استخراج ویژگی مبتنی بر تفکیک‌پذیری بیشتر رده‌ها با استفاده از طبقه‌بندهای کمکی

حمیدرضا غفاری* و آتنا جلالی مجاهد
گروه رایانه، دانشگاه آزاد اسلامی فردوس، فردوس، ایران

چکیده

طبقه‌بندی یک روش یادگیری ماشین است که برای پیش‌گویی برچسب یک نمونه خاص با کمترین خطا استفاده می‌شود. در این مقاله، از توانایی پیش‌گویی برچسب به‌کمک طبقه‌بند برای ایجاد ویژگی جدید استفاده شده است. امروزه روش‌های استخراج ویژگی زیادی مانند PCA و ICA وجود دارند که در زمینه‌های مختلف به‌طور وسیع استفاده می‌شوند و از هزینه بالای انتقال به فضای دیگر رنج می‌برند. در روش پیشنهادی، هدف این است که به‌کمک ویژگی جدید، قدرت تفکیک‌پذیری بیشتری بین رده‌های مختلف ایجاد شود و داده‌های درون رده‌ها به یکدیگر نزدیک‌تر و تمایز بیشتری بین داده‌های رده‌های مختلف به وجود آید تا کارایی طبقه‌بندها افزایش یابد. ابتدا به‌کمک یک یا چند طبقه‌بند، برچسب پیشنهادی برای مجموعه‌داده اولیه تعیین و به‌عنوان ویژگی جدید به مجموعه‌داده اولیه اضافه می‌شود. ایجاد مدل به‌کمک مجموعه‌داده جدید انجام می‌شود. ویژگی جدید برای مجموعه‌داده آموزش و آزمون به‌صورت جداگانه به‌دست آورده می‌شود. آزمایش‌ها بر روی بیست مجموعه‌داده استاندارد انجام شده و نتایج روش پیشنهادی با نتایج دو روش بیان‌شده در کارهای مرتبط نیز مقایسه شده است. نتایج نشان می‌دهد که روش پیشنهادی به‌طور قابل توجهی باعث بهبود دقت رده‌بندی شده است. در بخش دوم آزمایش‌ها، برای بررسی میزان مؤثربودن روش پیشنهادی، قدرت تفکیک‌پذیری ویژگی جدید بر اساس دو معیار بهره اطلاعاتی و شاخص جینی بررسی شده است. نتایج نشان می‌دهد که ویژگی به‌دست‌آمده در روش پیشنهادی در بیشتر موارد دارای بهره اطلاعاتی بیشتر و شاخص جینی کمتری است، زیرا بی‌نظمی کمتری دارد. در ادامه، جهت جلوگیری از افزایش ابعاد داده، ویژگی استخراج‌شده با بیشترین بار اطلاعاتی، جایگزین ویژگی با کمترین بار اطلاعاتی شده است. نتایج این مرحله نیز بیانگر افزایش میزان کارایی است.

واژگان کلیدی: استخراج ویژگی، طبقه‌بندی، بهره اطلاعاتی، شاخص جینی.

Feature extraction based on the more resolution of the classes using auxiliary classifiers

Hamid Reza Ghaffari*& Atena Jalali Mojahed
Department of Computer Engineering, University of Ferdows, Ferdows, Iran

Abstract

Classification is a machine learning method used to predict a particular sample's label with the least error. The present study was conducted using label prediction ability with the help of a classifier to create a new feature. Today, there are several feature-extraction methods like principal component analysis (PCA) and independent component analysis (ICA) that are widely used in different fields; however, they all suffer from the high cost of transferring to another space. The purpose of the proposed method was to create a higher distinction between various classes using the new feature in a way that, make the data in the classes closer to each other. As a result, for increasing the efficiency of classifiers, more differentiation is created between the data of various classes. Firstly, the suggested labels for the primary data set were determined using one or more classifiers and added to the primary data set as a new feature. The model was created using a new data set. The new feature for training and testing data

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

● تاریخ ارسال مقاله: ۱۳۹۷/۱۲/۲۳ ● تاریخ پذیرش: ۱۳۹۹/۵/۲۸ ● تاریخ انتشار: ۱۴۰۰/۰۷/۱۷ ● نوع مطالعه: پژوهشی

[DOI: 10.52547/jsdp.18.2.29]

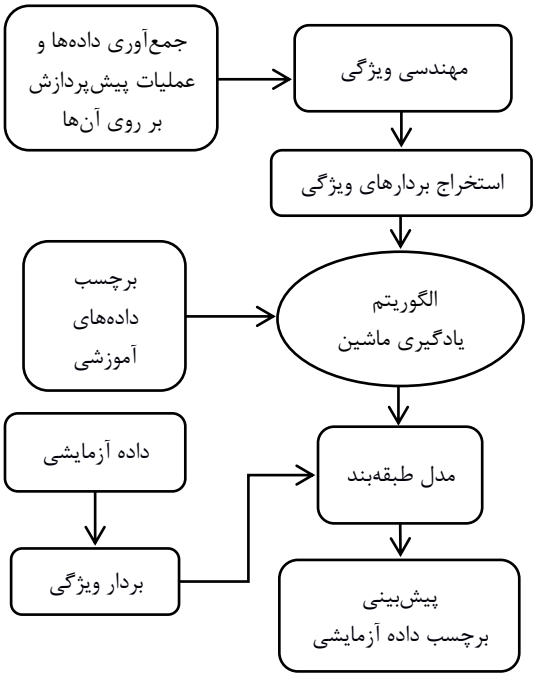
sets was provided separately. The tests were performed on 20 standard data sets and the results of the proposed method were compared with those of the two methods described in the related studies. The outputs indicated that the proposed method has significantly improved the classification accuracy. In the second part of the tests, the resolution of the new feature was examined according to two criteria, namely Information Gain and Gini Index, for examining the effectiveness of the proposed method. The results showed that the feature obtained in the proposed method has higher Information Gain and lower Gini Index in most cases, as it has less irregularity. To prevent the increase in data dimensions, the feature with the least Information Gain was replaced with the feature extracted with the most Information Gain. The results of this step showed an increase in efficiency as well.

Keywords: Feature extraction, classification, information gain, Gini index.

۱- مقدمه

حوزه یادگیری ماشین (ML[1])، یکی از پرکاربردترین حوزه‌های دانش و مهندسی در مفاهیم هوش مصنوعی[2] است. این علم به رایانه‌ها این امکان را می‌دهد که بدون اینکه به‌صراحت برنامه‌ریزی شوند، یاد بگیرند. شناسایی الگو (PR[3])، نیز یکی از مباحث یادگیری ماشین است که در آن ماشین، الگوهای ناشناخته ورودی را دریافت و بر اساس ویژگی‌های آن‌ها، در مورد رده یا کلاس آن‌ها تصمیم‌گیری می‌کند [1].

به‌طورکلی برای حل یک مسأله در مبحث یادگیری ماشین از نوع با ناظر[4]، مطابق با شکل (۱) عمل می‌کنیم.



(شکل-۱): مراحل اجرای مدل یادگیری با ناظر
(Figure-1): steps of implementing the learning model with the supervisor

در مرحله مهندسی ویژگی اقدامات زیر انجام می‌شود:

۱. استخراج ویژگی از نمونه‌های اولیه
۲. بررسی کارایی ویژگی برای مدل مورد نظر
۳. اگر این ویژگی مناسب نباشد، گام‌ها دوباره تکرار می‌شوند.

استخراج ویژگی فرایندی است که در آن با انجام یک سری عملیات بر روی داده‌ها ویژگی‌های بارز و تعیین‌کننده آن مشخص می‌شود. هدف از استخراج ویژگی این است که داده‌های خام به شکل قابل استفاده‌تری برای پردازش‌های آماری بعدی آماده شوند. به‌طورکلی، استخراج ویژگی از مجموعه‌داده‌ها با یک یا دو هدف زیر انجام می‌شود [2]:

۱. افزایش کارایی و سرعت روش‌های دسته‌بندی با کاهش ابعاد به‌خصوص جهت به‌کارگیری برخی روش‌های دسته‌بندی که گام آموزش آن‌ها هزینه و سربار زمانی یا حافظه‌ای بالایی دارند، مانند ماشین بردار پشتیبان.
۲. افزایش دقت روش‌های دسته‌بندی با حذف ویژگی‌های نوفه‌ای (که وجود آن‌ها باعث افزایش خطای دسته‌بندی برای داده‌های جدید می‌شوند) و استخراج ویژگی‌های مناسب (که باعث نزدیک‌شدن داده‌های درون رده‌ها و تمایز بیشتر بین داده‌های رده‌های مختلف می‌شوند.)

ازآنجاکه در مسائل یادگیری با ناظر نیز با یک مجموعه از ویژگی‌ها و برچسب نمونه‌های آموزشی روبه‌رو هستیم، بنابراین کیفیت ویژگی‌های استخراج‌شده نقش کلیدی و مهمی را در عملکرد الگوریتم‌های طبقه‌بندی ایفا می‌کنند. یک ویژگی نامرتبط و زائد اثر منفی بر صحت طبقه‌بندی دارد و پیچیدگی الگوریتم‌های طبقه‌بندی را افزایش داده و زمان اجرای آن‌ها را بالا می‌برد [3]؛ بنابراین گام مهندسی ویژگی، یکی از مهم‌ترین گام‌ها در مراحل یادگیری ماشین است که تفاوت بزرگی را بین یک مدل خوب و یک مدل بد ایجاد می‌کند. در این گام برای بهبود عملکرد طبقه‌بندها، در جستجوی مجموعه مناسبی از ویژگی‌ها هستیم که قدرت تمایز و تفکیک‌پذیری بالایی را

[1] Machine learning
[2] Artificial Intelligence
[3] Pattern Recognition
[4] Supervised Learning

[DOI: 10.52547/jsdp.18.2.29]

در بین نمونه‌های مجموعه‌داده در رده‌های مختلف ایجاد کند.

در ادامه، این مقاله به‌صورت زیر سازمان‌دهی شده است. در بخش ۲، روش‌های مختلف استخراج ویژگی بررسی می‌شود. در بخش ۳، مدل پیشنهادی، مبتنی بر افزودن ویژگی مناسب به مجموعه‌داده آموزشی و آزمایشی برای بالا بردن کارایی طبقه‌بندها، ارائه شده است. در بخش ۴، نتایج آزمایش‌ها بر روی چندین مجموعه‌داده استاندارد گزارش و در این بخش، میزان اثرگذاری ویژگی‌های جدید با توجه به هدف طبقه‌بندی، همراه با تحلیل ریاضی نیز ارائه شده است. نتیجه‌گیری نیز در بخش ۵ بیان شده است.

۲- مرور ادبیات

در این بخش، ابتدا به معرفی اجمالی رایج‌ترین روش‌های طبقه‌بندی می‌پردازیم؛ سپس در بخش کارهای مرتبط، پژوهش‌های انجام‌شده در زمینه استخراج ویژگی را بررسی می‌کنیم.

