Journal of Statistical Sciences, Autumn and Winter, 2022
Vol. 16, No. 2, pp 331-348
DOI: 10.29252/jss.16.2.331

Introducing a New Method for the Split Criteria of Decision Trees

Chaji, A.
Department of Electrical Engineering, Shohadaye Hoveizeh Campus of Technology, Shahid Chamran University of Ahvaz, Ahvaz, Iran.

Corresponding author: A. Chaji, a.chaji@scu.ac.ir

Received: 1/5/2022 **Revised:** 9/11/2022 **Accepted and Published Online:** 13/11/2022.

Introduction

A decision tree is a flow chart-like graph structure of decision nodes, branches, and leaf nodes starting from the root and ending at the leaves. The node is the independent variable on which the test is performed, and the root node is placed at the top of the tree, while the leaf is the dependent variable (answer) or category label, which is the last node of the tree. The decision at each stage of tree construction depends on the previous branching operation, which is crucial to the predictive capability of the tree. The branching method of information gain, by using the concept of entropy, measures how much information a feature provides about a class and decides to split the tree at each node.

Material and Methods

The first dataset describes the diagnosis of cardiac Single Proton Emission Computed Tomography (SPECT) image contains 267 SPECT image sets (patients), and 22 Attributes such that each of the patients is classified into two categories: normal and abnormal. The second data set contains 1024 binary attributes (molecular fingerprints) used to classify 1687 chemicals into two classes (binder to androgen receptor/positive, non-binder to androgen receptor /negative). Also, a real-world dataset used that contains 90 instances and 7 attributes include gender, blood pressure, blood sugar, cholesterol, smoking, weight and occupation of the patient, which were collected to predict the determination of the treatment method (medical treatment or angiography). A new approach is proposed to produce a decision tree based

on the T-entropy criterion. The method applied on the three datasets, examined by 11 evaluation criteria and compared with the well-known methods of Gini index, Shannon, Tsallis, and Renyi entropies for splitting the decision tree, with a proportion of 75 % for training and 25 % for testing and for 300 times of execution (each time of execution leads to the production of a decision tree). Also, a comparison is made between T-entropy and other discussed splitting measures in terms of the area under the ROC curve (AUC).

Results and Discussion
The performance of splitting methods of the Gini index, Shannon, Tsallis, Renyi entropies and T-entropy are examined. The evaluation criteria of accuracy (A), sensitivity, specificity, positive predictive value (PPV)(or precision), F-S ore index (F\), negative predictive value (NPV), false discovery rate (F R), false positive rate (FPR), false negative rate (FNR) and Mean square error for the three data sets were calculated. The maximum values for the first six criteria and the minimum values for the second four criteria indicate the better performance of the decision tree based on the introduction method. Also, the AUC value of the t-entropy method presented for all three data sets is higher than other methods, which indicates that the value of the true positive rate is higher than the false positive rate in the t-entropy method compared to other methods.

Conclusion
The results suggest that the proposed node-splitting method based on T-entropy has better behaviour than the other discussed methods for both low and high numbers of samples. Also, today because of the increasing growth of the big data problem and, on the other hand, the superiority of the T-entropy splitting method over the other investigated methods on different sizes of the dataset, the benefit of this method is twofold.

Keywords: Decision tree, Entropy, T-entropy, Splitting method, Evaluation criteria.

Mathematics Subject Classification (2010): 94A17, 54C70.

©The Author(s). The Publisher is Iranian Statistical Society.
This is an open access article distributed under the terms and conditions of (CC BY-NC 4.0)

مجله علوم آماری، پاییز و زمستان ۱۴۰۱
جلد ۱۶، شماره ۲، ص ۳۳۱ – ۳۴۸
DOI: 10.29252/jss.16.2.331
مقاله پژوهشی

معرفی روشی جدید برای نقطه انشعاب در تولید درختان تصمیم

علیرضا چاجی

گروه برق، دانشگاه شهید چمران اهواز–پردیس صنعتی شهدای هویزه

نویسنده مسئول : نام نویسنده مسئول مقاله، a.chaji@scu.ac.ir

تاریخ دریافت: ۱۴۰۱/۲/۱۱ تاریخ بازنگری: ۱۴۰۱/۸/۱۸ تاریخ پذیرش و انتشار: ۱۴۰۱/۸/۲۲

چکیده: قابلیت تفسیر پذیری بالا و سادگی فهم درختان تصمیم، آنها را به یکی از پرکاربردترین الگوریتم‌های یادگیری ماشین تبدیل کرده است. موضوع کلیدی در ساخت درختان تصمیم کارامد و موثر، بکارگیری روش انشعاب مناسب است. این مقاله یک روش انشعاب جدید جهت تولید درخت مبتنی بر معیار تی– آنتروپی برای نقطه انشعاب پیشنهاد می‌کند. روش ارایه شده روی سه مجموعه داده توسط ۱۱ معیار ارزیابی، مورد بررسی قرارگرفته است. نتایج نشان می‌دهد که روش معرفی شده در ساخت درخت تصمیم نسبت به روش‌های معروف شاخص جینی، آنتروپی‌های شانون، تیسالیس و رنی عملکرد دقیقتری دارد و می‌تواند به عنوان روش جایگزین در تولید درخت تصمیم مورد استفاده قرار گیرد.

واژه‌های کلیدی: درخت تصمیم، آنتروپی، تی–آنتروپی، روش انشعاب، معیار ارزیابی .
کد موضوع‌بندی ریاضی (۲۰۱۰): 94A17، 54C70.

۱ مقدمه

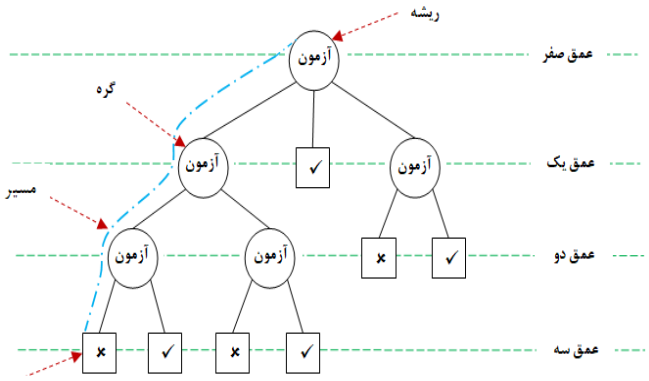
درختان تصمیم یکی از طبقه‌بندهای محبوب برای تحلیل کلان داده‌ها هستند (لیو و همکاران، ۲۰۱۶) و الگویی برای بیان مفاهیم علمی (کشف دانش) فراهم می‌نمایند (مک کلین، ۲۰۰۳). درخت تصمیم یک مدل پیش‌بینی کننده است که از یک گراف درختی، برای نگاشت داده‌های مشاهده شده در یک موضوع تا نتیجه‌گیری مقدار هدف استفاده می‌نماید (لیو و همکاران، ۲۰۱۶). برای مسائلی که دارای منابع اطلاعاتی محدود هستند مانند تخمین ژست انسان

©نویسند(گان). ناشر انجمن آمار ایران است.
این مقاله با دسترسی آزاد تحت شرایط و ضوابط (CC BY-NC 4.0) توزیع شده است.

