فصلنامه (علمی – پژوهشی)

زیست شناسی تکوینی

سال دهم، شماره ۳، تابستان ۱۳۹۷

تعیین مهم‌ترین ویژگی‌های کمی و کیفی متمایزکننده سرده *Rubus* L. در ایران با استفاده از الگوریتم‌های دسته‌بندی و انتخاب ویژگی

محمدجواد شیخ‌زاده *

دانشکده علوم پایه و فنی مهندسی، دانشگاه گنبد کاووس، گنبد کاووس، ایران

* Email: sheikhzadeh@gonbad.ac.ir

تاریخ دریافت: ۹۷/۰۱/۰۷ تاریخ پذیرش: ۹۷/۰۴/۲۰

چکیده

جنس سرده *Rubus* L. متعلق به تیره Rosaceae و زیرتیره Rosoideae شامل حدود ۷۵۰ گونه در دنیا است. این جنس در اکثر نواحی جهان پراکنش دارد. در فلور ایران تعداد هشت گونه و پنج هیبرید (دو رگه) گزارش شده است. تمشک یکی از پر چالش‌ترین جنس‌ها در بین گیاهان گل‌دار می‌باشد. وجود عواملی از جمله پلی‌پلوئیدی، آپومیکسی و دورگه‌گیری زیاد در این جنس باعث مشکلاتی در تشخیص از نظر ریخت شناسی شده است. جمع آوری داده‌های کمی و کیفی جهت تشخیص گونه‌ها و روش‌های جمع‌آوری مولفه‌های ریخت شناسی بسیار زمانبر و پر هزینه است. بنابراین، بکارگیری روش‌های متفاوت در جهت کاهش زمان و هزینه‌ها همواره مطرح می‌باشد. در بسیاری از موارد، جهت آنالیز داده‌های زیستی روش‌های داده کاوی بکار گرفته می‌شود. در این مقاله، از ترکیب الگوریتم‌های مختلف انتخاب ویژگی و دسته‌بندی برای تشخیص ویژگی‌های متمایزکننده بین گونه‌های سرده *Rubus* L. استفاده شد. با بکارگیری روش دسته‌بندی Random Forest و مدل انتخاب ویژگی InfoGainAttributeEval با تعداد ۲۸ ویژگی به دقت ۹۴/۰۵ درصد در دسته‌بندی رسیدیم که بهترین الگوریتم از نظر دقت می‌باشد و با استفاده از روش MLP و مدل انتخاب ویژگی SymetricalAttributeEval، با تعداد ۴ ویژگی دقت دسته‌بندی ۸۴/۳۲ درصد حاصل شد که بهترین الگوریتم از نظر تعداد اندک ویژگی‌های انتخاب شده است. چهار ویژگی فوق توسط اکثر الگوریتم‌های استفاده شده در این مقاله انتخاب شدند. تمامی این ویژگی‌ها کیفی هستند و جهت بدست آوردن آنها نیازی به هزینه اندازه‌گیری آزمایشگاهی نمی‌باشد. بنابراین می‌توانند معیار مناسبی جهت کلید شناسایی باشند.

کلیدواژه‌ها: الگوریتم، داده کاوی، ریخت شناسی، کلید شناسایی.

مقدمه

جنس تمشک متعلق به تیره Rosaceae و زیرتیره Rosoideae می‌باشد و شامل حدود ۷۵۰ گونه در دنیا است [۱۲،۲۲،۲۹،۳۱] که در اکثر نواحی جهان به جز ناحیه قطب جنوب پراکنش دارد [۹،۱۰،۱۱،۱۶]. این تاکسون از جنبه اقتصادی و بوم‌شناسی حائز اهمیت