۲-۱- طبقه‌بندی با ناظر

طبقه‌بندی با ناظر یکی از روش‌های یادگیری ماشین است که در آن مجموعه‌ای از داده‌های ورودی و برچسب‌های خروجی به ماشین داده می‌شود و ماشین سعی می‌کند تا رابطه بین ورودی و خروجی را یاد بگیرد و از آن برای ارائه برچسب به نمونه‌های دیده‌نشده استفاده کند. از الگوریتم‌های طبقه‌بندی رایج می‌توان به الگوریتم‌های k نزدیک‌ترین همسایه، ماشین بردار پشتیبان، درخت تصمیم و جنگل تصادفی اشاره کرد. گفتنی است که در این مبحث، هر نمونه داده x_i با یک بردار ویژگی نشان داده می‌شود که این نمونه در یک مسأله دو رده می‌تواند بر اساس برچسب y_i، عضو کلاس A و یا B باشد.

$$y_i = \begin{cases} +1 & x_i \in A \\ -1 & x_i \in B \end{cases}$$

بنابراین مجموعه داده آموزشی X با تعداد N تا نمونه به‌صورت زیر بیان می‌شود:

$$X = \{(x_i, y_i) \mid i = 1,2,..., N\}$$

الگوریتم طبقه‌بندی k نزدیک‌ترین همسایه (KNN[1])[4]، یکی از محبوب‌ترین طبقه‌بندها است که در آن برچسب نمونه آزمایشی بر اساس رأی‌گیری از برچسب k نزدیک‌ترین همسایه‌اش تعیین می‌شود. در این روش برای یافتن k نزدیک‌ترین همسایه باید تمام نمونه‌های آموزشی مجموعه‌داده در حافظه قرار گیرد و فاصله اقلیدسی نمونه آزمایشی با تمام نمونه‌های آموزشی محاسبه و سپس به‌طور صعودی مرتب شود؛ بنابراین در مواقعی که تعداد نمونه‌های آموزشی زیاد است، پیچیدگی مکانی و زمانی این الگوریتم بالا خواهد بود. از کاربردهای این الگوریتم می‌توان به تشخیص غلط املایی و دزدی ادبی اشاره کرد.

الگوریتم طبقه‌بندی ماشین بردار پشتیبان (SVM[2]) [5,6]، به‌عنوان یکی از برجسته‌ترین الگوریتم‌های یادگیری با ناظر محسوب می‌شود که در آن یک ابرصفحه در فضای ویژگی رسم می‌شود که بیشترین فاصله را با نزدیک‌ترین نمونه آموزشی در هر رده دارد؛ چونرا که بیشترین حاشیه، کمترین خطای تعمیم‌پذیری را در طبقه‌بندی به همراه خواهد داشت. از آنجا که اندازه حاشیه برابر با $2/\|w\|$ است و هدف بیشینه‌کردن آن است، بنابراین تابع هدف در این روش با رابطه (۱) تعریف می‌شود:

$$\min \varphi(w) = \frac{1}{2}\|w\|^2 \qquad (۱)$$
$$st: y_i(w^T x_i + b) \geq 1, i = 1,2,..., l$$

b، عرض از مبدأ و w، راستای عمود بر ابرصفحه جداکننده است. با حل مسأله بهینه‌سازی به فرم دوال طبق رابطه (۲)، نقاط بردار پشتیبان یعنی نقاطی که ضریب لاگرانژ (α) آن‌ها بزرگ‌تر از صفر است به دست می‌آیند:

$$\max l = \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x'_j \qquad (۲)$$

سپس w و b برای ابر صفحه جداکننده از رابطه (۳) و (۴) محاسبه می‌شوند:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \qquad (۳)$$

$$b = \frac{1}{n}\sum_{i=1}^{n}(y_i - w x_i) \qquad (۴)$$

n، تعداد نقاط بردار پشتیبان است؛ درنهایت از تابع تصمیم با رابطه (۵) برای پیش‌گویی برچسب نمونه آزمایشی x استفاده می‌شود:

$$f_{w,b}(x) = sign(w^T x + b) \qquad (۵)$$

[1] K-Nearest Neighbor

[2] Support Vector Machine

[DOI: 10.52547/jsdp.18.2.29]

از کاربردهای ماشین بردار پشتیبان، می‌توان به تشخیص چهره و تشخیص صدا اشاره کرد.

الگوریتم طبقه‌بندی درخت تصمیم (DT[1]) [7,8]، یک الگوریتم حریصانه از بالا به پایین است که مجموعه‌داده‌های آموزشی X، با مجموعه‌ویژگی‌های $\{A_1,A_2,\ldots,A_p\}$ را به‌طور متوالی به زیرمجموعه‌های کوچک‌تر $\{X_1,X_2,\ldots,X_n\}$ تقسیم می‌کند تا وقتی‌که همه نمونه‌ها در هر مجموعه‌داده X_i به رده C_i تعلق یابند. این روش مطابق با نظریه اطلاعات عمل می‌کند؛ یعنی در هر زمان یک ویژگی از داده‌های آموزشی به‌عنوان گره انتخاب می‌شود که بیشترین بهره اطلاعاتی[2] را دارد، یعنی بیشتر موجب کاهش بی‌نظمی[3] می‌شود. میزان بهره اطلاعاتی هر ویژگی A_k، از رابطه (۶) به‌دست می‌آید.

$$\begin{aligned} Gain(A_k, X) &= I(X) - \sum_{i=1}^{n} \frac{|X_i|}{|X|} I(X_i) \\ I(X) &= -\sum_{j=1}^{m} p_j \log_2 p_j \\ p_j &= \frac{|X \cap C_j|}{|X|} \\ I(X_i) &= -\sum_{j=1}^{m} \frac{|X_i \cap C_j|}{|X_i|} \log_2 \frac{|X_i \cap C_j|}{|X_i|} \end{aligned} \qquad (۶)$$

در این رابطه، m، تعداد رده‌ها و P_j احتمال رخداد رده C_j، در مجموعه X است. $|X_i|$ تعداد نمونه‌هایی در مجموعه‌داده آموزشی X است که مقادیر V_i را برای صفت A_k دارند. n، تعداد مقادیر ممکن صفت A_k را مشخص می‌کند. $I(X_i)$، توزیع تصادفی نمونه‌ها در زیرمجموعه X_i را با توجه به رده‌های ممکن اندازه می‌گیرد. بعد از ساخت درخت، یک نمونه آزمایشی به درخت داده می‌شود تا با توجه به مقادیر ویژگی‌هایش از ریشه به سمت برگ‌ها حرکت کرده تا در رده موردنظر قرار گیرد. از کاربردهای درخت تصمیم، می‌توان به پردازش زبان طبیعی و مهندسی نرم‌افزار اشاره کرد [9].

الگوریتم طبقه‌بندی جنگل تصادفی (RF[4])[10]، یک الگوریتم ترکیبی قدرتمند است که از تعدادی درخت تصمیم به‌عنوان الگوریتم‌های پایه استفاده می‌کند. هرکدام از این درختان با زیرمجموعه‌ای تصادفی از مجموعه‌داده اولیه به همراه زیرمجموعه‌ای تصادفی از مجموعه ویژگی‌های موجود ساخته می‌شوند، بنابراین تنوع در درختان را خواهیم داشت. این الگوریتم از معیار جینی[5]، مطابق با رابطه (۷)، برای انتخاب ویژگی‌ها استفاده می‌کند و میزان ناخالصی یک ویژگی را نسبت به رده‌ها محاسبه می‌کند.

$$Gini(X) = 1 - \sum_{i=1}^{m} p_j^{\,2} \qquad (۷)$$

در این رابطه، p_j، احتمال تعلق یک نمونه در زیرمجموعه‌ای از X، به رده C_j است و m، تعداد رده‌ها است. شاخص جینی، یک تقسیم‌بندی دودویی (X_1,X_2) را برای هر ویژگی A_k در نظر می‌گیرد؛ بنابراین شاخص جینی با توجه به این تقسیم‌بندی به‌صورت مجموع وزن ناخالصی هر قسمت طبق رابطه (۸) بیان می‌شود.

$$Gini_A(X) = \frac{X_1}{X} Gini(X_1) + \frac{X_2}{X} Gini(X_2) \qquad (۸)$$

میزان کاهش ناخالصی نیز از رابطه (۹) محاسبه می‌شود.

$$\Delta Gini(A) = Gini(X) - Gini_A(X) \qquad (۹)$$

بنابراین زیرمجموعه‌ای که کمینه شاخص جینی را برای صفت A به‌دست می‌آورد، به‌عنوان زیرمجموعه تقسیم‌کننده آن انتخاب می‌شود. پس از اینکه درختان آماده شدند، نمونه‌های جدید توسط هرکدام از این درختان برچسب‌گذاری شده و برچسب نهایی هر نمونه به کمک رأی‌گیری مشخص می‌شود [11].

در مدل پیشنهادی این پژوهش نیز، از هر یک از الگوریتم‌های طبقه‌بندی بالا برای ایجاد یک مدل کمکی استفاده می‌شود و ابتدا توسط آن‌ها ویژگی‌های مناسبی، مبتنی بر پیش‌گویی برچسب از مجموعه‌داده اولیه، استخراج و مجموعه‌داده جدید آموزشی و آزمایشی تشکیل می‌شود.

۲-۲- کارهای مرتبط

استخراج ویژگی‌های مناسب از یک مجموعه‌داده نقش حیاتی در بهبود کیفیت و کارایی روش‌های یادگیری ماشین دارد؛ زیرا یک ویژگی مناسب قدرت تفکیک‌پذیری بیشتری را بین طبقه‌های مختلف ایجاد می‌کند. به‌عبارت‌دیگر باعث نزدیک‌شدن داده‌های درون طبقه‌ها و تمایز بیشتر بین داده‌های طبقه‌های مختلف می‌شود.

روش‌های استخراج ویژگی از تنوع بسیاری برخوردار هستند. در بخش نخست، روش‌هایی را مطرح می‌کنیم که سعی می‌کنند با کمینه‌کردن ماتریس پراکندگی درون‌رده‌ای و بیشینه‌کردن ماتریس برون‌رده‌ای، طبقه‌ها را از هم جدا کنند، مانند تحلیل اجزای اصلی (PCA[6]) [12]،

[1] Decision Tree
[2] Information Gain
[3] Entropy
[4] Random Forest
[5] Gini Index
[6] Principal Component Analysis

[DOI: 10.52547/jsdp.18.2.29]

تحلیل اجزای مستقل (ICA[1]) [13] و تحلیل تفکیک خطی (LDA[2]) [14]. این الگوریتم‌ها در پیش‌پردازش داده‌های مربوط به پردازش سیگنال، پردازش صوت، تشخیص چهره و غیره بسیار پرکاربرد هستند.