در زمان واقعی (شوتن و همکاران، ۲۰۱۱) و یا مسائلی که نمونه‌های آموزشی نامتعادل و کم دارند از قبیل زمان ردیابی یک شی متحرک (ژانگ و همکاران، ۲۰۱۷)، درختان تصمیم مناسب هستند. درختان تصمیم اثر خود را روی انواع مسایل طبقه‌بندی (باروز و همکاران، ۲۰۱۲)، اتصال رکوردی (آقامحمدی و رضایی قهرودی، ۱۴۰۱) و بینایی ماشین (شوتن و همکاران، ۲۰۱۱؛ ژانگ و همکاران، ۲۰۱۷) نشان داده‌اند. همچنین به کاربردهای درخت تصمیم در منابع طبیعی و آبخیزداری (لوسه و همکاران، ۲۰۱۹)، نظامی (کلوسه و همکاران، ۲۰۲۱)، پزشکی (عاشوری، ۱۳۹۶؛ شنبه زاده و همکاران، ۲۰۲۲) و کسب و کار و بازاریابی می‌توان اشاره نمود. ساخت یک درخت تصمیم کارا و موثر به علت سادگی و انعطاف‌پذیری آن یک موضوع کلیدی در یادگیری ماشین است (ونگ و چیا، ۲۰۱۷). مساله یافتن درخت تصمیم بهینه، با الگوریتم‌های شناخته شده قابل اجرا در زمان چندجمله‌ای امکان‌پذیر نیست، (هیفیل و ریوست، ۱۹۷۶) لذا رویکرد مرسوم یافتن درخت تصمیم، جستجوی کامل درخت با یک روش آموزشی سطح به سطح حریصانه برای پیدا کردن نقطه انشعاب است (بریمن و همکاران، ۱۹۸۴).

۲ درخت تصمیم

درخت تصمیم ساختاری گراف مانند است که از ریشه آغاز شده و در برگ‌ها خاتمه می‌یابد. گره متغیر مستقلی است که آزمون روی آن انجام می‌شود و گره ریشه در بالاترین نقطه درخت قرار می‌گیرد در حالی‌که برگ متغیر وابسته (پاسخ) یا برچسب دسته و آخرین گره از درخت است. انشعاب به مفهوم تفکیک یک گره بر مبنای شرایط آزمون به تعدادی زیر گره است و انشعاب گره با توجه به شرایط آزمون همان گره انجام می‌گیرد و هر مسیر از ریشه تا برگ یک تصمیم یا نتیجه آزمون است (شکل ۱). قیود شرطی این مسیر یک قانون و برگ‌های درخت مبین کلاس‌ها یا



شکل ۱. ساختار یک درخت تصمیم با برچسب دسته دودویی

توزیع‌های کلاس هستند. هدف از ساخت درخت تصمیم، پیش‌بینی متغیر هدف بر مبنای متغیر ورودی با یادگیری قوانین تصمیم‌گیری است (شوبها و رنگازومی، ۲۰۱۸). فرآیند ساخت درخت تصمیم زمانی خاتمه می‌یابد که هر زیر

مجموعه شامل نمونه داده‌های متعلق به یک کلاس واحد باشد. در عمل وقتی درخت رشد می‌نماید، ابتدا با مجموعه کوچکی از داده‌ها یک درخت اولیه ساخته می‌شود، سپس از این درخت برای طبقه‌بندی داده‌های باقی‌مانده در مجموعه آموزشی استفاده می‌گردد. اگر داده‌ها به درستی طبقه‌بندی شوند، درخت رضایت‌بخش خواهد بود. چنانچه مواردی از طبقه‌بندی نادرست وجود داشته باشند، این موارد به مجموعه کاری اضافه شده و یک درخت جدید بر مبنای این مجموعه افزوده شده، ساخته می شود. این فرآیند به طور مکرر انجام می‌گیرد تا زمانی که درخت تصمیم رضایت‌بخش به‌دست آید. شایان ذکر است، مشکل بیش‌برازش مدل که منجر به از دست رفتن قابلیت پیش‌بینی آن می‌گردد، ممکن است با هرس کردن درخت، که شامل جایگزینی درختان فرعی (زیر درخت) با برگ‌ها است، برطرف گردد (مک کلین، ۲۰۰۳). لذا ساخت درخت با گره ریشه شروع شده، با اجرای آزمون و حرکت روی مسیر ادامه می‌یابد و اجرای آزمون تا زمانی تکرار می‌شود تا به گره برگ یا پایان برسد.

اگر متغیر هدف درخت تصمیم، مجموعه متناهی از مقادیر بگیرد، بعنوان درخت طبقه‌بند شناخته می‌شود در حالی که اگر متغیر هدف از مجموعه پیوسته، مقدار بگیرد، درخت رگرسیون نامیده می‌شود. درختان تصمیم روی مجموعه داده‌های بزرگ صحت پایین‌تری دارند، اما اساساً مقرون به صرفه هستند (ههن و همکاران، ۲۰۲۰). در جدول ۱ مزایا و معایب درختان تصمیم بیان شده است (شوبها و رنگازومی، ۲۰۱۸؛ نیسبت و همکاران، ۲۰۱۸).

جدول ۱. مزایا و معایب درخت تصمیم

ردیف	مزایا	معایب
۱	سادگی درک، تفسیر و تجسم درختان تصمیم که موجب تسهیل بیان مدل می‌گردد و برخلاف برخی الگوریتم دیگر مانند شبکه عصبی دارای جعبه سیاه نیست.	درخت تصمیم بعلت عدم تعمیم خوب داده‌ها مستعد بیش‌برازش است (برخی روش‌های هرس می‌تواند این مشکل را کاهش دهد).
۲	قابلیت اجرا بروی هر دو نوع متغیر طبقه‌ای و عددی را دارد.	بکارگیری متغیرهای طبقه‌ای با سطوح چندگانه، حصول بهره اطلاعات و سطوح بیشتر را به همراه دارد.
۳	درخت تصمیم در پیش‌بینی متغیر هدف در مدل‌های غیرخطی کاراست.	الگوریتم حریصانه ممکن است درخت تصمیم بهینه کلی را نتیجه ندهد. برای رفع این ایراد، نمونه‌ها بطور تصادفی برای آموزش چندین درخت، انتخاب می‌شوند.
۴	سریع آموزش می‌بینند.	ساخت درخت با دست نیاز به محاسبات دارد.
۵	امکان استخراج مجموعه قواعد از درخت آموزش‌دیده نهایی به کمک الگوریتم‌های C۴.۵ و C۵ وجود دارد.	تغییر کوچک در داده‌ها منجر به درختی متفاوت گردیده و این ناپایداری در درخت تصمیم را نشان می‌دهد که این مشکل به کمک روش‌های تقویت و بسته‌بندی کاهش می‌یابد.
۶	بطور ضمنی انتخاب ویژگی متغیر را انجام می‌دهد.	در طبقه‌بندی چندین کلاس خروجی مشکل دارند.
۷	درختان تصمیم نسبت به اتخاذ سناریوهای جدید انعطاف‌پذیری بالایی دارند و تکنیک‌های تصمیم‌گیری بروی درختان تصمیم قابل ترکیب هستند.	تشکیل درخت مغرضانه وقتی برخی کلاس‌ها غالب هستند، اتفاق می‌افتد. در این حالت از روش‌های متعادل‌سازی مجموعه داده آموزش مانند نمونه‌گیری مجدد، استخراج ویژگی و تکنیک یادگیری حساس به هزینه استفاده می‌شود. (هایجانگ و همکاران، ۲۰۱۷).
۸	آماده‌سازی داده‌ها توسط کاربر انجام می‌شود و افزودن نظر و ترجیحات متخصص در درخت تصمیم آسان است.	—

