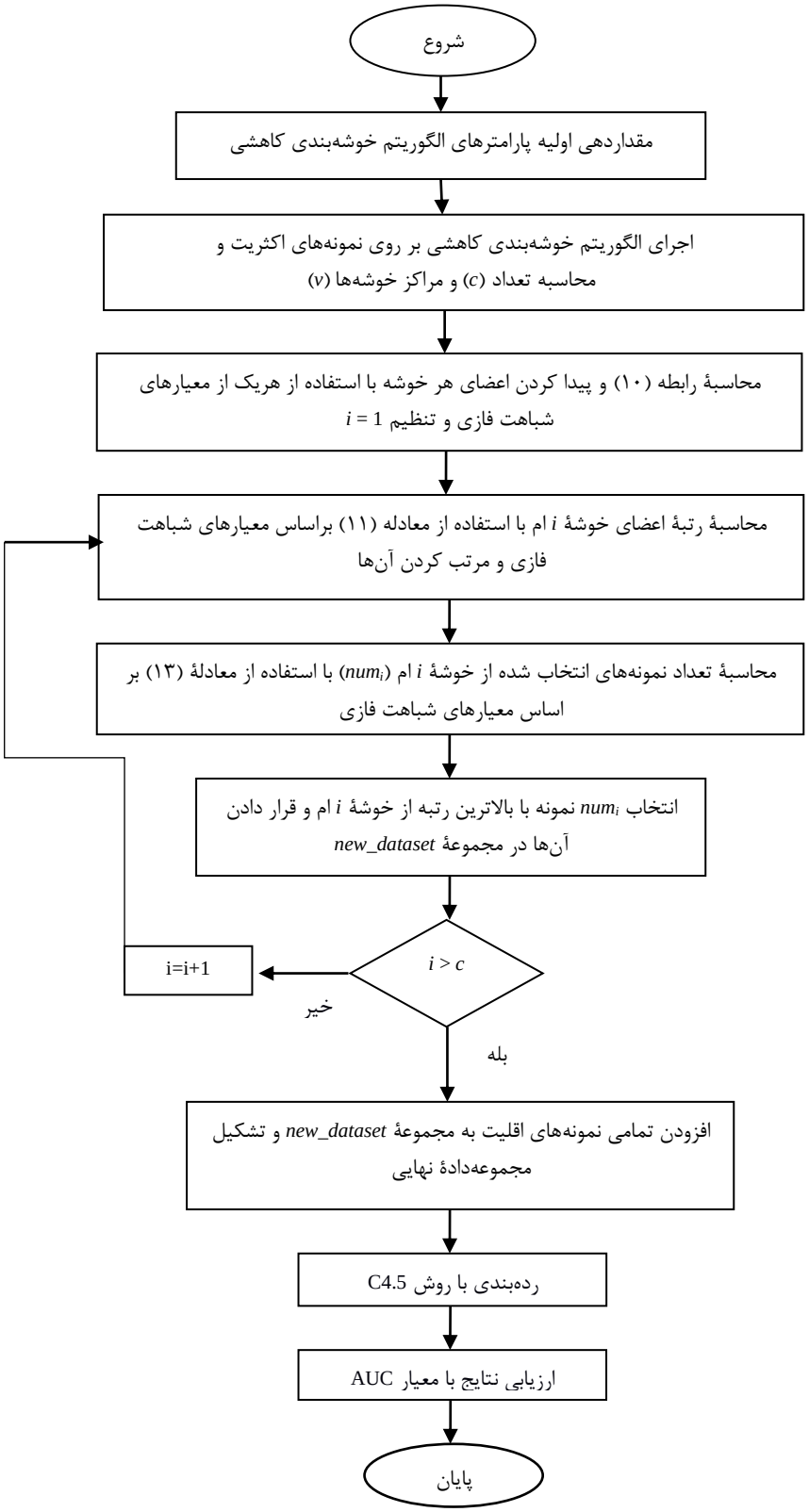
گفتنی است ازآنجاکه به‌منظور به‌دست‌آوردن شباهت از معیارهای شباهت فازی استفاده شده است و معیارهای شباهت فازی روی مجموعه‌های فازی کار می‌کنند، در این پژوهش به‌منظور به‌دست‌آوردن مجموعه‌های فازی در ابتدا داده‌ها بین بازه صفر و یک نرمال شده‌اند.