در روش PCA، محورهای مختصات جدیدی برای داده‌ها تعریف شده و داده‌ها بر اساس این محورهای مختصات جدید بیان می‌شوند. نخستین محور باید در جهتی قرار گیرد که واریانس داده‌ها بیشینه شود (در جهتی که پراکندگی داده‌ها بیشتر است). دومین محور باید عمود بر محور نخست به‌گونه‌ای قرار گیرد که واریانس داده‌ها بیشینه شود. به همین ترتیب برای محورهای بعدی این روند ادامه دارد. در روش ICA نیز یک نگاشت خطی انجام می‌گیرد، اما بردارهای این نگاشت لزوماً بر یکدیگر عمود نیستند. در کلیه این روش‌ها یک زیرفضای مناسب m بعدی از فضای اصلی d بعدی ویژگی‌ها، تعیین می‌شود به‌طوری‌که m<=d است؛ بنابراین این روش‌ها سربار پردازشی زیادی دارند و خروجی آن‌ها نیز به‌راحتی برای کاربران قابل تفسیر نیست.

روش LDA، نسبت ماتریس پراکندگی بین‌رده‌ای (S_b) را به ماتریس پراکندگی درون کلاسی (S_w) بیشینه می‌کند. ویژگی استخراج‌شده با بیشینه‌کردن معیار $tr(S_w^{-1}S_b)$ به‌دست می‌آید. محدودیت اصلی LDA این است که وقتی تعداد نمونه‌های آموزشی محدود باشد، ماتریس پراکندگی درون‌دسته‌ای منفرد شده و کارایی LDA به‌شدت پایین می‌آید.

در بخش نخست، روش‌هایی را در زمینه استخراج ویژگی بررسی کردیم که به‌طورعمومی محدودیت‌هایی دارند. به‌عنوان مثال الگوریتم PCA با اینکه یک روش بسیار خوبی در بیش‌تر مسائل هست، اما یکی از بارزترین معایب آن این است که رویکرد این الگوریتم بر اساس پراکندگی داده‌ها است. به این معنی که از نظر این الگوریتم ویژگی‌هایی که پراکندگی کمتری دارند حاوی اطلاعات مهمی نیستند؛ بنابراین در ساخت ویژگی‌های جدید به آن‌ها سهم کمتری داده می‌شود. این در حالی است که ممکن است، یک ویژگی‌ای که پراکندگی کمتری دارد، قدرت تفکیک‌پذیری بهتری داشته باشد و بتوان با کمک این ویژگی، داده‌های دو رده یا چند رده را با دقت بالاتری دسته‌بندی کرد؛ بنابراین ممکن است این الگوریتم در بعضی مسائل کارایی خوبی نداشته باشد. برخلاف الگوریتم PCA، الگوریتم LDA روشی است که بر اساس معیار تفکیک‌پذیری، داده‌ها را از فضای اصلی به فضای جدید نگاشت می‌دهد، اما هنگامی‌که تعداد نمونه‌ها کم است کارایی این الگوریتم LDA نیز کاهش می‌یابد؛ بنابراین انگیزه اصلی ما در بخش دوم، بررسی روش‌هایی است که بر اساس معیار تفکیک‌پذیری عمل کرده و هم‌زمان بر روی تعداد نمونه‌های کم آموزشی نیز قابل اجرا باشند و هزینه تولید ماتریس پراکندگی و انتقال داده‌ها به فضای دیگر را نیز نداشته باشند. روش پیشنهادی مقاله از این روش‌ها الهام گرفته و در بخش آزمایش‌ها با آن‌ها نیز از نظر کارایی مقایسه می‌شود.

در مقاله [15]، از روشی موسوم به TANN[3] برای به‌دست‌آوردن ویژگی جدید استفاده شده است. در این روش ابتدا به‌کمک الگوریتم خوشه‌بندی K میانه[4]، مراکز خوشه‌ها را به‌دست آورده، سپس فاصله هر نمونه با تمام مراکز خوشه‌ها و فاصله مراکز خوشه‌ها نیز نسبت به هم محاسبه می‌شوند؛ سپس با درنظرگرفتن فاصله هر نمونه با دو مرکز خوشه و فاصله همان دو مرکز، یک مثلث تصور می‌شود که با جمع‌کردن این سه فاصله بر روی سه ضلع مثلث، یک ویژگی جدید برای مجموعه‌داده‌ها استخراج می‌شود؛ درنهایت صحت مجموعه‌داده جدید با الگوریتم k نزدیک‌ترین همسایه مورد ارزیابی قرار گرفته است. عیب اصلی این روش آن است که فقط برای مجموعه‌داده‌هایی مناسب است که تعداد رده‌های آن‌ها کم است.

مقاله [16]، با روشی CANN[5]، ویژگی جدید را استخراج کرده است. در این روش ابتدا به‌کمک الگوریتم خوشه‌بندی k میانه، مراکز خوشه‌ها را به‌دست آورده، سپس مجموع فاصله میان هر نمونه با مراکز خوشه‌ها و نزدیک‌ترین همسایه‌اش در همان خوشه را به‌عنوان ویژگی جدید به کار گرفتند؛ سپس از الگوریتم k نزدیک‌ترین همسایه و ماشین بردار پشتیبان برای طبقه‌بندی مجموعه‌داده‌ها با ویژگی جدید استفاده کردند.

در مقاله [17]، در روش DCNN[6]، بیان شده است که معیار فاصله استفاده‌شده در روش CANN به‌تنهایی معیار خوبی برای بیان یک ویژگی نیست؛ بنابراین علاوه‌بر معیار فاصله، از چگالی داده‌ها نیز به‌عنوان یک ویژگی دیگر استفاده شده است. در این روش برای تعیین چگالی هر نمونه، یک دایره به مرکز آن نمونه با یک شعاع که

[1] Independent Component Analysis
[2] Linear Discriminant Analysis
[3] Triangle Area based NearestNeighbors
[4] K-means
[5] Cluster Center and Nearest Neighbors
[6] Density Cluster centers and Nearest Neighbors

[DOI: 10.52547/jsdp.18.2.29]

به‌صورت دستی تنظیم می‌شود در نظر گرفته می‌شود؛ بنابراین چگالی هر نمونه با تعداد نمونه‌های درون آن دایره برابر است. در این روش نیز از الگوریتم k نزدیک‌ترین همسایه جهت ارزیابی مجموعه‌داده جدید استفاده شده است.

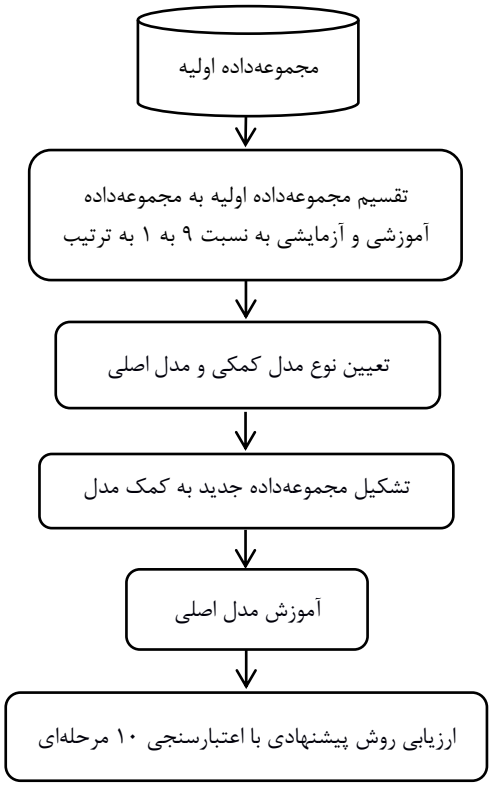
۳- روش پیشنهادی

از آنجا که در مسائل یادگیری با ناظر، ما با دسته‌ای از داده‌ها مواجه هستیم که توسط یک ناظر، طبقه‌بندی شده‌اند و در رده‌های درست قرار گرفته‌اند؛ بنابراین ما نیز در این پژوهش از پیش‌گویی برچسب نمونه‌ها به‌عنوان یک ویژگی خوب با قدرت تفکیک‌پذیری و تمایز بالا در مجموعه‌داده‌ها استفاده کردیم، به هدف این‌که کارایی الگوریتم‌های طبقه‌بندی را بهبود دهیم. روش پیشنهادی که ما آن را به‌اختصار CBC[1] نامیدیم، سعی می‌کند تا یک مجموعه‌داده جدید آموزشی و آزمایشی تشکیل دهد. در این روش برای ساخت مجموعه‌داده جدید از یک مدل کمکی استفاده کردیم؛ لذا در مدل پیشنهادی علاوه‌بر نیاز به یک رده‌بند اصلی، به یک رده‌بند کمکی هم نیاز است.

همان‌طور که در شکل (۲) مشاهده می‌شود، ابتدا مجموعه‌داده اولیه را به دو مجموعه آموزشی و آزمایشی به نسبت ۹ به ۱ به‌ترتیب، تقسیم و برای تعیین مجموعه‌داده جدید، از یک مدل کمکی استفاده می‌کنیم. مدل کمکی یکی از طبقه‌بندهای رایج است. برای ساخت مدل اصلی از مجموعه‌داده جدید که حاصل مدل کمکی است، استفاده شده است.

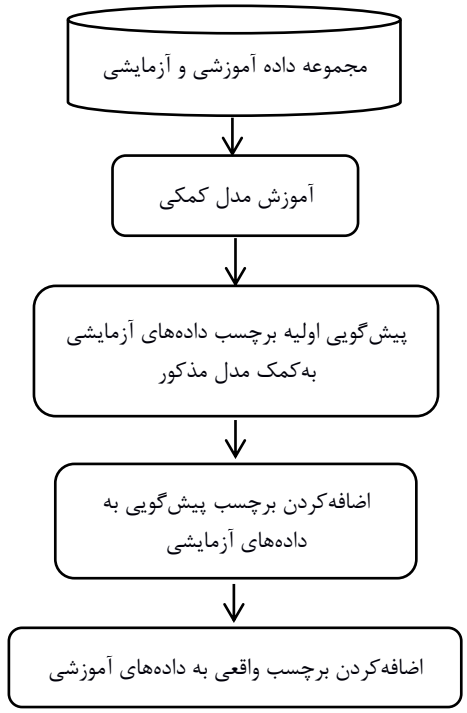
برای ساخت مجموعه‌داده جدید، مانند شکل (۳)، به‌کمک نمونه‌های آموزشی، مدل کمکی ساخته و برچسب مجموعه‌داده آزمایشی به‌کمک مدل کمکی تعیین می‌شود. برچسب‌های پیش‌بینی‌شده را به‌عنوان یک ویژگی جدید به مجموعه‌داده آزمایشی اضافه و از طرف دیگر برچسب واقعی داده‌های آموزشی را نیز به‌عنوان یک ویژگی جدید به نمونه‌های آموزشی اضافه می‌کنیم. تا اینجا ما یک مجموعه‌داده جدید آموزشی و آزمایشی را تشکیل داده‌ایم.