۲.۱ روش‌های انشعاب

انتخاب اصلی در الگوریتم‌های درخت، یافتن بهترین انشعاب است بطوری‌که هر انشعاب نمونه را به دو بخش یا بیشتر تقسیم می‌نماید و هر بخش می‌تواند یک کلاس یا بیشتر را شامل شود. انشعاب گره جزء اصلی‌ترین موضوعات ساختار درخت تصمیم است (کوتسیس، ۲۰۱۳). مطالعه منابع گسترده‌تر معیارهای انشعاب توسط صفویان و لندگرب انجام شده است (صفویان و لندگرب، ۱۹۹۱). روش‌های انشعاب متعددی برای درختان تصمیم وجود دارد که براساس نوع متغیر هدف به دو دسته تقسیم می‌گردند (جدول ۲). برای انشعاب در درختان تصمیم با هدف طبقه‌بندی روش‌های گوناگونی بر مبنای آنتروپی ارایه شده‌اند (شارما و همکاران، ۲۰۱۳؛ ونگ و همکاران، ۲۰۱۷؛ ردی و چیتنینی، ۲۰۲۱). در جدول ۲، p_i احتمال (فراوانی نسبی) کلاس C_i در زیرافراز حاصل از افرازبندی مجموعه کل داده‌ها بر اساس نقطه شکست تعیین شده و α و q برترتیب پارامترهای آنتروپی رنی و تیسالیس هستند. همچنین Actual مقدار واقعی مشاهده شده و Expected مقدار پیش‌بینی شده به کمک درخت تصمیم است. انواع مختلفی از درختان تصمیم تاکنون توسط محققین معرفی شده است. جدول ۳ درختان تصمیم معروف و روش انشعاب آن‌ها را نشان می‌دهد. در ادامه روش یافتن نقطه انشعاب بهینه در یک گره بیان شده است. فرض کنید بخواهیم مجموعه D به r

جدول ۲. انواع روش‌های انشعاب

نوع متغیر هدف	روش انشعاب	روش کار	معیار
پیوسته	کاهش واریانس	استفاده از واریانس	$Var = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$
	شاخص جینی	استفاده از شاخص جینی	$GI = 1 - \sum_{i=1}^{n} p_i^2$
طبقه‌ای	بهره اطلاعات	استفاده از آنتروپی	$\begin{cases} \text{Shanon Entrop} = \sum_{i=1}^{n} p_i \log p_i \\ \text{Renyi Entrop} = \frac{1}{1-\alpha}\log(\sum_{i=1}^{n} p_i^{\alpha}) \\ \text{Tsallis Entrop} = \frac{1}{q-1}(1 - \sum_{i=1}^{n} p_i^{q}) \end{cases}$
	مربع کای	استفاده از والد و فرزند	$\chi^2 = \sum \frac{(\text{Actual}-\text{Expected})^2}{\text{Expected}}$

زیر مجموعه $s(D) = (D_1, \ldots, D_r)$ افراز شود. ابتدا معیار (گرابچوسکی، ۲۰۱۴)

$$\triangle I(s, D) = I(D) - \sum_{i=1}^{r} \frac{|D_i|}{|D|} I(D_i) \qquad (1)$$

تحت عنوان معیار خلوص در نظر گرفته می‌شود، که در آن $|D|$ و $|D_i|$ بیانگر تعداد اعضای مجموعه‌های D و D_i ، s بیانگر افراز (شکاف) مجموعه D و تابع $I(\cdot)$ براساس معیار انشعابی که در نظر می‌گیریم متفاوت است. با

جدول ۳. درختان تصمیم معروف و پرکاربرد

ردیف	نام درخت (نام اختصاری)	مرجع	روش انشعاب
۱	Dichotomiser Iterative۳ (ID3)	(کوییلن، ۱۹۷۹)	بهره اطلاعات
۲	Detection Interaction Automatic Chi-squared (CHAID)	(کاس، ۱۹۸۰)	مربع کای
۳	Tree Regression And Classification (CART)	(بریمن و همکاران، ۱۹۸۴)	شاخص جینی
۴	C۴.۵ جانشین ID۳ وشناخته شده با نام C۵.۰	(کوییلن، ۱۹۹۴)	بهره اطلاعات
۵	Trees decision of INndution PaRallelizable Scalable (SPRINT)	(شافر و همکاران، ۱۹۹۶)	شاخص جینی
۶	Quest In Learning Supervised (SLIQ)	(مهتا و همکاران، ۱۹۹۶)	شاخص جینی
۷	Ctree	(هوتورن و همکاران، ۲۰۰۶)	کاهش واریانس

توجه به جداول ۳ برای معیار آنتروپی شانون و جینی روابط

$$I_E(D) = -\sum_{i=۱}^{r} P(c_i|D)\log P(c_i|D)$$
$$I_G(D) = ۱ - \sum_{i=۱}^{r} P^۲(c_i|D).$$

به‌دست می‌آید. وقتی به ازای مجموعه‌های مختلف $D_۱, \ldots, D_r$ رابطه (۱) محاسبه می‌شود، بعد از یک جستجوی کامل روی تمام متغیرهایی که امکان انشعاب دارند، بهترین نقطه انشعاب برای مجموعه D، انشعابی است که $\triangle I(s, D)$ را ماکسیمم کند. فرض کنید یک گره دارای r کلاس و P انشعاب است که به صورت

	$c_۱$	$\ldots$	c_r	
۱	$n^۱_{c_۱}$	$\ldots$	$n^۱_{c_r}$	$n^۱$
۲	$n^۲_{c_۱}$	$\ldots$	$n^۲_{c_r}$	$n^۲$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p	$n^p_{c_۱}$	$\ldots$	$n^p_{c_r}$	n^p
	$n_{c_۱}$	$\ldots$	n_{c_r}	n

نشان دادهمی‌شود، که در آن $n^j_{c_i}$ بیانگر تعداد اعضای مجموعه D واقع در کلاس c_i و انشعاب jام است، بنابراین

$$I_E(D) = -\sum_{i=۱}^{r} \frac{n_{c_i}}{n}\log\frac{n_{c_i}}{n} = \log n - \frac{۱}{n}\sum_{i=۱}^{r} n_{c_i}\log n_{c_i}.$$

با توجه به رابطه (۱) وقتی معیار، آنتروپی شانون است:

$$\triangle I_E(s,D) = \log n + \frac{۱}{n}\big(- \sum_{i=۱}^{r} n_{c_i} \log n_{c_i} - \sum_{j=۱}^{p} n^j \log n^j + \sum_{j=۱}^{p}\sum_{i=۱}^{r} n^j_{c_i} \log n^j_{c_i}\big)$$

و بهترین نقطه انشعاب وقتی بدست می‌آید که $\triangle I_E(s,D)$ ماکسیمم شود.