است، چون دارای میوه‌های خوراکی وگیاهانی با ارزش تزئینی هستند. این جنس به واسطه داشتن فرم رویشی درختچه‌ای و رونده، ساقه خاردار و میوه خوراکی از نوع شفت مجتمع متشکل از شفتچه‌هایی کوچک به آسانی قابل تشخیص است. اگرچه تشخیص گونه‌های آن به واسطه تنوع ریخت‌شناسی در جمعیت‌های گونه بسیار مشکل است. این جنس در ۱۲ زیرجنس طبقه‌بندی می‌شود [۱۱-۹]. بزرگترین زیرجنس آن *Rubus* با ٦ بخش و ۱۳۲ گونه می‌باشد و کوچکترین زیر جنس آن *Chamaemorus* با یک گونه می‌باشد. زیر جنس *Rubus* شامل گونه‌هایی هستند که در اروپا، آسیا و آمریکای شمالی یافت می‌شوند [۳]. خاتم‌ساز هشت گونه و پنج هیبرید (دورگه) را از ایران نام برده است [۲۰]، که به دو زیر جنس *Cylactis* و *Rubus* تعلق دارد. *R. saxatilis* به‌عنوان تنها گونه‌ی علفی جنس در ایران متعلق به زیرجنس *Cylactis* و ۷ گونه *R. caesius*، *R. discolor*، *R .dolichocarpus*، *R. hirtus*، *R .hyrcanus*، *R. persicus* و *R. sanctus* متعلق به زیرجنس *Rubus* می‌باشند. تمشک یکی از پرچالش‌ترین جنس‌ها در بین گیاهان گل‌دار است [۱،۲۲،۲۷،۲۹] که این چالش‌ها به خاطر وجود عوامل بیولوژیکی از جمله پلی‌پلوئیدی، آپومیکسی و دورگه‌گیری زیاد بین گونه‌های یک زیر-جنس یا گونه‌های زیرجنس‌های مختلف بوده و به نظر می‌رسد استفاده از ریخت‌شناسی ظاهری به‌تنهایی در روشن شدن وضعیت رده‌بندی این جنس کارایی نداشته باشد [۱۳،۳۲]. از سوی دیگر، استخراج ویژگی‌های کمی و کیفی یک گیاه کاری بسیار زمانبر و پرهزینه است. از طرفی، بسیاری از آنها یا بلااستفاده هستند و یا اینکه بار اطلاعاتی چندانی ندارند. از این رو هرچه تعداد این ویژگی‌ها کمتر باشد، هزینه و زمان صرف شده برای استخراج و همچنین ارزیابی آنها کاهش می‌یابد. وجود داده‌های بزرگ در برخی از مسائل که ارزیابی و تحلیل آنها سخت، پرهزینه و زمان‌بر است و دقت پایین الگوریتم‌ها باعث می‌شود تا همچنان پیدا کردن الگوریتمی که بتواند به طوری کارآمد ویژگی‌های تفکیک کننده را مشخص کرده و با استفاده از آنها نمونه‌ها را دسته‌بندی کند حائز اهمیت باشد [۳۳،۳٤].

هدف از این پژوهش، معرفی روشی بهینه برای تشخیص مهم‌ترین ویژگی‌های متمایزکننده و همچنین دسته‌بندی گونه‌های مختلف سرده تمشک در ایران است. طوری که با استفاده از تعدادی کمتر از ویژگی‌ها با دقت بیشتری نمونه‌ها را دسته‌بندی کرد. در بسیاری از موارد، جهت آنالیز داده‌های زیستی (از جمله مطالعات آناتومی و ریخت شناسی) روش‌های داده کاوی بکار گرفته می‌شود. بنابراین در این مقاله، از ترکیب الگوریتم‌های مختلف انتخاب ویژگی و دسته‌بندی برای تشخیص ویژگی‌های متمایزکننده بین گونه‌های سرده *Rubus* L. استفاده می‌شود.

مواد و روش‌ها

در این مقاله، به منظور انجام مطالعات ریخت‌شناسی روی جمعیت‌های مختلف گونه‌های معرفی شده سرده *Rubus* در ایران (شامل *R. caesius*، *R. discolor*، *R. dolichocarpus*، *R. hirtus*، *R. hyrcanus*، *R. persicus* و *R. Sanctus*) از مناطق مختلف شمال ایران از سه استان گیلان، مازندران و گلستان نمونه‌هایی مطابق (جدول ۱) جمع‌آوری گردید. به این ترتیب در این پژوهش از ۵۷۸۱ داده خام حاصل از تعداد ٤٦ ویژگی کیفی و کمی استفاده شده است، که این ویژگی‌ها در (جدول ۲ و ۳) ارائه

گردیده است. ویژگی‌ها کدگذاری شده و بر اساس است [۱۸].
تعداد نمونه‌ها، میانگین و واریانس محاسبه گردیده

جدول ۱- نمونه‌های جمع‌آوری شده در سه استان مورد مطالعه [۱۸]

ردیف	گونه	محل جمع‌آوری	ارتفاع (متر)	جمع‌آوری کننده	مشخصات هرباریومی
۱	*R. caesius* L	مازندران، رامسر، قاسم آباد سفلی	۵۰	کسلخه، حبیبی	-803012 GKUH
۲	*R. caesius* L	گلستان، بندر گز، گز غربی	۹۵	کسلخه، حبیبی	-803025 GKUH
۳	*R. discolor* Weihe and Nees.	گیلان، جاده آستارا به اردبیل، گردنه حیران	۵۰۰	کسلخه، مهدی‌یانی	-803055 GKUH
٤	*R. dolichocarpus* Jaz.	گلستان، پارک ملی گلستان	۵۰۰	کسلخه، مهدی‌یانی	803120-GKUH
۵	*R. hirtus* Waldst and kit	مازندران، جنگل سنگده	۱۳٦۰	کسلخه، مهدی‌یانی	-803135 GKUH
٦	*R. hyrcanus* Juz.	گیلان، جاده اسالم به خلخال	۱۰۳۰	کسلخه، حبیبی	-803146 GKUH
۷	*R. persicus* Boiss.	گلستان، جنگل امام رضا (ع) کردکوی	۲۵۰	مهدی‌یانی	-803150 GKUH
۸	*R. sanctus* Schreber	گیلان، جیرنده، روستای بی‌ورزین، نزدیک زیارتگاه	۱۰٤۰	کسلخه، حبیبی	-803235 GKUH