در مرحله بعد طبق شکل (۴)، ساخت مدل اصلی به‌کمک مجموعه‌داده جدید صورت می‌گیرد. آموزش مدل اصلی با استفاده از مجموعه‌داده آموزشی جدید و ارزیابی مدل به‌کمک مجموعه‌داده آزمایشی جدید انجام می‌شود.



(شکل-۲): ساختار کلی مدل پیشنهادی CBC

(Figure-2): The general structure of the proposed CBC model

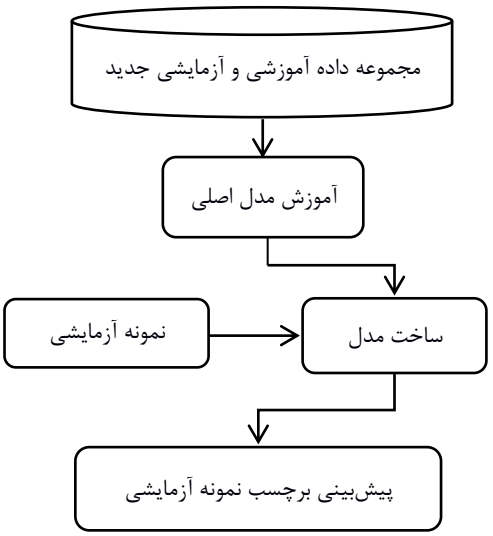


(شکل-۳): تشکیل مجموعه داده جدید مبتنی بر پیش‌گویی اولیه

(Figure-3): Creating a new dataset based on the initial prediction

[1] Classification By Classification

[DOI: 10.52547/jsdp.18.2.29]



(شکل-۴): مراحل آزمایش مدل پیشنهادی CBC
(Figure-4): Test steps of the proposed CBC model

۳-۱- استخراج ویژگی جدید در روش پیشنهادی CBC

در روش پیشنهادی CBC، ما سعی داریم تا به‌کمک طبقه‌بندهای کمکی مختلف، برچسب نمونه‌های آزمایشی را پیش‌گویی کرده و آن‌ها را به‌عنوان یک ویژگی جدید به ویژگی‌های نمونه‌های آزمایشی اضافه کنم. در روش نخست که آن را CBCsvm[1] نامیدیم، برای پیش‌بینی برچسب نمونه‌های آزمایشی از رده‌بند کمکی ماشین بردار پشتیبان استفاده کردیم. در روش دوم که آن را CBCknn[2] نامیدیم، برای پیش‌بینی برچسب نمونه‌های آزمایشی از رده‌بند k نزدیک‌ترین همسایه استفاده و در روش سوم که آن را CBCdt[3] نامیدیم، برای پیش‌بینی برچسب نمونه‌های آزمایشی از رده‌بند درخت تصمیم استفاده کردیم. در روش چهارم که آن را CBCskt[4] نامیدیم، از سه ویژگی قبلی یعنی ویژگی‌های حاصل از طبقه‌بند ماشین بردار پشتیبان، k نزدیک‌ترین همسایه و درخت تصمیم میانگین گرفته و به‌عنوان ویژگی جدید به ویژگی‌های نمونه‌های آزمایشی اضافه و در روش پنجم که آن را CBCrf[5] نامیدیم، برای پیش‌بینی برچسب نمونه‌های آزمایشی از رده‌بند جنگل تصادفی استفاده کردیم.

گفتنی است که در تمامی روش‌های بالا بعد از پیش‌بینی برچسب داده‌های آزمایشی، به مجموعه‌داده آموزشی نیز برچسب واقعی‌شان را به‌عنوان یک ویژگی جدید اضافه کردیم. در آخرین روش که آن را CBCstr[6] نامیدیم سه ویژگی حاصل از هر یک از روش‌های CBCsvm، CBCtd و CBCrf را به‌عنوان سه ویژگی جدید به ویژگی‌های نمونه‌های آزمایشی اضافه و در ضمن در همین روش، به مجموعه داده آموزشی برچسب واقعی‌شان و برچسب حاصل از اعمال الگوریتم ماشین بردار پشتیبان و درخت تصمیم را به‌عنوان سه ستون ویژگی جدید اضافه کردیم.

[1] CBC support vector machine
[2] CBC k-nearest neighbor
[3] CBC decision tree
[4] CBC svm knn tree
[5] CBC random forest

۴- آزمایش‌ها

در این مرحله، روش پیشنهادی CBC را بر روی بیست مجموعه‌داده استاندارد اعمال کرده و از چهار الگوریتم رده‌بند k نزدیک‌ترین همسایه، ماشین بردار پشتیبان، درخت تصمیم و جنگل تصادفی نیز به‌عنوان طبقه‌بند کمکی استفاده کردیم؛ سپس صحت مدل پیشنهادی را ارزیابی کرده‌ایم. درنهایت کارایی روش پیشنهادی را با کارایی چهار الگوریتم طبقه‌بند یادشده بر روی مجموعه داده‌های اولیه و همچنین با کارایی دو روش TANN و CANN که در بخش کارهای مرتبط معرفی کرده نیز مقایسه کردیم.

۴-۱- مشخصات مجموعه‌داده‌ها

در این پژوهش، از بیست مجموعه‌داده UCI [18] برای فرایند طبقه‌بندی و انجام آزمایش‌ها استفاده کردیم. جزئیات مشخصات این مجموعه‌داده‌ها در جدول (۱) بیان شده است.

(جدول-۱): مشخصات بیست مجموعه‌داده UCI
(Table-1): Specifications of 20 UCI Data Sets

مجموعه داده‌ها	تعداد نمونه‌ها	تعداد ویژگی‌ها	تعداد کلاس‌ها
Iris	150	4	3
Wine	178	13	3
Wdbc	569	30	2
Hheart	270	13	2
Banana	5300	2	2
Bupa	345	6	2
Sonar	208	60	2
Segmentation	210	18	7
Pima	768	8	2
mammographic	748	4	2
Ionosphere	351	32	2

[6] CBC svm tree random forest

[DOI: 10.52547/jsdp.18.2.29]

2	10	1024	Cloude
3	4	160	Hayes-Roth
3	10	5472	Page-Blocks
3	20	7200	Thyroid
2	10	2500	Magic
2	6	432	Monk-2
2	20	7400	Ring
2	20	7400	Twonorm
3	24	5456	Wall_Following Robot Navigation Data

در آزمایش‌ها، همه مجموعه‌داده‌ها با استفاده از روشی با عنوان یک رده در برابر همه رده‌ها (OAA[1]) [19]، به مسائل دو رده تبدیل شده‌اند؛ درنهایت، برای تخمین کارایی مدل پیشنهادی، از روش اعتبارسنجی ضربدری (گردشی)[2] [20]، با k برابر ۱۰، استفاده کردیم. ارزیابی با این روش مشخص می‌کند که نتایج یک تحلیل آماری بر روی یک مجموعه‌داده تا چه اندازه قابل تعمیم و مستقل از داده‌های آموزشی است. در این روش، داده‌ها را به ده زیرمجموعه مساوی افراز کرده و هر بار از نُه زیرمجموعه آن برای آموزش طبقه‌بند و از یک زیرمجموعه آن برای ارزیابی رده‌بند استفاده کردیم. این روال را ده بار تکرار کرده تا همه نمونه‌ها هم برای آموزش و هم برای ارزیابی و آزمایش مدل استفاده شوند. در آخر از نتایج اعتبارسنجی‌ها، میانگین گرفته و آن را به‌عنوان یک تخمین نهایی گزارش کردیم.

۲-۴- معیار ارزیابی

در مرحله ارزیابی سامانه پیشنهادی، از معیار صحت[3] که به‌کمک ماتریس درهم‌ریختگی[4] بیان می‌شود، استفاده کردیم. جدول (۲)، ماتریس درهم‌ریختگی را برای یک مسأله رده‌بندی دو رده نشان می‌دهد.

(جدول-۲): ماتریس درهم ریختگی [21]

(Table-2): Confusion matrix

		ردهٔ واقعی	
		مثبت	منفی
کلاس	مثبت	TP	FP
پیش‌بینی‌شده	منفی	FN	TN

چهار مقدار موجود در این جدول به‌شرح زیر تعریف می‌شوند:

[1] One-Against-All
[2] k-fold Cross Validation
[3] Accuracy
[4] Confusion Matrix

[5]TP: تعداد نمونه‌هایی که به‌درست در ردهٔ مثبت قرار گرفته‌اند.

[6]TN: تعداد نمونه‌هایی که به‌درست در ردهٔ منفی قرار گرفته‌اند.

[7]FP: تعداد نمونه‌هایی که به‌اشتباه در ردهٔ مثبت قرار گرفته‌اند.

[8]FN: تعداد نمونه‌هایی که به‌اشتباه در ردهٔ منفی قرار گرفته‌اند.

بنابراین طبق ماتریس بالا، صحت رده‌بندی برای محاسبه کارایی طبقه‌بندها از رابطه (۱۰) به‌دست می‌آید.

$$acc = \frac{TP + TN}{TP + FP + TN + FN} \qquad (۱۰)$$

بنابراین صحت رده‌بندی، نسبت تعداد نمونه داده‌های آزمایشی که به‌درستی به دو رده مثبت و منفی دسته‌بندی شده‌اند به کل نمونه‌های موجود در مجموعه داده‌های مورد آزمایش است؛ لذا عملکرد الگوریتمی بهتر است که صحت رده‌بندی بالاتری دارد.

۳-۴- ارزیابی روش پیشنهادی CBC

به‌طورکلی در مدل پیشنهادی CBC، از برچسب‌های پیش‌بینی‌شده به‌وسیلهٔ الگوریتم‌های رده‌بندی کمکی مختلف به‌عنوان برچسب نمونه‌های آزمایشی و از برچسب‌های اصلی نیز به‌عنوان ویژگی جدید برای داده‌های آموزشی استفاده و مجموعه‌داده جدیدی را ایجاد کردیم.

دوباره الگوریتم‌های رده‌بندی را به مجموعه‌داده جدید اعمال کردیم تا به این پرسش پاسخ دهیم که آیا ویژگی‌های افزوده‌شده به مجموعه‌داده اولیه می‌تواند کارایی رده‌بندها را در پیش‌گویی صحیح برچسب نمونه آزمایشی افزایش دهد؟ اگر پاسخ مثبت باشد به این معنی است که این ویژگی جدید، میزان تفکیک‌پذیری بین رده‌ها را افزایش داده است. در ادامه آزمایش‌ها نیز مناسب بودن ویژگی‌ها را به‌کمک دو مفهوم بهره اطلاعاتی و شاخص جینی بررسی کرده‌ایم.