۳ روش کار

درخت تصمیم یک تکنیک یادگیری ماشین برای پیش‌بینی یا تصمیم‌گیری است که با تقسیم مکرر داده آموزشی به نمونه‌های کوچک و کوچک‌تر ساخته می‌شود. روش تقسیم داده‌ها روی دقت و صحت درخت تصمیم تاثیرگذار است. تصمیم در هر مرحله از ساخت درخت به عملیات انشعاب قبلی وابسته است (شوبها و رنگازومی، ۲۰۱۸) و در نحوه پیش‌بینی/طبقه‌بندی درخت حیاتی است (مک کلین، ۲۰۰۳) . روش انشعاب بهره اطلاعات با بهره‌گیری از مفهوم آنتروپی، تصمیم به تقسیم درخت در هر گره می‌نماید. در این مقاله تی-آنتروپی (چاکرابورتی و همکاران، ۲۰۲۱) بعنوان یک روش انشعاب به صورت

$$\text{T-Entropy} = \sum_i p_i \tan^{-۱}(\frac{۱}{{p_i}^t}) - \frac{\pi}{۴}, \quad c > ۰$$

معرفی گردیده است، که در آن t پارامتری مثبت است و تابع $I(۰)$ برای این معیار به صورت

$$I_T(D) = -\sum_{i=۱}^{r} P(c_i|D) \tan^{-۱}(\frac{۱}{P^t(c_i|D)}) - \frac{\pi}{۴} = -\sum_{i=۱}^{r} \frac{n_{c_i}}{n} \tan^{-۱}\{(\frac{n_{c_i}}{n})^{-t}\} - \frac{\pi}{۴}$$

است و با توجه به رابطه (۱) معیار خلوص

$$\triangle I_T(s,D) = -\sum_{i=۱}^{r} \frac{n_{c_i}}{n} \tan^{-۱}\{(\frac{n_{c_i}}{n})^{-t}\} - \frac{\pi}{۴} - (\sum_{j=۱}^{p} \frac{n^j}{n} \sum_{i=۱}^{r}[\frac{n^j_{c_i}}{n^j} \tan^{-۱}\{(\frac{n^j_{c_i}}{n^j})^{-t}\} - \frac{\pi}{۴}]).$$

را داریم. چون مقدار $\frac{\pi}{۴}$ ثابت و در همه موارد جستجو وجود دارد می‌تواند حذف شود، در این صورت

$$\triangle I_T(s,D) = -\sum_{i=۱}^{r} \frac{n_{c_i}}{n} \tan^{-۱}\{(\frac{n_{c_i}}{n})^{-t}\} - (\frac{۱}{n} \sum_{j=۱}^{p}\sum_{i=۱}^{r}[n^j_{c_i} \tan^{-۱}\{(\frac{n^j_{c_i}}{n^j})^{-t}\}]).$$

الگوریتم ۱. .

گام ۱- درک داده و آماده‌سازی آن: در این گام سه مجموعه داده استاندارد که دو مجموعه داده از سه مجموعه داده انتخاب شده، از سایت UCI (به صورت دسترسی باز) و مجموعه داده دیگر با مراجعه به بیمارستان خاتم الانبیاء (ع) و بیمارستان علی ابیطالب(ع) زاهدان جمع‌آوری گردیده و مورد استفاده قرار گرفته‌اند. مقادیر این مجموعه داده‌ها به صورت دودویی بوده و در این مجموعه از داده‌ها، مقدار گم شده وجود نداشته است.

گام ۲- مدل‌سازی: در این مرحله درختان تصمیم بر اساس روش‌های انشعاب معروف (جدول ۲) از قبیل شاخص جینی، آنتروپی‌های شانون، تیسالیس، رنی و تی-آنتروپی بر روی مجموعه داده‌های مورد نظر و به نسبت ۷۵ درصد برای آموزش و ۲۵ درصد برای آزمون و به تعداد ۳۰۰ مرتبه اجرا (هر مرتبه اجرا منجر به تولید یک درخت تصمیم می‌شود)، ساخته شده‌اند. زبان برنامه‌نویسی پایتون نسخه ۳.۵ بوده است[1] .

گام ۳- ارزیابی: برای سنجش عملکرد روش‌های انشعاب شاخص جینی، آنتروپی‌های شانون، تیسالیس، رنی و روش پیشنهادی این مقاله (تی- آنتروپی)، شاخص‌های حساسیت یا نرخ مثبت درست (TPR)، شفافیت یا نرخ منفی درست (TNR)، دقت یا ارزش اخباری مثبت (PPV)، صحت (ACC)، شاخص F۱ (F_Score) ، ارزش اخباری منفی (NPV)، نرخ کشف غلط ، (FDR) نرخ مثبت غلط ، (FPR) نرخ منفی غلط (FNR) و میانگین توان‌های دوم خطا محاسبه شدند:

TP = تعداد داده‌های برچسب مثبتی که درست دسته‌بندی شده‌اند،

FP = تعداد داده‌های برچسب منفی که به نادرست مثبت دسته‌بندی شده‌اند،

FN = تعداد داده‌های برچسب مثبتی که به نادرست منفی دسته‌بندی شده‌اند،

TN = تعداد داده‌های برچسب منفی که درست دسته‌بندی شده‌اند.

$$TPR(True\ Positive\ Rate\ or\ Sensitivity\ or\ Recall) = \frac{TP}{TP+FN}$$
$$TNR(True\ Negative\ Rate\ or\ Specificity) = \frac{TN}{FP+TN}$$
$$PPV(PositivePredictiveValueorPrecision) = \frac{TP}{TP+FP}$$
$$ACC(Accuracy) = \frac{TP+TN}{TP+FN+FP+TN}$$
$$F۱(F_Score) = \frac{۲\times TP}{۲\times TP+FN+FP}$$
$$NPV(NegativePredictiveValue) = \frac{TN}{TN+FN}$$
$$FDR(FalseDiscoveryRate) = \frac{FP}{FP+TP} = ۱ - PPV$$
$$FPR(False\ Positive\ Rate\ or\ Fall-out) = \frac{FP}{FP+TN} = ۱ - TNR$$
$$FNR(Rate\ Miss\ or\ Rate\ Negative\ False) = \frac{FN}{TP+FN} = ۱ - TPR$$
$$MSE(Error\ Squared\ Mean) = \frac{\sum_{i=۱}^{n}(x_i - x_i^{'})^۲}{n}$$

در رابطه MSE مقادیر $x_i^{'}$ پیش‌بینی و x_i مقادیر واقعی هستند.

[1] https://github.com/Chajialireza/D-T_t-entropy.git

۴ نتایج

روش‌های مختلفی برای ارزیابی عملکرد طبقه‌بندی وجود دارند. صحت، ماتریس اغتشاش، دقت و حساسیت معیارهای ارزیابی پرکاربرد هستند (شوبها و رنگازومی، ۲۰۱۸). در این بخش به بررسی نتایج روش پیشنهادی بر مبنای الگوریتم ۱ روی سه مجموعه داده منتخب می‌پردازیم. از آنجایی‌که رویکرد مرسوم در یافتن درخت تصمیم، جستجوی کامل درخت برای پیدا کردن نقطه انشعاب است، مقادیر بهینه برای پارامترهای q ، α و t با جستجوی کامل در فضای حالت $\{۱, \ldots, ۳۰۰\}$ که برای α و q مقادیر از ۲ شروع شده‌اند، به‌دست آمده‌اند.