۵-نتایج و بحث

در این قسمت، نتایج حاصل از اعمال روش پیشنهادیِ بیان‌شده و به‌منظور بررسی دقیق‌تر کارایی روش پیشنهادی عملکرد آن روی هریک از معیارهای شباهت فازی نیز مورد تحلیل قرار گرفته است. برای دستیابی به این هدف ابتدا مجموعه داده‌های نامتوازن استفاده شده بیان می‌شود و سپس میزان دقت و کارایی هریک از معیارهای شباهت فازی با سایر الگوریتم‌های شناخته‌شده در زمینه under-sampling از جمله RSS، CNN، TL، NCL، SBC و OSS با استفاده از معیار AUC بررسی می‌شود. در این پژوهش از مجموعه‌دادگان نامتوازن مفروض در نرم‌افزار keel استفاده شده است [40]. مشخصات تمامیِ مجموعه‌داده‌هایی که در ساخت مدل‌های پیشنهادی استفاده شده‌اند در جدول (۱)

[DOI: 10.52547/jsdp.19.2.27]

فهرست شده است. در ستون نخست این جدول، نام مجموعه‌داده‌ها قرار دارد و در ستون‌ها دوم تا ششم به‌ترتیب تعداد نمونه‌ها، تعداد ویژگی‌ها، برچسب ردهٔ اقلیت و اکثریت، توزیع رده (برحسب درصد) و نسبت تعداد نمونه‌های ردهٔ اکثریت به تعداد نمونه‌های ردهٔ اقلیت (نرخ عدم توازن) نشان داده شده است.



(شکل-۱): الگوریتم روش پیشنهادی

(Figure-1): Algorithm of the proposed method

[DOI: 10.52547/jsdp.19.2.27]

(جدول-۱): مجموعه‌داده‌های مورداستفاده در مقاله

(Table-1): Datasets used in the article

نام مجموعه داده	تعداد نمونه	تعداد ویژگی	کلاس	توزیع کلاس ٪	نرخ عدم‌توازن
Ecoli0vs1	۲۲۰	۷	(im, cp)	(۳۵/۰۰,۶۵/۰۰)	۱/۸۶
Iris0	۱۵۰	۴	(Iris-Setosa, reminder)	(۳۳/۳۳,۶۶/۶۷)	۲/۰۰
Vehicle2	۸۴۶	۱۸	(bus, reminder)	(۲۸/۳۷,۷۱/۶۳)	۲/۵۲
Vehicle0	۸۴۶	۱۸	(van, reminder)	(۲۳/۶۴,۷۶/۳۶)	۳/۲۳
Ecoli1	۳۳۶	۷	(im, reminder)	(۲۲/۹۲,۷۷/۰۸)	۳/۳۶
New-thyroid2	۲۱۵	۵	(hypo, reminder)	(۱۶/۸۹,۸۳/۱۱)	۴/۹۲
New-thyroid1	۲۱۵	۵	(hyper, reminder)	(۱۶/۷۲,۸۳/۲۸)	۵/۱۴
Segment0	۲۳۰۸	۱۹	(brickface, reminder)	(۱۴/۲۶,۸۵/۷۴)	۶/۰۱
Glass6	۲۱۴	۹	(headlamps, reminder)	(۱۳/۵۵,۸۴/۴۵)	۶/۳۸
Ecoli3	۳۳۶	۷	(iMU, reminder)	(۱۰/۸۸,۸۹/۱۲)	۸/۱۹
Yeast2vs4	۵۱۴	۸	(cyt;me2)	(۹/۹۲,۹۰/۰۸)	۹/۰۸
Glass016vs2	۱۹۲	۹	(ve-win-float-proc;build-win-float-proc,build-win-non-float-proc,headlamps)	(۸/۸۹,۹۱/۱۱)	۱۰/۲۹
Yeast1vs7	۴۵۹	۸	(nuc,vac)	(۶/۷۲,۹۳/۲۸)	۱۳/۸۷
Shuttle2vs4	۱۲۹	۹	(Fpv Open; Bypass)	(۴/۶۵,۹۵/۳۵)	۲۰/۵۰
Yeast1458vs7	۶۹۳	۸	(vac; nuc, me2, me3, pox)	(۴/۳۳,۹۵/۶۷)	۲۲/۱۰
Yeast1289vs7	۹۴۷	۸	(vac;nuc,cyt,pox,erl)	(۳/۱۷,۹۶/۸۳)	۳۰/۵۶
Yeast5	۱۴۸۴	۸	(ME1, reminder)	(۲/۹۶,۹۷/۰۴)	۳۲/۷۸
Ecoli0137vs26	۲۸۱	۷	(pp, imL; cp, im, imU, imS)	(۲/۴۹,۹۷/۵۱)	۳۹/۱۵
Yeast6	۱۴۸۴	۸	(exc; reminder)	(۲/۴۹,۹۷/۵۱)	۳۹/۱۵