همان‌طور که جدول‌های (۳، ۴، ۵ و ۶) نشان می‌دهند به‌ترتیب از رده‌بند ماشین بردار پشتیبان، درخت تصمیم، جنگل تصادفی و k نزدیک‌ترین همسایه به‌عنوان رده‌بند اصلی استفاده کردیم؛ درضمن همین رده‌بندها را

[5] True Positive
[6] True Negative
[7] False Positive
[8] False Negative

[DOI: 10.52547/jsdp.18.2.29]

بر روی مجموعه‌داده اولیه (Base) و مجموعه‌داده حاصل از دو روش CANN و TANN نیز اعمال و برای مقایسه از بیست مجموعه‌داده و معیار صحت استفاده کرده و سپس نتایج را از صحت بالا به پایین رتبه‌بندی کردیم.

در مرحله ارزیابی، روشی عملکرد بهتری دارد که به‌طور میانگین صحت بالاتر و رتبه پایین‌تری داشته است. میانگین صحت روش‌ها و همچنین میانگین رتبه عملکرد آن‌ها در جدول‌های (۷، ۸، ۹ و ۱۰) آورده شده است.

مطابق با جدول (۷)، روش ماشین بردار پشتیبان بر روی مجموعه‌داده اولیه دارای صحت ۸۲/۷۸٪ است درحالی‌که وقتی با روش پیشنهادی CBCrf، یعنی روشی که از الگوریتم جنگل تصادفی به‌عنوان رده‌بند کمکی استفاده کرده است، مجموعه‌داده جدید را ساخته و الگوریتم ماشین بردار پشتیبان را به آن اعمال کردیم، صحت رده‌بندی برابر با ۸۵/۶۵٪ شده است؛ یعنی حدود چهار درصد، کارایی بهبود یافته است. البته همان‌طور که جدول (۷) نشان می‌دهد، تمامی روش‌های پیشنهادی CBC، نتایج رده‌بندی روش ماشین بردار پشتیبان را نسبت به مجموعه‌داده اولیه بهبود داده‌اند. این در حالی است که با دو روش CANN و TANN هیچ بهبودی حاصل نشده است.

طبق جدول (۸)، باز هم همه روش‌های پیشنهادی به‌جز روش CBCsvm و CBCsktr کارایی الگوریتم درخت تصمیم را بالا برده‌اند. به‌خصوص در روش CBCrf و CBCstrf که صحت رده‌بندی الگوریتم درخت تصمیم، حدوداً از ۸۷٪ به ۹۱٪ افزایش یافته است. در این حالت روش TANN نیز باعث کمی بهبود در عملکرد این الگوریتم شده است؛ اما طبق جدول (۹)، روش‌های پیشنهادی هیچ‌کدام بهبودی را در الگوریتم جنگل تصادفی ایجاد نکرده‌اند و این به‌خاطر ماهیت تصادفی این الگوریتم در انتخاب ویژگی‌ها در هنگام ساخت درختان تصمیم است. در جدول (۱۰) نیز انواع روش‌های پیشنهادی در CBC باعث بهبود عملکرد روش k نزدیک‌ترین همسایه شده‌اند. روش پیشنهادی CBCrf باز هم صحت رده‌بند K نزدیک‌ترین همسایه را از حدود ۸۷٪ به ۹۱٪ افزایش داده و در این حال روش CANN نیز باعث کمی بهبود در الگوریتم شده است.

در جدول (۱۱)، عملکرد روش‌های طبقه‌بندی k نزدیک‌ترین همسایه، ماشین بردار پشتیبان، درخت تصمیم و جنگل تصادفی را فقط بر روی بخش‌هایی از روش پیشنهادی CBC که نتایج خوبی داشتند را با حالت اولیه مجموعه داده، مقایسه کردیم. نتایج نشان می‌دهد که در بین روش‌های پیشنهادی که بیشتر آن‌ها خوب بودند، روش پیشنهادی CBCrf از همه مؤثرتر بوده و کارایی سه الگوریتم k نزدیک‌ترین همسایه، ماشین بردار پشتیبان و درخت تصمیم را حدود چهار درصد بهبود داده است؛ بنابراین بهترین الگوریتم طبقه‌بندی برای مرحله انتخاب مدل کمکی، رده‌بند جنگل تصادفی است

(جدول-۳): مقایسه صحت رده‌بند ماشین بردار پشتیبان بر روش پیشنهادی CBC و سایر روش‌ها به درصد، بر روی بیست مجموعه‌داده استاندارد و رتبه‌بندی نتایج (R)

(Table-3): Comparison of the support vector classifier classification on the proposed CBC method and other methods by percentage, on 20 standard datasets and ranking the results (R)

Dataset	Base	R	CBC svm	R	CBC knn	R	CB Cdt	R	CB Cskt	R	CB Crf	R	CBC sktr	R	CB Cstr	R	CA NN	R	TA NN	R
iris	97.33	2	97.33	2	96.66	4	96.44	5	96.88	3	96.44	5	96.44	5	96.44	5	82.66	6	**97.55**	1
wine	92.31	2	92.31	2	**92.51**	1	90.83	4	**92.51**	1	91.40	3	90.82	5	90.82	5	84.28	7	90.81	6
wdbc	90.85	2	90.76	4	90.67	5	89.17	8	**91.02**	1	**91.02**	1	90.14	9	90.84	3	90.32	7	90.50	6
heart	76.66	3	73.51	10	74.44	8	74.62	7	75.74	4	**79.44**	1	75.55	5	76.85	2	74.25	9	75.18	6
banana	89.60	3	89.60	3	88.65	6	87.39	7	89.85	2	88.94	5	88.94	5	88.94	5	**90.75**	1	89.50	4
bupa	66.25	7	66.68	6	62.92	10	65.92	9	67.97	4	72.17	2	71.73	3	**72.31**	1	66.22	8	66.94	5
sonar	**57.69**	1	**57.69**	1	**57.69**	1	56.99	3	**57.69**	1	57.22	2	56.99	3	56.99	3	56.98	4	**57.69**	1
Segmentation	94.69	8	94.69	8	95.64	2	95.10	6	95.57	3	**95.98**	1	95.37	5	95.44	4	91.44	9	94.76	7
pima	73.03	6	73.03	6	73.75	5	70.76	9	74.14	4	**74.67**	1	74.15	3	74.54	2	71.60	8	72.77	7
mammographic	62.96	9	63.02	8	74.20	5	73.06	6	**74.73**	1	74.47	2	74.33	4	74.40	3	53.81	10	63.46	7

[DOI: 10.52547/jsdp.18.2.29]

ionosphere	89.76	2	89.76	2	89.62	3	89.18	4	89.62	3	**89.90**	1	89.76	2	89.76	2	89.76	2	89.76	2
cloud	89.04	5	98.04	5	97.07	8	99.02	3	98.48	4	**99.21**	1	99.02	3	99.17	2	97.85	7	97.95	6
hayes_roth	65.08	9	68.65	7	77.54	6	84.59	4	82.35	5	**85.21**	1	84.63	3	84.81	2	62.75	10	67.79	8
page-blocks	94.55	8	96.37	6	97.66	5	**98.16**	1	98.02	4	98.10	3	98.00	5	98.12	2	94.51	9	94.57	7
thyroid	95.30	3	95.40	8	95.53	7	**97.90**	1	97.00	6	97.80	2	97.06	5	97.10	4	95.40	8	95.16	9
magic	81.00	10	81.49	8	86.84	4	83.70	6	85.79	5	**89.00**	1	88.69	3	88.85	2	81.60	7	81.14	9
monk-2	97.21	2	97.21	2	95.59	3	**100**	1	97.21	2	**100**	1	**100**	1	**100**	1	86.79	4	97.21	2
ring	84.60	2	84.60	2	**84.65**	1	80.89	6	**84.65**	1	83.15	5	80.54	7	80.54	7	84.20	4	84.49	3
twonorm	62.80	3	62.80	3	63.00	2	61.50	4	63.00	2	**63.04**	1	61.45	6	61.50	5	59.60	8	59.90	7
Wall_Following Robot Navigation Data	85.89	5	85.89	5	85.52	8	88.77	2	**88.89**	1	85.82	6	87.55	4	87.95	3	85.47	9	85.59	7

(جدول-۴): مقایسه صحت رده‌بند درخت تصمیم بر روش پیشنهادی CBC و سایر روش‌ها به درصد، بر روی بیست مجموعه داده استاندارد و رتبه‌بندی نتایج (R)

(Table-4): Comparing the accuracy of the decision tree classifier on the proposed CBC method and other methods as a percentage, on 20 standard datasets and ranking the results (R)

Dataset	Base	R	CBC svm	R	CBC knn	R	CB Cdt	R	CBC skt	R	CB Crf	R	CBC sktr	R	CB Cstr	R	CA NN	R	TA NN	R
iris	96.44	4	**97.55**	1	96.66	3	96.44	4	96.88	2	96.44	4	96.44	4	96.44	4	87.55	8	95.11	7
wine	94.38	6	92.31	7	97.94	3	94.38	6	**98.31**	1	97.75	2	92.31	7	97.75	2	94.57	5	94.75	4
wdbc	93.23	5	90.85	6	**96.75**	1	93.23	5	96.49	2	96.39	3	90.85	6	96.39	3	93.23	5	93.31	4
heart	75.55	6	76.66	5	77.77	4	75.55	6	79.44	3	**82.59**	1	81.11	2	**82.59**	1	75.00	7	74.62	8
banana	87.39	6	89.60	2	88.65	4	87.39	6	**89.85**	1	88.94	3	88.94	3	88.94	3	86.49	7	87.54	5
bupa	65.77	5	66.25	3	62.34	7	65.77	5	67.54	2	**72.61**	1	**72.61**	1	**72.61**	1	65.78	4	64.75	6
sonar	74.15	5	57.69	6	**82.81**	1	74.15	5	79.51	3	82.62	2	57.69	6	82.62	2	74.15	5	74.40	4
Segmentation	95.37	5	94.96	6	96.53	4	95.37	5	96.66	3	**97.34**	1	96.93	2	**97.34**	1	93.46	8	93.67	7
pima	70.89	7	73.03	4	73.95	2	70.89	7	74.40	2	**75.00**	1	**75.00**	1	**75.00**	1	71.28	6	71.54	5
mammographic	73.06	4	62.96	7	74.07	3	73.06	4	**74.67**	1	74.47	2	74.47	2	74.47	2	64.97	6	72.12	5
ionosphere	87.86	7	89.76	3	88.33	6	87.86	7	89.04	4	**93.02**	1	89.76	3	**93.02**	1	91.83	2	88.87	5
cloud	99.26	3	98.04	5	97.36	6	99.26	3	98.77	4	**99.61**	1	99.51	2	**99.61**	1	99.26	3	99.26	3
hayes_roth	84.17	4	65.08	8	77.54	7	84.17	4	82.14	5	85.27	2	85.05	3	85.27	2	78.31	6	**85.41**	1
page-blocks	98.24	2	94.55	6	97.84	5	98.24	2	98.22	3	**98.28**	1	98.24	2	**98.28**	1	98.20	4	98.20	4
thyroid	**99.66**	1	95.30	6	95.73	5	**99.66**	1	97.40	4	99.56	2	99.23	3	99.56	2	**99.66**	1	99.66	1
magic	83.40	6	81.49	7	86.79	2	83.40	6	86.04	3	**89.05**	1	**89.05**	1	**89.05**	1	84.05	4	83.70	5
monk-2	**100**	1	97.21	3	95.59	4	**100**	1	97.21	3	**100**	1	**100**	1	**100**	1	**100**	1	99.64	2
ring	83.59	8	84.60	5	84.10	7	83.59	6	88.40	4	93.09	3	84.60	5	93.09	3	**97.10**	1	96.94	2
twonorm	80.05	4	62.80	7	**96.30**	1	80.05	4	89.20	3	95.85	2	62.80	7	95.85	2	79.95	5	79.90	6
Wall_Following Robot Navigation Data	98.64	2	85.89	7	89.87	6	98.64	2	93.67	5	**98.77**	1	94.92	4	**98.77**	1	98.32	3	98.64	2