۴.۱ مجموعه داده اول

این مجموعه (دعا و گرف، ۲۰۱۹؛ کورگان و همکاران، ۲۰۱۹) شامل ۲۶۷ داده و ۲۲ متغیر است که در سال ۲۰۰۱ در سایت UCI قرار گرفته است.[1] این مجموعه داده در برگیرنده تصاویر دو بعدی و سه بعدی و داده‌های بالینی بیماران شامل : کد اختصاصی بیمار، جنسیت، وزن، قد، زمان، ۲۲ تشخیص جزیی و یک تشخیص کلی برای هر بیمار بصورت فایل متنی است. نتایج ارزیابی این مجموعه داده در جدول ۴ و شکل ۲(الف) نشان داده شده‌اند. نتایج حاصل نشان می‌دهد که تی-آنتروپی عملکرد بهتری از خود نشان داده است زیرا مقدار معیارهای حساسیت، شفافیت، دقت، صحت، شاخص F۱ و ارزش اخباری منفی برای این روش بیشترین مقدار خود و معیارهای نرخ کشف غلط، نرخ مثبت غلط، نرخ منفی غلط و میانگین توان‌های دوم خطا در این روش کمترین مقدار خود را نسبت به سایر روش‌ها دارند (بجز حالت برابری با آنتروپی رنی برای معیار حساسیت و نرخ منفی غلط). شکل ۲(ب) عملکرد معیار صحت که مهم ترین شاخص ارزیابی محسوب می‌شود را برای روش‌های مورد مقایسه نشان می‌دهد. همان‌طور که در شکل مشخص است، صحت تی-آنتروپی برای $t = ۳۱$ به مقدار بیشینه خود رسیده و با اختلاف زیاد بالاتر از سایر روش‌ها قرار گرفته است. همچنین برای این مجموعه داده معیارهای ارزیابی دو روش شاخص جینی و آنتروپی شانون برابر است یعنی درخت حاصل از این روش‌ها عملکرد مشابه دارند (شکل ۲(ب)).

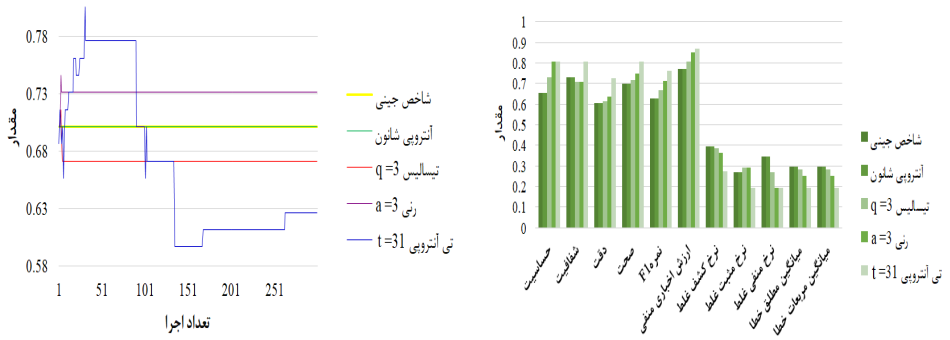
۴.۲ مجموعه داده دوم

این مجموعه داده (گریسونی و همکاران، ۲۰۱۹) مربوط به ۱۶۸۷ ماده شیمیایی شامل ۱۰۲۴ متغیر است که در شرایط آزمایشگاهی با توان عملیاتی بالا بدست آمده‌اند و در سال ۲۰۱۹ در سایت UCI قرار گرفته است.[2] نتایج ارزیابی در جدول ۵ و شکل ۳(الف) نمایش داده شده‌اند. این نتایج بیانگر آن است که تی-آنتروپی در معیارهای حساسیت، دقت، صحت، شاخص F۱ و ارزش اخباری منفی بیشینه مقدار و معیارهای نرخ کشف غلط، نرخ مثبت غلط، نرخ منفی غلط و میانگین توان‌های دوم خطا در این روش کمینه مقدار را نسبت به سایر روش‌ها دارند. لازم به ذکر است که تنها معیار شفافیت برای آنتروپی رنی بیشترین مقدار را دارد. در حالیکه تی - آنتروپی در ۹ معیار دیگر بویژه معیار

[1] https://archive.ics.uci.edu/ml/datasets/spect%2Bheart
[2] https://archive.ics.uci.edu/ml/datasets/QSAR+androgen+receptor

جدول ۴. مقادیر معیارهای ارزیابی برای مجموعه داده اول

ردیف	معیار ارزیابی	شاخص جینی	آنتروپی شانون	آنتروپی تی-سالیس	آنتروپی رنی	تی-آنتروپی
	پارامترهای مدل			q= ۳	$a = ۳$	t=۳۱
۱	حساسیت/نرخ مثبت درست ↑	۰/۶۵۳۸	۰/۶۵۳۸	۰/۷۳۰۴	۰/۸۰۷۷	۰/۸۰۷۷
۲	شفافیت/نرخ منفی درست ↑	۰/۷۳۱۷	۰/۷۳۱۷	۰/۷۰۷۳	۰/۷۰۷۳	۰/۸۰۴۹
۳	دقت/ارزش اخباری مثبت ↑	۰/۶۰۷۱	۰/۶۰۷۱	۰/۶۱۲۹	۰/۶۳۶۴	۰/۷۲۴۱
۴	صحت ↑	۰/۷۰۱۵	۰/۷۰۱۵	۰/۷۱۶۴	۰/۷۴۶۳	۰/۸۰۶۰
۵	شاخص F۱ ↑	۰/۶۲۹۶	۰/۶۲۹۶	۰/۶۶۶۷	۰/۷۱۱۹	۰/۷۶۳۶
۶	ارزش اخباری منفی↑	۰/۷۶۹۲	۰/۷۶۹۲	۰/۸۰۵۶	۰/۷۵۲۹	۰/۸۶۸۴
۷	نرخ کشف غلط ↓	۰/۳۹۲۹	۰/۳۹۲۹	۰/۳۸۷۱	۰/۳۶۳۶	۰/۲۷۵۹
۸	نرخ مثبت غلط ↓	۰/۲۶۸۳	۰/۲۶۸۳	۰/۲۹۲۷	۰/۲۹۲۷	۰/۱۹۵۱
۹	نرخ منفی غلط ↓	۰/۳۴۶۲	۰/۳۴۶۲	۰/۲۶۹۲	۰/۱۹۲۳	۰/۱۹۲۳
۱۰	میانگین توان‌های دوم خطا ↓	۰/۲۹۸۵	۰/۲۹۸۵	۰/۲۸۳۶	۰/۲۵۳۷	۰/۱۹۴۰



(ب) (الف)