در تمامی روش‌ها از رده‌بند C4.5 به‌عنوان الگوریتم رده‌بند پایه استفاده شده است. جدول (۲) میانگین نتایج AUC بر روی مجموعه‌داده‌های آزمایش را با استفاده از معیارهای فازی مختلف نشان می‌دهد. در این جدول، بهترین نتایج در هر مجموعهٔ داده به‌صورت پررنگ مشخص شده است. نتایج نشان می‌دهد که روش پیشنهادی در هر یک از معیارهای شباهت فازی بهترین میانگین را در بیش‌تر مجموعه‌داده‌ها به‌دست می‌آورد.

(جدول-۲): نتایج معیارهای شباهت فازی

(Table-2): Results of fuzzy similarity measures

مجموعه داده	OSS	SBC	NCL	TL	CNN	RUS	معیار $\frac{1}{1+DM(A,B)}$	معیار شباهت فازی $1-[DM]_r(A,B)$	معیار شباهت فازی $e^{-\alpha DM(A,B)}$	معیار شباهت فازی $sup_{x\in X}\, A \cap B(X)$	معیار شباهت فازی $\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$	معیار شباهت فازی $1-\lvert A \Delta B\rvert$
Ecoli0vs1	۰/۹۶	۰/۵۰	۰/۹۵	۰/۹۷	۰/۹۴	۰/۹۷	**۱**	۰/۹۸	**۰/۹۸**	**۰/۹۸**	**۰/۹۷**	**۰/۹۸**
Iris0	۰/۹۷	۰/۵۰	**۰/۹۹**	**۰/۹۹**	۰/۹۷	۰/۹۹	**۰/۹۹**	۰/۹۹	**۰/۹۹**	**۰/۹۹**	**۰/۹۹**	**۰/۹۹**
Vehicle 1	**۰/۷۵**	۰/۷۰	۰/۷۳	۰/۷۰	۰/۶۶	۰/۷۰	۰/۶۵	۰/۷۳	۰/۷۱	**۰/۷۶**	۰/۷۳	**۰/۷۲**
Vehicle2	**۰/۹۴**	۰/۹۳	**۰/۹۴**	**۰/۹۴**	۰/۹۳	۰/۹۳	**۰/۹۵**	**./۹۴**	۰/۹۳	۰/۹۲	**۰/۹۴**	**۰/۹۴**
Ecoli1	۰/۸۸	۰/۵۷	۰/۸۷	۰/۸۹	۰/۸۷	۰/۸۸	۰/۹۰	**۰/۹۰**	**۰/۹۱**	**۰/۹۲**	**۰/۹۰**	**۰/۸۹**
New-thyroid2	۰/۹۳	۰/۵۰	۰/۹۴	۰/۹۳	۰/۹۵	۰/۹۰	**۰/۹۸**	۰/۹۴	**۰/۹۸**	**۰/۹۸**	**۰/۹۸**	**۰/۹۸**
New-thyroid1	۰/۹۴	۰/۵۰	۰/۹۴	۰/۹۲	۰/۹۳	۰/۹۱	**۰/۹۷**	**./۹۵**	**۰/۹۷**	**۰/۹۷**	**۰/۹۷**	**۰/۹۷**
Segment0	**۰/۹۸**	۰/۵۰	**۰/۹۸**	**۰/۹۸**	۰/۹۷	۰/۹۷	**۰/۹۸**	**۰/۹۸**	**۰/۹۸**	۰/۹۷	۰/۹۷	**۰/۹۸**