[DOI: 10.52547/jsdp.18.2.29]

(جدول-۵): مقایسه صحت رده‌بند جنگل تصادفی بر روش پیشنهادی CBC و سایر روش‌ها به درصد، بر روی بیست مجموعه داده استاندارد و رتبه‌بندی نتایج (R)

(Table-5): Comparison of the accuracy of stochastic forest classification on the proposed CBC method and other methods in percentage, on 20 standard datasets and ranking the results (R)

Dataset	Base	R	CBCsvm	R	CBCknn	R	CBCdt	R	CBCskt	R	CBCrf	R	CBCsktr	R	CBCstr	R	CANN	R	TANN	R
iris	96.44	4	**97.55**	1	96.66	3	96.44	4	96.88	2	96.44	4	96.44	4	96.44	4	87.33	5	97.33	2
wine	97.75	5	93027	7	97.94	4	94.38	6	98.31	2	97.75	5	98.10	3	97.94	4	**98.51**	1	98.10	3
wdbc	96.39	5	91.30	9	96.75	2	93.40	8	96.49	4	96.39	5	**96.84**	1	96.31	6	96.66	3	96.13	7
heart	**82.59**	1	76.66	8	77.77	7	75.55	9	79.44	6	**82.59**	1	80.92	5	81.48	3	81.29	4	82.03	2
banana	88.94	4	89.60	3	88.65	5	87.39	7	**89.85**	1	88.94	4	88.94	4	88.94	4	89.75	2	88.29	6
bupa	**72.61**	1	66.25	5	62.34	7	65.77	6	67.54	4	**72.61**	1	**72.61**	1	**72.61**	1	72.46	2	70.43	3
sonar	82.62	2	63.67	8	62.81	9	74.62	7	79.75	5	82.62	2	81.44	4	79.51	6	**83.82**	1	81.95	3
Segmentation	97.41	2	95.17	7	96.59	5	95.51	6	96.73	4	97.41	2	**97.68**	1	97.41	2	96.73	4	97.14	3
pima	75.00	3	73.03	6	73.95	5	70.89	7	74.40	4	75.00	3	75.00	3	75.00	3	75.77	2	**76.43**	1
mammographic	74.47	3	62.96	7	74.07	4	73.06	6	74.67	2	74.47	3	74.47	3	74.47	3	74.00	5	**75.20**	1
ionosphere	93.02	3	89.90	6	88.33	8	88.00	9	89.18	7	93.02	3	91.46	4	91.18	5	93.71	2	**94.15**	1
cloud	**99.61**	1	98.34	5	97.55	6	99.26	3	98.82	4	**99.61**	1	**99.61**	1	**99.61**	1	99.60	2	**99.61**	1
hayes_roth	85.27	2	65.08	7	77.54	6	84.17	3	82.14	5	85.27	2	85.27	2	85.27	2	83.11	4	**86.48**	1
page-blocks	98.28	2	94.56	7	97.84	6	98.24	4	98.22	5	98.28	2	98.28	2	98.26	3	**98.34**	1	**98.34**	1
thyroid	99.56	3	95.76	8	95.93	7	**99.63**	1	97.46	6	99.65	3	99.60	2	99.56	3	99.36	4	99.33	5
magic	**89.05**	1	81.49	6	86.79	3	83.40	5	86.04	4	**89.05**	1	**89.05**	1	**89.05**	1	88.95	2	88.95	2
monk-2	**100**	1	97.21	3	95.59	4	**100**	1	97.21	3	**100**	1	**100**	1	**100**	1	99.53	2	**100**	1
ring	93.09	3	84.60	8	64.15	9	83.59	5	88.40	6	93.02	3	89.89	4	89.35	7	**97.84**	1	97.44	2
twonorm	95.85	3	62.80	9	**96.30**	1	80.05	8	89.20	6	95.85	3	92.09	3	86.64	7	95.70	4	96.25	2
Wall_Following Robot Navigation Data	**98.77**	1	88.37	7	91.19	6	98.72	2	94.42	5	**98.77**	1	98.67	3	98.72	2	98.62	4	98.62	4

(جدول-۶): مقایسه صحت رده‌بند k نزدیک‌ترین همسایه بر روش پیشنهادی CBC و سایر روش‌ها به درصد، بر روی بیست مجموعه‌داده استاندارد و رتبه‌بندی نتایج (R)

(Table-6): Comparison of the accuracy of the nearest neighbor classifier k on the proposed CBC method and other methods in percentage, on 20 standard datasets and ranking the results (R)

Dataset	Base	R	CBCsvm	R	CBCknn	R	CBCdt	R	CBCskt	R	CBCrf	R	CBCsktr	R	CBCstr	R	CANN	R	TANN	R
iris	96.66	3	**97.33**	1	96.66	3	96.44	4	96.88	2	96.44	4	96.66	3	96.66	3	84.22	5	96.66	
wine	97.94	4	92.31	7	97.94	4	94.38	6	98.31	2	97.75	5	**98.32**	1	98.12	3	**98.32**	1	97.94	
wdbc	96.75	2	91.65	7	96.75	2	93.49	6	96.58	3	96.39	5	**96.84**	1	96.75	2	96.39	5	96.40	
heart	77.77	4	77.59	5	77.77	4	77.40	6	80.00	3	**82.96**	1	81.66	2	81.16	2	73.88	7	77.59	
banana	88.65	5	89.60	2	88.65	5	87.39	7	**89.85**	1	88.94	4	88.99	3	88.94	4	88.29	6	88.65	
bupa	62.34	6	66.25	4	62.34	6	65.77	5	67.54	3	**72.61**	1	72.18	2	**72.61**	1	60.24	7	62.34	
sonar	82.81	3	65.30	8	82.81	3	79.70	7	80.95	5	**84.30**	1	81.15	4	80.68	6	83.38	2	82.81	

[DOI: 10.52547/jsdp.18.2.29]

Segmentation	96.53	5	94.82	8	96.53	5	95.44	6	96.66	4	**97.41**	1	97.34	2	97.27	3	95.30	7	96.66
pima	73.95	4	73.03	5	73.95	4	70.89	7	74.40	2	**75.00**	1	**75.00**	1	**75.00**	1	71.54	6	74.02
mammographic	74.07	5	62.96	8	74.07	5	73.06	6	**74.67**	1	74.47	3	74.54	2	74.54	2	69.18	7	74.14
ionosphere	88.33	8	89.76	5	88.33	8	89.46	6	89.04	7	**92.45**	1	90.74	4	91.16	3	91.89	2	87.91
cloud	97.36	6	98.04	5	97.36	6	99.26	3	98.77	4	**99.61**	1	99.56	2	99.56	2	97.07	8	97.26
hayes_roth	77.54	7	65.08	9	77.54	7	84.17	3	82.14	6	**85.27**	1	85.06	2	**85.27**	1	82.39	5	74.61
page-blocks	97.84	5	94.61	8	97.84	5	98.24	3	98.22	4	98.26	2	**98.28**	1	98.22	4	97.48	7	97.80
thyroid	95.73	9	95.37	7	95.73	6	**99.56**	1	97.40	5	99.36	3	99.10	4	99.40	2	95.33	8	95.73
magic	86.79	3	81.49	7	86.79	3	83.40	6	86.04	4	**89.05**	1	89.00	2	89.00	2	85.89	5	86.79
monk-2	95.59	3	97.21	2	95.59	3	**100**	1	97.21	2	**100**	1	**100**	1	**100**	1	88.63	5	90.49
ring	64.10	9	84.60	5	64.10	9	83.59	7	88.40	6	93.09	4	93.64	3	93.79	2	**97.29**	1	64.25
twonorm	96.30	2	62.80	8	96.30	2	80.05	7	89.20	6	95.85	3	95.25	4	90.45	5	**96.29**	1	96.30
Wall_Following Robot Navigation Data	89.87	7	79.86	9	89.87	7	98.49	2	93.67	5	**98.57**	1	97.09	4	98.44	3	89.07	8	89.92

(جدول-۷): میانگین نتایج صحت رده‌بند ماشین بردار پشتیبان بر روش پیشنهادی CBC و سایر روش‌ها به درصد، بر روی بیست مجموعه‌داده استاندارد و رتبه‌بندی نتایج

(Table-7): Mean Results of Support Vector Machine Classification Accuracy on CBC Proposed Method and Other Methods in Percentage, on 20 Standard Datasets and Results Ranking

نام طبقه‌بند	ماشین بردار پشتیبان									
نام روش	Base	CBCsvm	CBCknn	CBCtree	CBCskt	CBCrf	CBCsktr	CBCstr	CANN	TANN
صحت رده‌بند (%)	82.78	82.94	84.00	84.20	85.05	85.65	85.05	85.26	79.91	82.63
میانگین رتبه هر روش	5	5	5	5	3	2	4	3	7	6
رتبه	8	7	6	5	3	1	4	2	10	9

(جدول- ۸): میانگین نتایج صحت رده‌بند درخت تصمیم بر روش پیشنهادی CBC و سایر روش‌ها به درصد، بر روی بیست مجموعه داده استاندارد و رتبه‌بندی نتایج

(Table-8): Mean Results of Decision Tree Classifier Accuracy Based on Proposed CBC Method and Other Methods Percent, on 20 Standard Datasets and Results Ranking

نام طبقه‌بند	درخت تصمیم									
نام روش	Base	CBCsvm	CBCknn	CBCtree	CBCskt	CBCrf	CBCsktr	CBCstr	CANN	TANN
صحت رده‌بند (%)	87.05	82.82	87.84	87.05	88.69	90.83	86.62	90.83	86.65	87.60
میانگین رتبه هر روش	5	5	4	4	3	2	3	2	5	4
رتبه	7	10	4	6	3	1	9	2	8	5

(جدول-۹): میانگین نتایج صحت رده‌بند جنگل تصادفی بر روش پیشنهادی CBC و سایر روش‌ها به درصد، بر روی بیست مجموعه‌داده استاندارد و رتبه‌بندی نتایج