شکل ۲. الف- نمودار معیارهای ارزیابی و ب- نمودار صحت برای مجموعه داده اول

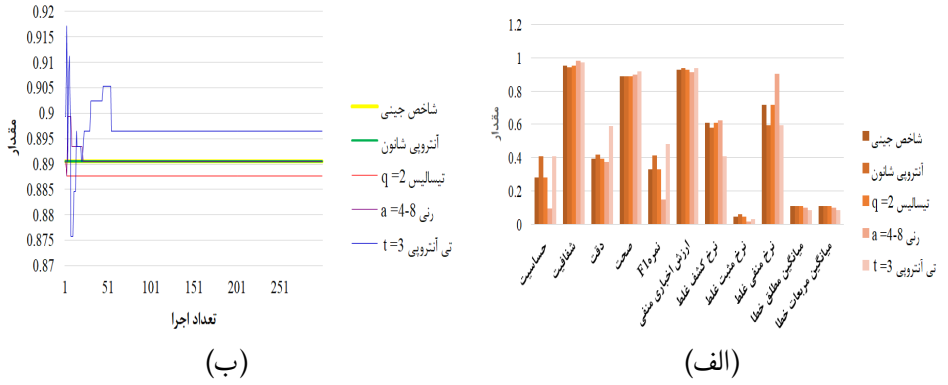
صحت، نسبت به سایر روش‌ها پیشرو بوده است. شکل ۳(ب) نمودار معیار صحت را برای روش‌های تحت بررسی نشان می‌دهد. همان‌طور که در شکل مشخص است علی رغم حجیم بودن این مجموعه داده و تعداد زیاد متغیرها، صحت تی-آنتروپی در $t = ۳$ به بالاترین مقدار خود رسیده است که نشان‌دهنده کارایی روش پیشنهادی برای داده‌های با حجم بالاست. لازم به ذکر است روش‌های شاخص جینی حالت خاصی از آنتروپی تیسالیس برای $q = ۲$ است لذا نتایج جدول ۴ برای این دو روش یکسان شده‌اند.

۴.۳ مجموعه داده سوم

این مجموعه شامل ۹۰ داده و ۷ متغیر است که به روش سرشماری (شامل همه موارد) در فاصله زمانی زمستان ۹۳ تا تابستان ۹۴ بدست آمده است. متغیرهای این مجموعه داده شامل: جنس، فشار خون، قند خون، کلسترول، مصرف سیگار، وزن و شغل بیمار می‌باشند که با هدف پیش‌بینی تعیین روش درمان (درمان طبی یا آنژیوگرافی) جمع آوری

جدول ۵. مقادیر معیارهای ارزیابی برای مجموعه داده دوم

ردیف	معیار ارزیابی	شاخص جینی	آنتروپی شانون	آنتروپی تی-سالیس	آنتروپی رنی	تی-آنتروپی
	پارامترهای مدل			q= ۲	$a = ۴-۸$	t=۳
۱	حساسیت/نرخ مثبت درست ↑	۰/۲۸۱۳	۰/۴۰۶۳	۰/۲۸۱۳	۰/۰۹۳۸	۰/۴۰۶۳
۲	شفافیت/نرخ منفی درست ↑	۰/۹۵۴۲	۰/۹۴۱۲	۰/۹۵۴۲	۰/۹۸۳۷	۰/۹۷۰۶
۳	دقت/ارزش اخباری مثبت ↑	۰/۳۹۱۳	۰/۴۱۹۴	۰/۳۹۱۳	۰/۳۷۵	۰/۵۹۰۹
۴	صحت ↑	۰/۸۹۰۵	۰/۸۹۰۵	۰/۸۹۰۵	۰/۸۹۹۴	۰/۹۱۷۲
۵	شاخص F۱ ↑	۰/۳۲۷۳	۰/۴۱۲۷	۰/۳۲۷۳	۰/۱۵	۰/۴۸۱۵
۶	ارزش اخباری منفی↑	۰/۹۲۷۰	۰/۹۳۸۱	۰/۹۲۷۰	۰/۹۱۲۱	۰/۹۳۹۹
۷	نرخ کشف غلط ↓	۰/۶۰۸۷	۰/۵۸۰۶	۰/۶۰۸۷	۰/۶۲۵	۰/۴۰۹۱
۸	نرخ مثبت غلط ↓	۰/۰۴۵۸	۰/۰۵۸۸	۰/۰۴۵۸	۰/۰۱۶۳	۰/۰۲۹۴
۹	نرخ منفی غلط ↓	۰/۷۱۸۸	۰/۵۹۳۸	۰/۷۱۸۸	۰/۹۰۶۳	۰/۵۹۳۸
۱۰	میانگین توان‌های دوم خطا ↓	۰/۱۰۹۵	۰/۱۰۹۵	۰/۱۰۹۵	۰/۱۰۰۶	۰/۰۸۲۸



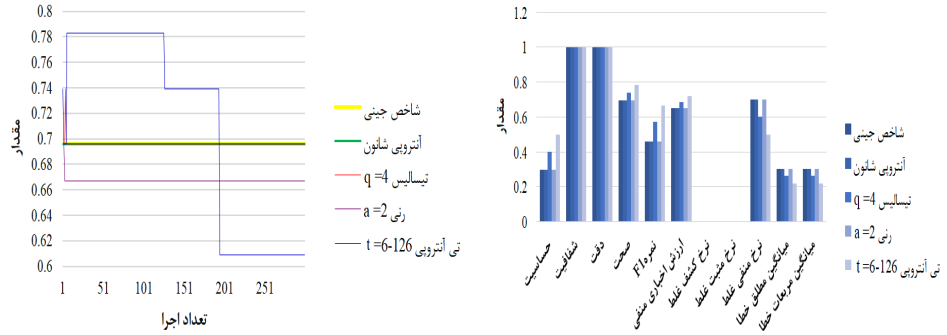
(ب) (الف)

شکل ۳. الف- نمودار معیارهای ارزیابی و ب- نمودار صحت برای مجموعه داده دوم

شده‌اند. نتایج گزارش شده در جدول ۶ و شکل ۴(الف) گویای آن است که عملکرد روش تی - آنتروپی در معیارهای حساسیت، صحت، شاخص $F۱$ ، ارزش اخباری منفی ، نرخ منفی غلط و میانگین توان‌های دوم خطا نسبت به سایر روش‌ها بهتر بوده است همچنین برای معیارهای شفافیت، دقت، نرخ کشف غلط و نرخ مثبت غلط، مقادیر بدست آمده برای تمام روش‌ها برابر هستند. از آنجایی‌که معیار صحت دارای اهمیت بیشتری است مطابق شکل ۴(ب) مقدار این معیار برای روش تی-آنتروپی با اختلاف زیادی بیشتر از سایر روش‌ها است. روی این مجموعه داده روش‌های شاخص جینی، آنتروپی‌های شانون و رنی رفتار یکسانی دارند. شکل ۴(ب) نشان می‌دهد که تی - آنتروپی برای $t = ۶$ تا $t = ۱۲۶$ بیشینه مقدار خود را دارد و پس از آن مقدار صحت به صورت پلکانی کاهش می‌یابد این در حالی است که روش تیسالیس به بیشینه مقدار خود در $q = ۴$ می‌رسد و پس از آن رو به کاهش است. همچنین سایر روش‌ها در تمامی ۳۰۰ اجرا رفتار یکسان دارند.