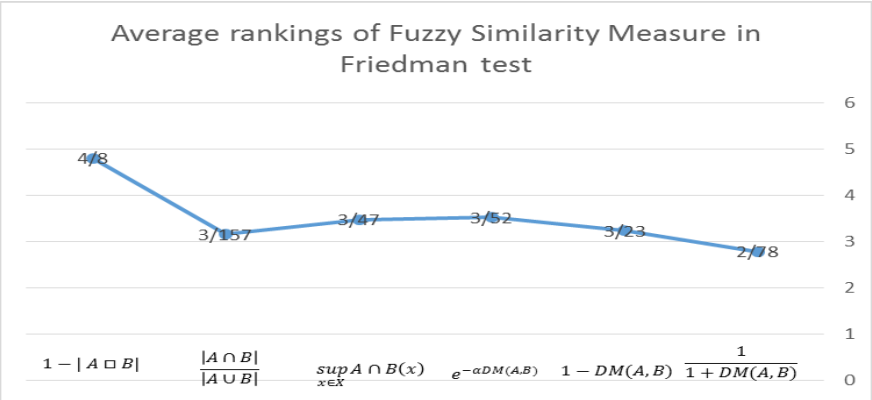
[DOI: 10.52547/jsdp.19.2.27]

Glass6	۰/۸۱	۰/۵۰	۰/۸۹	۰/۸۶	۰/۸۸	۰/۸۸	۰/۹۲	**۰/۹۴**	**۰/۹۳**	**۰/۹۸**	**۰/۹۲**	۰/۸۶
Ecoli3	۰/۷۵	۰/۵۰	۰/۸۲	۰/۸۴	۰/۷۰	۰/۸۴	۰/۸۶	**۰/۸۶**	**۰/۸۵**	**۰/۸۵**	**۰/۸۸**	**۰/۸۳**
Yeat2vs4	۰/۸۹	۰/۵۰	۰/۸۶	۰/۸۴	۰/۸۱	۰/۸۸	۰/۹۶	**۰/۹۴**	**۰/۹۲**	**۰/۹۴**	**۰/۹۷**	**۰/۸۹**
Glass016vs2	۰/۵۲	۰/۵۰	۰/۶۲	۰/۶۰	۰/۶۴	**۰/۶۵**	۰/۸۴	**۰/۹۴**	۰/۶۲	**۰/۸۷**	**۰/۸۴**	۰/۵۷
Yeast1vs7	۰/۶۳	۰/۵۰	۰/۶۲	۰/۵۸	۰/۶۴	۰/۶۵	۱	**۰/۸۹**	۱	**۰/۷۷**	۱	**۰/۷۷**
Shuttle2vs4	۱	۰/۵۰	۱	۱	۱	۰/۹۸	۰/۹۰	۱	۰/۹۰	۱	۱	۰/۷۰
Yeast1458vs7	۰/۵۰	۰/۵۰	۰/۴۹	۰/۵۰	۰/۴۷	۰/۵۷	۱	**۰/۹۲**	۱	**۰/۶۹**	۱	**۰/۶۹**
Yeast1289vs7	۰/۶۱	۰/۵۰	۰/۵۲	۰/۵۴	۰/۵۹	۰/۶۰	۱	۱	**۰/۹۸**	**۰/۹۶**	**۰,۹۴**	**۰/۷۱**
Yeast5	۰/۹۰	۰/۵۰	۰/۹۰	۰/۸۹	۰/۸۶	۰/۹۳	۰/۹۶	**۰/۹۴**	**۰/۹۵**	**۰/۹۵**	**۰/۹۸**	۰/۸۸
Ecoli0137vs26	۰/۵۴	۰/۵۰	**۰/۸۴**	**۰/۸۴**	۰/۷۱	۰/۷۸	۰/۷۰	۰,۷۶	۰/۵۶	۰/۵۳	۰/۷۰	۰/۶۰
Yeast6	۰/۷۷	۰/۵۰	۰/۷۷	۰/۷۹	۰/۵۵	۰/۸۱	**۰/۹۸**	**۰/۹۷**	**۰/۹۸**	**۰/۹۸**	**۰/۹۵**	۰/۷۵
میانگین	۰/۸۰	۰/۵۴	۰/۸۲	۰/۸۲	۰/۷۹	۰/۸۳	**۰/۹۲**	**۰/۹۲**	۰/۹۰	۰/۸۹	**۰/۹۲**	۰/۸۲