(Table-9): Mean results of random forest classification accuracy on the proposed CBC method and other methods in percentage, on 20 standard datasets and ranking results

نام طبقه‌بند	جنگل تصادفی									
نام روش	Base	CBCsvm	CBCknn	CBCtree	CBCskt	CBCrf	CBCsktr	CBCstr	CANN	TANN
صحت رده‌بند (%)	90.83	83.37	85.93	87.10	88.75	90.83	90.31	89.88	90.55	91.11
میانگین رتبه هر روش	3	6	5	5	4	3	3	3	3	3
رتبه	3	10	9	8	7	2	5	6	4	1

[DOI: 10.52547/jsdp.18.2.29]

(جدول-۱۰): میانگین نتایج صحت رده‌بند k نزدیک‌ترین همسایه بر روش پیشنهادی CBC و سایر روش‌ها به درصد، بر روی بیست مجموعه‌داده استاندارد و رتبه‌بندی نتایج

(Table-10): Mean results of the nearest neighbor classifier k class on the proposed CBC method and other methods to percentage, on 20 standard datasets and ranking results

نام رده‌بند	k نزدیک‌ترین همسایه									
نام روش	Base	CBCsvm	CBCknn	CBCtree	CBCskt	CBCrf	CBCsktr	CBCstr	CANN	TANN
صحت رده‌بند (%)	86.84	83.32	86.84	87.50	88.79	**90.88**	90.52	90.37	87.10	86.41
میانگین رتبه هر روش	5	6	5	5	4	2	2	3	5	5
رتبه	8	10	7	5	4	1	2	3	6	9

(جدول-۱۱): نتیجه‌گیری نهایی برای انتخاب بهترین روش CBC و رده‌بند کمکی

(Table-11): Final conclusion for selecting the best CBC method and auxiliary classifier

نام رده‌بند	Base	CBCrf	CBCstr	CBCskt	CBCsktr	TANN
ماشین بردار پشتیبان	82.78	**85.65**	85.26	85.05	85.05	82.63
درخت تصمیم	87.05	**90.83**	90.83	88.69	86.62	87.60
جنگل تصادفی	90.83	**90.83**	89.88	88.75	90.31	91.11
K نزدیک‌ترین همسایه	86.84	**90.88**	90.37	88.79	90.52	86.41

۴-۴- ارزیابی بهره اطلاعاتی و شاخص جینی ویژگی‌های حاصل از روش پیشنهادی CBC

در این بخش سعی داریم تا با آزمایش‌های دیگری روی ویژگی‌های افزوده‌شده در روش‌های مختلف CBC، علت بهبود کارایی الگوریتم‌های رده‌بندی را پژوهش کنیم. از آنجا که کارایی یک رده‌بند، رابطه مستقیم با وجود ویژگی‌هایی دارد که برای هدف مورد نظر اطلاعات لازم و کافی را در برداشته باشد؛ لذا به‌کمک روش‌هایی که در پژوهش [22]، ارائه شده است به بررسی میزان اطلاعات موجود در ویژگی‌های اولیه مربوط به مجموعه داده‌ها و ویژگی‌های افزوده‌شده به‌کمک روش CBC می‌پردازیم. بدین منظور از دو معیار بهره اطلاعاتی و شاخص جینی، مطابق با رابطه‌های (۷ و ۶) معرفی‌شده در بخش ۲-۱ استفاده کردیم.

بهره اطلاعاتی [23]، مشخص‌کننده میزان اثرگذاری یک ویژگی با توجه به هدف دسته‌بندی است. به‌عبارت دیگر میزان اطلاعاتی است که یک ویژگی درباره یک رده می‌دهد. بدون تردید، ویژگی نامرتبط با رده داده، هیچ اطلاعاتی را به ما نمی‌دهد. در این روش ویژگی x بر ویژگی y برتری دارد اگر اطلاعات به‌دست‌آمده از ویژگی x بیشتر از اطلاعاتی باشد که از y به‌دست می‌آید.

در جدول (۱۲)، مقادیر بهره اطلاعاتی بر روی ویژگی‌های مجموعه‌داده Iris با چهار ویژگی، طول کاسبرگ، پهنای کاسبرگ، طول گلبرگ، پهنای گلبرگ و ویژگی‌های اضافه‌شده به‌وسیلهٔ روش پیشنهادی CBC آورده شده‌اند. همان‌طور که نتایج نشان می‌دهد میزان اطلاعات موجود در ویژگی‌های اضافه‌شده در روش‌های پیشنهادی مقاله قابل رقابت با ویژگی‌های اولیه مجموعه داده یادشده هستند. به‌عنوان مثال میزان بهره اطلاعاتی ویژگی حاصل از روش CBCrf، مقدار ۷۸۷۳/. است که از بهره اطلاعاتی ویژگی پهنای کاسبرگ با مقدار ۰/۶۴۸۱ بیشتر است.

(جدول-۱۲): میانگین بهره اطلاعاتی بر روی مجموعه‌داده Iris

(Table-12): Mean information gain on the Iris dataset

ویژگی‌ها	میانگین بهره اطلاعاتی
ویژگی طول کاسبرگ	**0.8243**
ویژگی پهنای کاسبرگ	**0.6481**
ویژگی طول گلبرگ	**0.8960**
ویژگی پهنای گلبرگ	**0.8916**
ویژگی استخراج شده از CBCsvm	**0.7914**
ویژگی استخراج شده از CBCknn	**0.7766**
ویژگی استخراج شده از CBCtree	**0.7613**
ویژگی استخراج شده از CBCrf	**0.7873**
ویژگی استخراج شده از CBCskt	**0.7882**

شاخص جینی [24]، میزان ناخالصی یک ویژگی را نسبت به رده‌ها در نظر می‌گیرد. از آنجا که هر چه ناخالصی بیشتر باشد، بی‌نظمی نیز بیشتر است، بنابراین هر چه رتبه به‌دست‌آمده در این شاخص پایین‌تر باشد نتیجه بهتر است.

در جدول (۱۳) نیز مقادیر شاخص جینی بر روی ویژگی‌های مجموعه‌داده Iris و ویژگی‌های اضافه‌شده در روش‌های پیشنهادی CBC آورده شده‌اند.

[DOI: 10.52547/jsdp.18.2.29]

(جدول-۱۳): میانگین شاخص جینی بر روی مجموعه‌داده Iris

(Table-13): Mean Gini index on Iris data set

ویژگی‌ها	میانگین شاخص جینی
ویژگی طول کاسبرگ	**6.66**
ویژگی پهنای کاسبرگ	**7.93**
ویژگی طول گلبرگ	**3.2**
ویژگی پهنای گلبرگ	**4.2**
ویژگی استخراج‌شده از CBCsvm	**2.86**
ویژگی استخراج‌شده از CBCknn	**3.86**
ویژگی استخراج‌شده از CBCtree	**4.63**
ویژگی استخراج‌شده از CBCrf	**5.33**
ویژگی استخراج‌شده از CBCskt	**6.3**

همان‌طور که مشاهده می‌شود، ویژگی‌های پیشنهادی مقاله دارای شاخص جینی بهتری نسبت به ویژگی‌های اولیه مجموعه‌داده Iris است. به‌عنوان مثال شاخص جینی ویژگی حاصل از روش CBCrf، مقدار ۵/۳۲ است که از شاخص جینی ویژگی پهنای کاسبرگ با مقدار ۷/۹۳ کمتر است.

در ادامه جهت بررسی بیشتر، نتایج میزان بهره اطلاعاتی و شاخص جینی بر روی مجموعه‌داده Hayes-roth در جدول‌های (۱۴ و ۱۵) نیز آورده شده است.

در جدول (۱۴)، مقادیر بهره اطلاعاتی بر روی ویژگی‌های مجموعه‌داده Hayes-roth با چهار ویژگی سرگرمی، سن، سطح تحصیلات، وضعیت تأهل و ویژگی‌های اضافه‌شده به‌وسیلهٔ روش پیشنهادی CBC آورده شده‌اند. همان‌طور که نتایج نشان می‌دهد، میزان اطلاعات موجود در ویژگی‌های اضافه‌شده در روش‌های پیشنهادی مقاله قابل رقابت با ویژگی‌های اولیه مجموعه‌داده یادشده هستند. به‌عنوان مثال میزان بهره اطلاعاتی ویژگی حاصل از روش CBCrf، مقدار ۰/۴۵۶۵ است که از بهره اطلاعاتی ویژگی سرگرمی با مقدار ۰/۱۲۱۷ خیلی بیشتر است.

در جدول (۱۵) نیز مقادیر شاخص جینی بر روی ویژگی‌های مجموعه‌داده Hayes-roth و ویژگی‌های اضافه شده در روش‌های پیشنهادی CBC آورده شده‌اند. همان‌طور که مشاهده می‌شود ویژگی‌های پیشنهادی مقاله دارای شاخص جینی بهتری نسبت به ویژگی‌های اولیه مجموعه‌داده Hayes-roth است. به‌عنوان مثال شاخص جینی ویژگی حاصل از روش CBCrf، مقدار ۲/۹۳ است که از شاخص جینی ویژگی سرگرمی با مقدار ۷/۴۰ خیلی کمتر است.

(جدول-۱۴): میانگین بهره اطلاعاتی بر روی مجموعه‌داده Hayes-roth

(Table-14): Mean information gain on Hayes-roth dataset

ویژگی‌ها	میانگین بهره اطلاعاتی
ویژگی سرگرمی	**0.1217**
ویژگی سن	**0.2696**
ویژگی سطح تحصیلات	**0.2639**
ویژگی وضعیت تأهل	**0.2518**
ویژگی استخراج شده از CBCsvm	**0.3005**
ویژگی استخراج شده از CBCknn	**0.2185**
ویژگی استخراج شده از CBCtree	**0.4292**
ویژگی استخراج شده از CBCrf	**0.4565**
ویژگی استخراج شده از CBCskt	**0.3864**

(جدول-۱۵): میانگین شاخص جینی بر روی مجموعه‌داده Hayes-roth

(Table-15): Mean Gini index on Hayes-roth dataset

ویژگی‌ها	میانگین شاخص جینی
ویژگی سرگرمی	**7.40**
ویژگی سن	**5.40**
ویژگی سطح تحصیلات	**6.13**
ویژگی وضعیت تأهل	**5.96**
ویژگی استخراج‌شده از CBCsvm	**4.96**
ویژگی استخراج‌شده از CBCknn	**5.33**
ویژگی استخراج‌شده از CBCtree	**2.30**
ویژگی استخراج‌شده از CBCrf	**2.93**
ویژگی استخراج‌شده از CBCskt	**4.56**

با توجه به نتایج آزمایش‌ها بر روی دیگر مجموعه داده‌ها می‌توان نتیجه گرفت که در تقریباً همه آن‌ها، ویژگی پیشنهادی از یک ویژگی خاص در هر دو مورد یعنی هم از نظر بار اطلاعاتی و هم از نظر شاخص جینی بهتر است.