جدول ۶. مقادیر معیارهای ارزیابی برای مجموعه داده سوم

ردیف	معیار ارزیابی	شاخص جینی	آنتروپی شانون	آنتروپی تی-سالیس	آنتروپی رنی	تی-آنتروپی
	پارامترهای مدل			q= ۴	$a = ۲$	t=۶-۱۲۶
۱	حساسیت/نرخ مثبت درست ↑	۰/۳۰۰۰	۰/۳۰۰۰	۰/۴۰۰۰	۰/۳۰۰۰	۰/۵۰۰۰
۲	شفافیت/نرخ منفی درست ↑	۱	۱	۱	۱	۱
۳	دقت/ارزش اخباری مثبت ↑	۱	۱	۱	۱	۱
۴	صحت ↑	۰/۶۹۵۷	۰/۶۹۵۷	۰/۷۳۹۱	۰/۶۹۵۷	۰/۷۸۲۶
۵	شاخص F۱ ↑	۰/۴۶۱۵	۰/۴۶۱۵	۰/۵۷۱۴	۰/۴۶۱۵	۰/۶۶۶۷
۶	ارزش اخباری منفی↑	۰/۶۵۰۰	۰/۶۵۰۰	۰/۶۸۴۲	۰/۶۵۰۰	۰/۷۲۲۲
۷	نرخ کشف غلط ↓	۰	۰	۰	۰	۰
۸	نرخ مثبت غلط ↓	۰	۰	۰	۰	۰
۹	نرخ منفی غلط ↓	۰/۷۰۰۰	۰/۷۰۰۰	۰/۶۰۰۰	۰/۷۰۰۰	۰/۵۰۰۰
۱۰	میانگین توان‌های دوم خطا ↓	۰/۳۰۴۳	۰/۳۰۴۳	۰/۲۶۰۹	۰/۳۰۴۳	۰/۲۱۷۴



(ب) (الف)

شکل ۴. الف- نمودار معیارهای ارزیابی و ب- نمودار صحت برای مجموعه داده سوم

بحث و نتیجه‌گیری

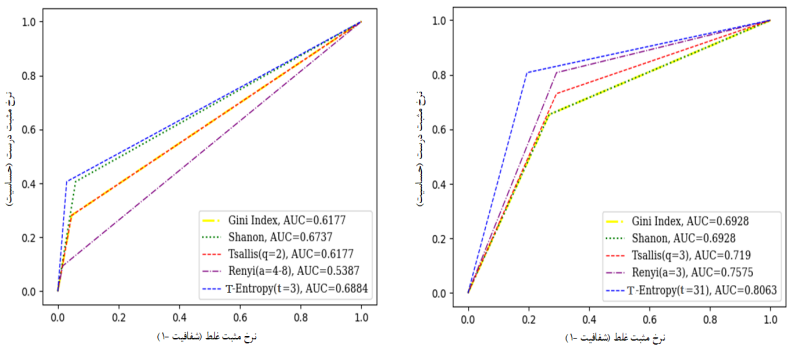
بررسی نتایج حاصل شده نشان‌دهنده عملکرد بهتر روش معرفی شده تی-آنتروپی در مقایسه با سایر روش‌های مورد بحث، بویژه برای داده‌های با حجم بالا و تعداد متغیر زیاد است. نتایج تحلیل سه مجموعه داده بر اساس ۱۰ معیار ارزیابی نشان می‌دهد که روش تی-آنتروپی رفتاری بهتر نسبت به سایر روش‌ها دارد. یکی از راه‌های ارزیابی کلی عملکرد یک آزمون استفاده از نمودار [۱] (ROC) است. این نمودار نمایش گرافیکی نرخ مثبت درست (یا حساسیت) در مقابل نرخ مثبت غلط (یا همان شفافیت منهای یک) و نشان‌دهنده توانایی ارزیابی سیستم دسته‌بندی باینری با آستانه تشخیص متغیر است. نقاطی که در این نمودار بالای خط نیمساز قرار دارند، مقدار حساسیت یا نرخ مثبت درست آن‌ها بیشتر از نرخ مثبت غلط بوده که جزو نقاط مطلوب به حساب می‌آیند. نقاط روی خط نیمساز و یا

[1] Receiver Operating Characteristic

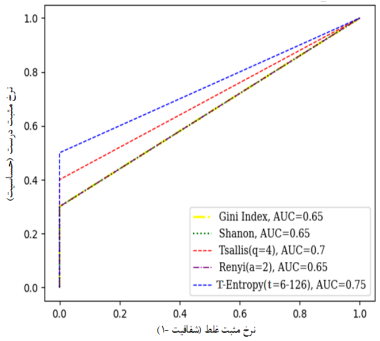
پایین آن دارای مقدار عددی نرخ مثبت درست، برابر و یا کمتر از نرخ مثبت غلط می‌باشند که بعنوان نقاط نامطلوب محسوب می‌شوند. نمودارهای ROC برای سه مجموعه داده در شکل ۵ نشان داده شده است. سطح زیر نمودار ROC (یا همان AUC) عددی بین صفر و یک است که قدرت تشخیص آزمون را نشان می‌دهد. مقادیر بزرگتر از $\frac{1}{2}$ برای این معیار، بیانگر بیشتر بودن نرخ مثبت درست نسبت به نرخ مثبت غلط، مقادیر AUC نزدیک به $\frac{1}{2}$ یا برابر با آن برابری نرخ مثبت درست و نرخ مثبت غلط و مقادیر کمتر از $\frac{1}{2}$ ، نرخ مثبت غلط بیشتر در مقایسه با نرخ مثبت درست است. در شکل ۵(الف) مقدار AUC برای روش‌های شاخص جینی و آنتروپی شانون برابر است. شکل ۵(ب) برابری مقدار AUC برای دو روش جینی و تیسالیس $(q = ۲)$ را مشابه جدول ۴ نشان می‌دهد. شکل ۵(ج) و جدول ۵ عملکرد یکسان روش‌های شاخص جینی و آنتروپی‌های شانون و رنی $(\alpha = ۲)$ را از طریق برابری مقدار AUC این سه روش تایید می‌نماید. مطابق شکل ۵، مقدار AUC در روش تی-آنتروپی ارایه شده برای هر سه مجموعه داده بیشتر از سایر روش‌ها است که این نشان‌دهنده بیشینه بودن مقدار نرخ مثبت درست نسبت به نرخ مثبت غلط در روش تی-آنتروپی نسبت به سایر روش‌ها است. در شکل ۵(الف) و (ج) درخت تصمیم تولید شده با روش انشعاب تی-آنتروپی با اختلاف زیادی بهتر از سایر روش‌ها است در حالی‌که در شکل ۵(ب)، درخت تصمیم تولید شده با انشعاب تی-آنتروپی، اختلاف کمتری نسبت به سایر روش‌ها دارد. همان‌طور که در شکل ۵ مشخص است، مقادیر بهینه برای پارامترهای q، α و t برای مجموعه داده‌های مختلف، مقادیری متفاوت هستند، لذا نمی‌توان رابطه‌ای جهت برآورد این پارامترها بدست آورد. معیارهای ارزیابی جهت برتری یک روش انشعاب مورد استفاده شده تاکنون صحت، دقت، حساسیت، میانگین توان‌های دوم خطا و سطح زیر نمودار ROC (شارما و همکاران، ۲۰۱۳؛ ونگ و همکاران، ۲۰۱۷؛ ردی و چیتنینی، ۲۰۲۱) هستند در حالی‌که مقاله حاضر علاوه بر پنج معیار مذکور، شش معیار دیگر را نیز مد نظر قرار داده است. در این مقاله، عملکرد روش‌های انشعاب شاخص جینی، آنتروپی‌های شانون، تیسالیس، رنی و تی-آنتروپی با بکارگیری معیارهای ارزیابی صحت، دقت، حساسیت، شفافیت، شاخص F۱ و ارزش اخباری منفی، نرخ کشف غلط، نرخ مثبت غلط، نرخ منفی غلط و میانگین توان‌های دوم خطا بررسی شد. مقادیر بیشینه شش معیار اول و مقادیر کمینه چهار معیار دوم بیانگر عملکرد بهتر درخت تصمیم معرفی شده است. روش انشعاب تی-آنتروپی با هر ترکیب متفاوتی از حجم داده و تعداد متغیر، برای مجموعه داده‌های منتخب رفتار بهتری نسبت به سایر روش‌ها از خود نشان می‌دهد. لذا پیشنهاد می‌گردد روش معرفی شده برای انشعاب در تولید درختان تصمیم بهینه در جهت پیش‌بینی/طبقه‌بندی دقیق تر مورد توجه قرار گیرد. همچنین با توجه به حجم داده‌ها و رشد روزافزون مساله کلان داده‌ها در سال‌های اخیر و از طرف دیگر برتری روش انشعاب تی-آنتروپی نسبت به سایر روش‌های مورد بررسی روی این گونه از داده‌ها، اهمیت این روش دو چندان است.