در جدول (۲) معیارهای شباهت فازی مختلف با یکدیگر مقایسه می‌شوند تا مشخص شود که کدام‌یک از معیارهای شباهت فازی نسبت به سایر معیارها از عملکرد بهتری در ترکیب با خوشه‌بندی کاهشی برخوردار است. نتایج نشان می‌دهد که روش پیشنهادی به‌همراه معیار شباهت‌های فازی Kozcy، Sanitini و Gregson عملکرد بهتری دارد و بهترین میانگین را در بیش‌تر مجموعه‌داده‌ها به‌دست می‌آورد. گرچه ارزیابی، تنها بر اساس مشاهده مشخص نمی‌کند که آیا میان روش‌های مختلف، تفاوت چشم‌گیری وجود دارد یا خیر؛ به همین منظور از آزمون‌های آماری برای اطمینان از وجود تفاوتی معنی‌دار میان روش‌ها استفاده می‌شود [41, 42]. منظور از تفاوت معنی‌دار میان روش‌ها این است که اختلاف میان روش‌ها به حد کافی بزرگ باشد تا بتوان اطمینان حاصل کرد که نتایج به‌طور تصادفی یا براثر شانس، بهتر نشده‌اند.

برای ارزیابی کارایی روش پیشنهادی از آزمون فریدمن استفاده شده است که مبتنی بر مقایسهٔ چندین روش است. این آزمون، هر الگوریتم را به‌طور جداگانه برای هر مجموعه‌داده رتبه‌بندی می‌کند و روشی که بهترین عملکرد را داشته باشد، دارای کمترین رتبه و بدترین روش، بیشترین رتبه را به خود اختصاص می‌دهد. در ازمون فریدمن فرضیه صفر بیان می‌کند که تمامی الگوریتم‌ها یکسان هستند؛ درحالی‌که رد این فرضیه وجود تفاوت میان الگوریتم‌های مورد بررسی را نشان می‌دهد [43].

شکل (۲) رتبه‌های معیارهای شباهت فازی مختلف در آزمون فریدمن را نشان می‌دهد. آشکار است که میانگین رتبهٔ نسبت داده‌شده به معیار شباهت فازی Koczy کمتر از رتبه‌های روش‌های دیگر است.



(شکل-۲): میانگین رتبه معیارهای شباهت فازی با استفاده از آزمون فریدمن

(Figure-2): Average ranking of fuzzy similarity measures using Friedman test

سپس عملکرد بهترین معیار شباهت فازی در روش پیشنهادی به همراه روش‌های شناخته‌شده در زمینه نمونه‌گیری کاهشی دوباره در آزمون فریدمن قرار داده شده است. میانگین رتبه‌های به‌دست‌آمده برای هر روش با استفاده از آزمون فریدمن در جدول (۳) نشان داده شده است. مقدار احتمال فریدمن به‌دست‌آمده ۴۴/۵۷ با درجهٔ آزادی ۶ است و *p-value* محاسبه شده با این آزمون صفر است. به‌دلیل این‌که مقدار احتمال محاسبه شده، از *p*-

[DOI: 10.52547/jsdp.19.2.27]

value بزرگ‌تر است فرضیه صفر رد می‌شود. با توجه به جدول (۳) آشکار است که میانگین رتبه نسبت داده‌شده به روش پیشنهادی که ترکیب خوشه‌بندی کاهشی به‌همراه معیار شباهت Koczy است، کمتر از رتبه‌های روش‌های دیگر است؛ بنابراین روش پیشنهادی به‌عنوان بهترین روش (و همچنین روش کنترلی در آزمون‌های هلم و فینر) در نظر گرفته می‌شود.

(جدول-۳): میانگین رتبه الگوریتم ها با استفاده از آزمون فریدمن

(Table-3): Average rank of algorithms using Friedman test

الگوریتم‌ها	رتبه
RUS	۳/۴
CNN	۴/۷
TL	۳/۶
SBC	۶/۷
OSS	۳/۷
NCL	۳/۵
Proposed Method (hybrid of subtractive clustering and Kozcy fuzzy measures)	**۲/۲**

بعد از اطمینان از وجود تفاوت با استفاده از آزمون فریدمن، از آزمون Post hoc استفاده می‌کنیم تا تشخیص دهیم که آیا روش کنترلی (بهترین روش در آزمون فریدمن) تفاوت آماری با روش‌های دیگر شرکت‌کننده در مقایسه دارد یا خیر. Post hoc های استفاده شده در این مقاله آزمون‌های هلم و فینر [44] می باشد. نتایج حاصل از اعمال آزمون post hoc در جدول (۴) شان داده شده است. رویه‌های هلم و فینر تمام فرضیه‌هایی را که مقدار p-value≤ 0.05 رد می‌کنند [44] و این شرط در جدول (۴) برقرار است و فرضیه صفر رد می‌شود؛ بنابراین روش پیشنهادی نسبت به سایر روش‌ها عملکرد بهتری دارد و باعث بهبود کارایی می‌شود.