طبق مقاله [25] و بررسی‌های انجام‌شده در این پژوهش، دو معیار بار اطلاعاتی و شاخص جینی در دو درصد موارد به نتایج متفاوتی می‌رسند؛ بنابراین نمی‌توان تصمیم گرفت که کدام‌یک از این دو آزمایش بهتر انجام می‌شود؛ بنابراین انجام هر دو آزمایش ضرورت پیدا می‌کند؛ اما از نظر سرعت محاسبات شاخص جینی بهتر است زیرا در آن محاسبات لگاریتمی وجود ندارد. از طرف دیگر برای انتخاب ویژگی مناسب‌تر، بررسی هر دو مقدار می‌تواند به تصمیم‌گیری بهتر کمک کند.

۵-۴- کاهش بعد در مجموعه‌داده‌ها

در این بخش، ابتدا ویژگی استخراج‌شده با بیشترین بار اطلاعاتی را به مجموعه داده‌ها اضافه و سپس برای جلوگیری از افزایش بعد، ویژگی‌ای با کمترین بار اطلاعاتی را از مجموعه‌داده‌ها حذف کردیم. نتایج این

[DOI: 10.52547/jsdp.18.2.29]

آزمایش در جدول (۱۶) نشان می‌دهد که می‌توان ویژگی استخراج‌شده را جایگزین ویژگی ضعیف‌تر از نظر بار اطلاعاتی کرد؛ زیرا کارایی عملکرد طبقه‌بندها کاهش نمی‌یابد.

(جدول-۱۶): مقایسه صحت طبقه‌بندی مجموعه‌داده‌ها به‌اضافه ویژگی استخراج‌شده قبل و بعد از حذف ویژگی با بار اطلاعاتی کم

(Table-16): Comparison of the classification accuracy of the data set plus the extracted property before and after the removal of the feature with low information gain

classifier	Datasets	Acc%
SVM	Iris + CBCrf	96.44
	Iris + CBCrf – feature2	96.88
	Wine + CBCrf	91.40
	Wine + CBCrf – feature4	94.22
	Hayes-roth+ CBCrf	85.21
	Hayes-roth + CBCrf – feature1	85.55
Decision Tree	Iris + CBCrf	96.44
	Iris + CBCrf – feature2	96.88
	Wine + CBCrf	97.75
	Wine + CBCrf – feature4	98.48
	Hayes-roth+ CBCrf	85.27
	Hayes-roth + CBCrf – feature1	85.55
Random Forest	Iris + CBCrf	96.44
	Iris + CBCrf – feature2	96.88
	Wine + CBCrf	97.75
	Wine + CBCrf – feature4	98.48
	Hayes-roth+ CBCrf	85.27
	Hayes-roth + CBCrf – feature1	85.55
Knn	Iris + CBCrf	96.44
	Iris + CBCrf – feature2	96.88
	Wine + CBCrf	97.75
	Wine + CBCrf – feature4	98.48
	Hayes-roth+ CBCrf	85.27
	Hayes-roth + CBCrf – feature1	85.55

در مجموعه‌داده Iris، ویژگی دوم، در مجموعه‌داده Wine، ویژگی چهارم و در مجموعه‌داده Hayes-roth، ویژگی نخست دارای کمترین بار اطلاعاتی بودند، بنابراین از آن‌ها برای کاهش بعد استفاده کردیم.

۵- نتیجه‌گیری

در این مقاله بیان شد که هدف از یک سامانه تشخیص الگو، قراردادن الگوها با کمترین خطا، در طبقه مربوط به خودشان است. در راستای چنین هدفی استخراج ویژگی مناسب که باعث تفکیک‌پذیری و تمایز بیشتر رده‌ها از هم می‌شود تأثیر به‌سزایی در بهبود کارایی رده‌بندها دارد. بدین منظور، ما نیز در این پژوهش با ارائه روش پیشنهادی CBC و افزودن ویژگی‌های مناسب به مجموعه‌داده اولیه، یک مجموعه‌داده آموزشی و آزمایشی جدید ساختیم.

در روش پیشنهادی برای ساخت یک مجموعه‌داده جدید، ابتدا از الگوریتم‌های طبقه‌بندی مانند جنگل تصادفی، ماشین بردار پشتیبان، درخت تصمیم و k نزدیک‌ترین همسایه، به‌عنوان یک رده‌بند کمکی برای پیش‌گویی برچسب نمونه‌های آزمایشی استفاده کرده و صحت رده‌بندی را نیز محاسبه و سپس برچسب‌های پیش‌گویی‌شده را به‌عنوان یک ویژگی جدید به داده‌های آزمایشی اضافه و در مرحله بعدی نیز، برچسب درست نمونه‌های آموزشی را به‌عنوان یک ویژگی به نمونه‌های آموزشی اضافه کردیم. پس از آماده‌شدن مجموعه‌داده جدید، دوباره طبقه‌بندهای بالا را بر آن اعمال و صحت رده بندی را محاسبه کرده‌ایم؛ سپس با مقایسه صحت رده بندی روی دو مجموعه‌داده اولیه و جدید، نشان دادیم که کارایی الگوریتم‌های طبقه‌بندی ماشین بردار پشتیبان، درخت تصمیم و k نزدیک‌ترین همسایه بر روی مجموعه داده جدید، حدود چهار درصد افزایش می‌یابد. این نتیجه به این معنی است که ویژگی‌های اضافه‌شده در روش پیشنهادی در بیشتر موارد، ویژگی‌های مناسب با قدرت تمایز خوبی بوده‌اند، به‌خصوص در روشی که از طبقه‌بند جنگل تصادفی به‌عنوان طبقه‌بند کمکی برای پیش‌گویی برچسب نمونه‌های آزمایشی استفاده شده است.

در عین حال برای بررسی خوب‌بودن ویژگی‌های به دست‌آمده، میزان اطلاعات موجود در ویژگی‌های افزوده‌شده و نیز میزان خلوص ویژگی‌ها را با دو معیار بهره اطلاعاتی و شاخص جینی به‌ترتیب اندازه گرفتیم. نتایج در این قسمت نیز نشان می‌دهد که ویژگی‌های به‌دست‌آمده در روش پیشنهادی در بیشتر موارد از مقادیر مناسب و خوبی نسبت به ویژگی‌های اولیه مجموعه‌داده برخوردار است؛ بنابراین ویژگی استخراج‌شده با بیشترین بار اطلاعاتی را جایگزین ویژگی با بار اطلاعاتی کم کرده و از این طریق هم از افزایش بعد جلوگیری کرده و هم به نتایج بهتری در میزان کارایی الگوریتم‌های رده‌بند دست یافتیم.

6- References ۶- مراجع

[1] S. J.Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia, Pearson Education Limited, 2016.

[2] R. P. Duin and D. M. J. Tax, "Statistical pattern recognition", In *Handbook of Pattern Recognition and Computer Vision*, pp. 3-24, 2005.

[3] Guyon, S. Gunn, M.Nikravesh, L. A. Zadeh, and editors, *Feature extraction: foundations and applications*, Vol. 207. Springer, 2008.

[4] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21-27, 1967.

[DOI: 10.52547/jsdp.18.2.29]

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995.

[6] J. Showe-Taylor and N. Christianini, *Support vector machines and other kernel-based learning methods*, 2000.

[7] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[8] B. W. Silverman and M. C. Jones, "An important contribution to nonparametric discriminant analysis and density estimation," *International statistical review/revue Internationale de statistique*, pp. 233-238, 1989.

[9] R. C.Barros, M. P. Basgalupp, A. C. De Carvalho, and A. A. Freitas, "A survey of evolutionary algorithms for decision-tree induction," *IEEETransactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no 3, pp. 291-312, 2012.

[10] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[11] M.Woźniak, M.GrañaandE. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, vol. 16, pp. 3-17, 2014.

[12] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[13] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vo. 36, no. 3, pp. 287–314, 1994.

[14] K. Fukunaga, "Introduction to Statistical Pattern Recognition," *San Diego*: *Academic Press Inc*, 1990.

[15] C. F. Tsai and C. Y Lin, "A triangle area based nearest neighbors approach to intrusion detection," *Pattern recognition*. vol. 43, no. 1, pp. 222-229, 2010.

[16] W. C.Lin, S. W. Ke and C. F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors", *Knowledge-based systems*, no. *78*, pp. 13-21, 2015.

[17] X. Wang, C. Zhang and K. Zheng, "Intrusion detection algorithm based on density, cluster centers, and nearest neighbors", *China Communications*, vol. 13, no. 7, pp. 24-31, 2016.

[18] A. Asuncion and D. J. Newman, UCI Machine Learning Repository, University of California, 2007.

https://archive.ics.uci.edu/ml/index.php

[19] C. W. Hsua and C. J. Lin, "A comparison of methods for multiclass support vector machines*", IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415-425, 2002.

[20] T. T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", *Pattern Recognition*, vol. 48, no. 9, pp. 2839-2846, 2015.

[21] J. T. Townsend, "Theoretical analysis of an alphabetic confusion matrix", *Perception & Psychophysics*, vol. 9, no. 1, pp. 40-50, 1971.

[22] M. Dash and H. Liu, "Consistency-based search in feature selection", *Artificial intelligence*, vol. 151, no. 1-2, pp. 155-176, 2003.

[23] J. R. Quinlan, "Induction of decision trees", *Machine learning*, vol. 1, no. 1, pp. 81-106, 1986.

[24] L. Breiman, "Classification and regression trees".*Routledge*, 2017.

[25] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria", *Annals of Mathematics and Artificial Intelligence*, vol. 41, no. 1, pp.77-93, 2004.

حمیدرضا غفاری تحصیلات خود را در مقطع کارشناسی در رشته رایانه در دانشگاه صنعتی شریف و کارشناسی ارشد را در دانشگاه تهران جنوب و دکترای خود را در دانشگاه فردوسی به پایان رساند. وی هم‌اکنون عضو هیئت علمی و استادیار دانشکده مهندسی رایانه دانشگاه آزاد اسلامی فردوس است. موضوعات مورد علاقه ایشان یادگیری ماشین، شناسایی الگو و پردازش تصویر است.

نشانی رایانامه ایشان عبارت است از:

hghaffaripaper@ferdowsiau.ac.ir

آتنا جلالی مجاهد تحصیلات خود را در مقطع کارشناسی در دانشگاه بیرجند و در مقطع ارشد در دانشگاه آزاد اسلامی مشهد به پایان رساند. وی هم‌اکنون دانشجوی دکترا در دانشکده مهندسی رایانه دانشگاه آزاد اسلامی فردوس است. موضوعات مورد علاقه ایشان یادگیری ماشین و شناسایی الگو است.

نشانی رایانامه ایشان عبارت است از:

st.ajalalia@ferdowsiau.ac.ir

[DOI: 10.52547/jsdp.18.2.29]