تقدیر و تشکر

نویسنده مقاله مراتب قدردانی و سپاس خود را از پیشنهادات سازنده داوران، سردبیر و ویراستار محترم مجله در جهت ارائه بهتر مطالب و ارتقای سطح کیفی مقاله، اعلام می‌دارد.



(ب) (الف)



(ج)

شکل ۵. نمودار ROC الف- مجموعه داده اول، ب- مجموعه داده دوم، ج- مجموعه داده سوم

مراجع

آقامحمدی، ژ، رضایی قهرودی، ز. (۱۴۰۱)، اتصال رکوردی با روش های یادگیری ماشین، مجله علوم آماری ، **(۱)۱۶**، ۱-۲۴.

عاشوری، م. (۱۳۹۶)، ارایه مدلی برای پیش‌بینی نوع صافی همودیالیز با تکنیک‌های داده‌کاوی، مدیریت سلامت، **۲۰(۶۷)**، ۱۱۰-۹۹.

Barros, R., Basgalupp, M., Carvalho, A. and Freitas, A. (2012), A Survey of Evolutionary Algorithms for Decision-tree Induction, *IEEE Transactions on Systems, Man, and Cybernetics*, **42**, 291-312.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification And Regression Trees*, Wadsworth International Group.

Chakraborty, S., Paul, D. and Das, S. (2021), T-Entropy: a New Measure of Uncertainty with Some Applications, *IEEE Symposium on Information Theory (ISIT)*.

Dua, D. and Graff, C. (2019), UCI Machine Learning Repository http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Grabczewski, K. (2014), *Meta-learning in Decision Tree Induction*, Springer, New York.

Grisoni, F., Consonni, V. and Ballabio, D. (2019), Machine Learning Consensus to Predict the Binding to the Androgen Receptor within the CoMPARA Project, *Journal of Chemical Information and Modeling*,**59**, 1839-1848

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. and Bing, G. (2017), Learning from Class-imbalanced Data: Review of Methods and Applications, *Expert Systems with Applications*. **73**, 220-239.

Hehn, T., Kooij, J. and Hamprecht, F. (2020), End-to-End Learning of Decision Trees and Forests, *International Journal of Computer Vision*, **128**, 997-1011.

Hothorn, T., Hornik , K. and Zeileis, A. (2006), Unbiased Recursive Partitioning: A Conditional Inference Framework, *Journal of Computational and Graphical Statistics*, **15**(3), 651-674.

Hyafil, L. and Rivest, R. (1976), Constructing Optimal Binary Decision Trees is NP-complete, *Information Processing Letters*, **5**(1), 15-17.

Kass, G. (1980), An Exploratory Technique for Investigating Large Quantities of Categorical Data, *Journal of the Royal Statistical Society*, **29**(2), 119-127.

Kolose, S., Stewart, T., Hume, P. and Tomkinson, G.R. (2021), Prediction of Military Combat Clothing Size Using Decision Trees and 3D Body Scan Data, *Applied Ergonomics*, **95**, 1-9.

Kotsiantis, S. (2013), Decision Trees: A Recent Overview, *Artificial Intelligence Review*, **39**, 261-283.

Kurgan, L.A., Cios, K.J., Tadeusiewicz, R., Ogiela, M. and Goodenday, L.S. (2001), Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis, *Artificial Intelligence in Medicine*, **23**(2), 149-169.

Liu, S., McGree, J., Ge, Z. and Xie, Y. (2016), *Computational and Statistical Methods for Analysing Big Data with Applications*, Academic Press.

Lohse, J., Doulgeris, A.P. and Dierking, W. (2019), An Optimal Decision-tree Design Strategy and Its Application to Sea Ice Classification from SAR Imagery, *Remote Sens*, **11**, 1574 – 1589.

McClean, S. (2003), *Data Mining and Knowledge Discovery, in Encyclopedia of Physical Science and Technology(Third Edition)*, New York: Academic Press.

Mehta, M., Agrawal , R. and Rissanen, J. (1996), SLIQ: A Fast Scalable Classifier for Data Mining, *International Conference on Extending Database Technology (Advances in Database Technology).*

Nisbet, R., Miner, G. and Yale, K. (2018), *Handbook of Statistical Analysis and Data Mining Applications*, Academic Press.

Quinlan, J. (1979), Discovering Rules by Induction from Large Collections of Examples, *Expert Systems in the Micro Electronic Age*, Edinburgh University Press, Edinburgh, UK, 168-201.

Quinlan, J. (1994), *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers.

Reddy, G. S. and Chittineni, S. (2021), Entropy Based C4.5-SHO Algorithm with Information Gain Optimization in Data Mining, *PeerJ Computer Science*, **17**, 1-22.

Safavian, S. and Landgrebe, D. (1991), A Survey of Decision Tree Classifier Methodology, *IEEE Transactions on Systems, Man, and Cybernetics* **21(3)**, 660-674.

Shafer, J., Agrawal , R. and Mehta, M. (1996), SPRINT: A Scalable Parallel Classifier for Data Mining, *Proceedings of the 22th International Conference on Very Large Data Bases*.

Shanbehzadeh,M., Nopour, R. and Kazemi-Arpanahi, H. (2022), Using Decision Tree Algorithms for Estimating ICU Admission of COVID-19 Patients, *Informatics in Medicine Unlocked*, **30**, 1-9.

Sharma, S., Agrawal, J. and Sharma, S. (2013), Classification Through Machine Learning Technique: C4. 5 Algorithm Based on Various Entropies, *International Journal of Computer Applications*, **82**(16), 28-32.

Shobha, G. and Rangaswamy, S. (2018), *Machine Learning*, in: Handbook of Statistics, Elsevier, **38**, 197–228.

Shotton, J., Fitzgibbon, A., Cook, K., Sharp, T., Finocchio,M. and Moore, M. (2011), Real-time Human Pose Recognition in Parts from Single Depth Images CVPR 2011, 1297-1304.

Wang, Y. and Xia, S. (2017), Unifying Attribute Splitting Criteria of Decision Trees by Tsallis Entropy, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2507-2511.

Wang, Y., Xia, S-T. and Wu, J. (2017), A Less-greedy Two-term Tsallis Entropy Information Metric Approach for Decision Tree Classification, *Knowledge-Based Systems*, **120**(c), 34–42.

Zhang, L., Varadarajan, J., Suganthan, P., Ahuja, N. and Moulin, P. (2017), Robust Visual Tracking Using Oblique Random Forests. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.