(جدول-۴): مقایسه post hoc معیارهای شباهت فازی

(Table-4): Post hoc comparison of fuzzy similarity measures

الگوریتم	Z	P	هلم	فینر	فرضیه
SBC	۶/۱۵۷۷۰۲	۰	۰/۰۰۸۳۳۳	۰/۰۰۸۵۱۲	Reject
CNN	۳/۵۶۶۹۶۱	۰/۰۰۰۳۶۱	۰/۰۱	۰/۰۱۶۹۵۲	Reject
TL	۱/۹۸۹۹۸۹	۰/۰۴۶۵۹۲	۰/۰۱۶۶۶۷	۰/۰۳۳۶۱۷	Reject
OSS	۲/۱۰۲۶۳	۰/۰۳۵۴۸۹	۰/۰۱۲۵	۰/۰۲۵۳۲۱	Reject
NCL	۱/۸۷۷۳۴۸	۰/۰۶۰۴۷	۰/۰۲۵	۰/۰۴۱۸۴۴	Reject
RUS	۱/۶۵۲۰۶۶	۰/۰۹۸۵۲۱	۰/۰۵	۰/۰۵	Reject

۶- نتیجه‌گیری

در این پژوهش مسألهٔ مجموعه‌داده‌های نامتوازن دودویی بررسی و روش جدیدی برای نمونه‌گیری از ردهٔ اکثریت ارائه شده و همچنین در این راستا نیز معیار های شباهت فازی از نظر کارآمدی در رده‌بندی داده‌های نامتوازن مورد تحلیل و بررسی قرار گرفتند. در این روش ابتدا نمونه‌های اکثریت با استفاده از خوشه‌بندی کاهشی به تعدادی خوشه تقسیم شده‌اند و سپس با استفاده از رتبه‌بندی به‌وسیله معیارهای شباهت فازی، نمونه‌ها مرتب شده و از هر خوشه با توجه به اندازهٔ آن تعدادی نمونه انتخاب شده‌اند. در این بررسی از مجموعه‌داده‌هایی دودویی و نامتوازن نرم‌افزار Keel استفاده و پس از اعمال پیش‌پردازش الگوریتم رده‌بندی C4.5 برای انجام رده‌بندی به کار برده شده است. نتایج شبیه‌سازی‌شده بر روی مجموعه‌داده‌های آزمون نشان می‌دهد که در بیش‌تر مجموعه‌داده‌های مورد بررسی، روش پیشنهادی خروجی بهتری داشته است. به‌منظور تحلیل دقیق‌تر آزمون‌های آماری به کار گرفته شده‌اند. به خدمت گرفتن این آزمون‌های آماری و مشاهده نتایج نشان‌دهنده برتری معیار شباهت فازی Koczy به‌همراه خوشه‌بندی کاهشی نسبت به سایر روش‌های شناخته‌شده در رده‌بندی داده‌های نامتوازن است.

۷- مراجع 7- References

[1] A. Braun, and et al, "Landslide Susceptibility Mapping in Tegucigalpa, Honduras, Using Data

[DOI: 10.52547/jsdp.19.2.27]

Artificial Intelligence (ICTAI), Vietri sul Mare, , 2015, pp. 705-712.

[13] G.E. Batista, R.C. Prati, and M.C. Monard, "A study of the behavior of several methods for balancing machine learning training data", *ACM SIGKDD explorations newsletter*, vol. 6(1), pp. 20-29, 2004.

[14] P. Hart, "The condensed nearest neighbor rule (Corresp.)", *IEEE transactions on information theory*, vol. 14(3), pp. 515-516, 1968.

[15] I.Tomek, "Two modifications of CNN", *IEEE Trans. Systems, Man and Cybernetics*, vol.6, pp. 769-772, 1976.

[16] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution", in Conference on Artificial Intelligence in Medicine in Europe, Springer. 2001.

[17] S.-J.Yen, and Y.-S. Lee, "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset", *in Intelligent Control and Automation*, Springer. pp. 731-740, 2006.

[18] M. Kubat, and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection", in Icml. 1997. Nashville, USA.

[19] S. Gazzah , A.H., N. Essoukri Ben Amara, "A hybrid sampling method for imbalanced data", pp. 1-6, 2015.

[20] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning", in International conference on intelligent computing, 2005, Springer.

[21] H. He, et al, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008.

[22] G. Cohen, et al., "Learning from imbalanced data in surveillance of nosocomial infection", *Artificial intelligence in medicine*, vol. 37(1), pp. 7-18, 2006.

[23] S. Tang, and S-.p. Chen, "The generation mechanism of synthetic minority class examples", in 2008 International Conference on Information Technology and Applications in Biomedicine, IEEE, 2008,.

[24] J. Stefanowski, and S. Wilk, "Selective pre-processing of imbalanced data for improving classification performance", in International Conference on Data Warehousing and Knowledge Discovery, Springer, 2008.

Mining Methods", in IAEG/AEG Annual Meeting Proceedings, San Francisco, California, 2018-Volume 1. 2019. Springer.

[2] S.Fotouhi, S. Asadi, and M.W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data", *Journal of biomedical informatics*, 2019.

[3] N. Junsomboon, and T. Phienthrakul, "Combining over-sampling and under-sampling techniques for imbalance dataset", in Proceedings of the 9th International Conference on Machine Learning and Computing. 2017. ACM.

[4] S.A. Golder, B.A. Huberman, "Usage patterns of collaborative tagging systems", *Journal of information science*, vol. 32(2), pp. 198-208. 2006.

[5] Y. Sun, and et al., "Cost-sensitive boosting for classification of imbalanced data", *Pattern Recognition*, vol. 40(12), pp. 3358-3378, 2007.

[6] Z.-H. Zhou, X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem", *IEEE Transactions on Knowledge & Data Engineering*, pp. 63-77. 2006.

[7] N.V. Chawla, and et al., "SMOTE: synthetic minority over-sampling technique", *Journal of artificial intelligence research*, vol. 16, pp. 321-357. 2002.

[8] E. Fernandes, and et al., "Ensemble of Classifiers based on MultiObjective Genetic Sampling for Imbalanced Data", IEEE Transactions on Knowledge and Data Engineering, 2019.

[9] A. Roy, et al. "A study on combining dynamic selection and data preprocessing for imbalance learning", *Neurocomputing*, pp. 179-192, 2002.

[10] W. Xie, G.Liang, Z. Dong, B. Tan, and B. Zhang, "Mathematical Problems in Engineering; An Improved Oversampling Algorithm Based on the Samples", Selection Strategy for Classifying Imbalanced Data. 2019.

[11] V.C. Silvia Cateni, M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems", Neurocomputing,Elsevier.

[12] T. M. Khoshgoftaar, A.F., D. J. Dittman and A. Napolitano, "Ensemble vs. Data Sampling: Which Option Is Best Suited to Improve Classification Performance of Imbalanced Bioinformatics Data?" *2015 IEEE 27th International Conference on Tools with*

[DOI: 10.52547/jsdp.19.2.27]

[25] D.M.B. Tarigan, and D.P. Rini, "Particle Swarm Optimization–Based on Decision Tree of C4. 5 Algorithm for Upper Respiratory Tract Infections (URTI) Prediction", *in Journal of Physics: Conference Series*, IOP Publishing, 2019.

[26] D. Devi, and B. Purkayastha, "Redundancy-driven modified Tomek-link based undersampling: a solution to class imbalance", *Pattern Recognition Letters*, vol.۹۳, pp. 3-12, ۲۰۱۷.

[27] K. Javed, R. Gouriveau, and N. Zerhouni, "A new multivariate approach for prognostics based on extreme learning machine and fuzzy clustering", *IEEE transactions on cybernetics*, vol.45(12), pp. 2626-2639, 2015.

[28] X.L. Xie, and G. Beni, "A validity measure for fuzzy clustering", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol.(8), pp. 841-847. 1991

[29] K.Bataineh, M. Naji, and M. Saqer, "A comparison study between various fuzzy clustering algorithms", *Editorial Board*, vol. 5, pp. 335, 2011.

[30] Y. Ding, and X. Fu, "Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm", *Neurocomputing*, vol.188, pp. 233-238, 2016.

[31] R.R.Yager, and D.P. Filev, "Generation of fuzzy rules by mountain clustering", *Journal of Intelligent & Fuzzy Systems*, vol. 2(3), pp. 209-219. 1994.

[32] S.L. Chiu, "Fuzzy model identification based on cluster estimation", *Journal of Intelligent & Fuzzy Systems*,vol. 2(3), pp. 267-278. 1994.

[33] D. W.Kim, et al., "A kernel-based subtractive clustering method", *Pattern Recognition Letters*, vol. 26(7), pp. 879-891, 2005.

[34] M. Y Chen, "A hybrid ANFIS model for business failure prediction utilizing particle swarm optimization and subtractive clustering", *Information Sciences*, vol.220, pp. 180-195. 2013.

[35] S. Zeng, S. M. Chen, .M. O.Teng, "Fuzzy forecasting based on linear combinations of independent variables, subtractive clustering algorithm and artificial bee colony algorithm", *Information Sciences*, vol.484, pp.350-366, 2019.

[36] I. Beg, and S. Ashraf, "Similarity measures for fuzzy sets", *Appl. and Comput. Math*, vol.8(2), pp. 192-202, 2009.

[37] L.T. Kóczy, and D. Tikk, "Fuzzy rendszerek", TypoTEX, Budapest, 2000.

[38] J. Williams, and N. Steele, "Difference, distance and similarity as a basis for fuzzy decision support based on prototypical decision classes", *Fuzzy sets and systems*, vol.131(1), pp. 35-46. 2002.

[39] S. Santini, and R. Jain, "Similarity is a geometer", *Multimedia Tools and Applications*, vol. 5(3), pp. 277-306, 1997.

[40] R. Zwick, E. Carlstein, and D.V. Budescu, "Measures of similarity among fuzzy concepts: A comparative analysis", *International Journal of Approximate Reasoning*, vol. 1(2), pp. 221-242,1987.

[41] S. García, et al., "A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: a case study on the CEC'2005 special session on real parameter optimization", *Journal of Heuristics*, vol.15(6), pp. 617-644, 2009.

[42] O.T. Yıldız, Ö. Aslan, and E. Alpaydın, "Multivariate statistical tests for comparing classification algorithms," *in Learning and Intelligent Optimization*, Springer, pp. 1-15, 2011.

[43] D.J. Sheskin, Handbook of parametric and nonparametric statistical procedures. 2003: Chapman and Hall/CRC.

[44] S.García, and et al., "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power", *Information Sciences*, vol.180(10), pp. 2044-2064, 2010.

سید احسان یثربی نائینی مدرک کارشناسی خود را در سال ۱۳۷۹ در رشته مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه تهران و مدرک کارشناسی ارشد خود را در سال ۱۳۸۲ از دانشگاه فردوسی مشهد دریافت است. وی در بین سال‌های ۸۲ تا ۸۵ عضو هیأت علمی دانشگاه آزاد اسلامی واحد فردوس و در بین سال‌های ۸۵-۹۰ عضو هیأت علمی مؤسسه آموزش عالی توس و در حال حاضر عضو هیأت علمی دانشگاه سراسری تربت حیدریه است. حوزه‌های تخصصی ایشان شامل داده‌کاوی، یادگیری ماشین و کاربرد روش‌های هوشمند در زمینه کشاورزی و پزشکی است. وی تاکنون چندین مقاله و کتاب منتشر کرده است.

نشانی رایانامه ایشان عبارت است از:

e.yasrebi@torbath.ac.ir

[DOI: 10.52547/jsdp.19.2.27]

مهلا حاتمی مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش نرم‌افزار در سال ۱۳۸۸ از دانشگاه بیرجند دریافت کرده است. وی تحصیلات خود را در سال ۱۳۹۳ در دانشگاه شهید باهنر کرمان در رشته مهندسی کامپیوتر گرایش هوش مصنوعی به پایان رسانده است. زمینه‌های پژوهشی مورد علاقه ایشان داده‌کاوی، الگوریتم‌های تکاملی، حل مسائل چندهدفه و یادگیری ماشین است.

نشانی رایانامه ایشان عبارت است از:

Hatami.mahla@gmail.com

[DOI: 10.52547/jsdp.19.2.27]