جدول ۲- لیست ویژگی‌های کیفی سرده *Rubus* و کد گذاری آنها در مطالعات ریخت‌شناسی [۱۸]

ردیف	ویژگی‌ها	شرح	انحراف معیار	واریانس
۱	فرم رویشی	بوته‌ای (۰)، درختچه‌ای (۱)	۰/۳۷۵۰	۰/۸۳۰٦
۲	فرم ساقه	خوابیده (۰)، خمیده (۱)، افراشته (۲)، بالارونده (۳)	۰/۷٦٤۵	۱/٤۳۵٤
۳	شکل شاخه گل‌دهنده	گرد (۰)، تا حدودی زاویه‌دار (۱)، زاویه‌دار (۲)	۰/۷٦۳۸	۱/٤۲۷٤
٤	شکل خار شاخه گل‌دهنده	مستقیم کوتاه (۰)، خمیده کوچک (۱)، خمیده درشت (۲)، سوزنی (۳)، خمیده سوزنی (٤)، خمیده و مستقیم (۵)	۱/۳۰۲۰	۲/۲٦٦۱
۵	حضور کرک در شاخه گل‌دهنده	بدون کرک (۰)، تا حدودی کرک‌دار (۱)، کرک‌دار (۲)	۰/۳۵۲۲	۱/۸۵٤۸
٦	نوع کرک در شاخه گل‌دهنده	تار (۰)، ستاره (۱)، تار و ستاره (۲)، نمدی (۳)	۰/۵٤٦۲	۲/۵
۷	حضور غده پایکدار در شاخه	حضور غده پایکدار(۰)، عدم حضور غده (۱)، حضور غده بدون پایک (۲)	۰/۳۷٤٤	۰/۸۵٤۸
۸	شکل برگچه جانبی	تخم‌مرغی (۰)، واژتخم‌مرغی (۱)، بیضوی (۲)، بیضوی پهن (۳)، لوزی (٤)، دایره‌ای (۵)، تخم‌مرغی- بیضوی (٦)، واژتخم‌مرغی- بیضوی (۷)	۲/۷۹۳۹	۳/۹۸۳۸
۹	شکل برگچه انتهایی	تخم‌مرغی (۰)، واژتخم‌مرغی (۱)، بیضوی (۲)، لوزی (۳)، بیضوی پهن با عرض بیش از ۲.۵ برابر طول(٤)، دایره‌ای(۵)	۱/۲٦۵٤	۱/۷۸٦۲
۱۰	نوک برگچه جانبی	بدون نوک (۰)، تیز یا کمی کشیده (۱)، دم‌دار (۲)، نوک کند (۳)	۰/۸۳۷۳	۱/۳۷۰۹
۱۱	نوک برگچه انتهایی	بدون نوک (۰)، تیز یا کمی کشیده (۱)، دم‌دار (۲)، نوک کند (۳)	۰/۸٤٤۰	۱/٤۲۷٤

ردیف	ویژگی‌ها	شرح	انحراف معیار	واریانس
۱۲	قاعده برگچه جانبی	قلبی (۰)، دایره‌ای (۱)، قلبی- دایره‌ای (۲)	۰/٦۱۲۲	۱/٥٦٤٥
۱۳	قاعده برگچه انتهایی	قلبی (۰)، دایره‌ای (۱)، قلبی- دایره‌ای (۲)	۰/۸۱٦۱	۱/٥٥٦٤
۱٤	رنگ سطح تحتانی برگچه	سبز (۰)، سفید (۱)، تا حدودی سبز (۲)، تا حدودی سفید (۳)	۰/٦۷٦۷	۰/۹٥۹٦
۱٥	پوشش کرک سطح فوقانی برگچه	بدون کرک (۰)، پراکنده (۱)، متوسط (۲)، فشرده (۳)	۰/۷٥۸٥	۱/۱۹۳٥
۱٦	پوشش کرک سطح تحتانی برگچه	بدون کرک (۰)، پراکنده (۱)، متوسط (۲)، فشرده (۳)	۰/٥۹۳۲	۲/٦۰٤۸
۱۷	نوع کرک سطح فوقانی برگچه	بدون کرک (۰)، اکثرا تار (۱)، اکثرا ستاره (۲)، تار و ستاره (۳)، نمدی (٤)	۱/۰۷۳۳	۱/٤٦۷۷
۱۸	نوع کرک سطح تحتانی برگچه	بدون کرک (۰)، اکثرا تار (۱)، اکثرا ستاره (۲)، نمدی (۳)	۰/۰۸۹٤	۲/۹۹۱۹
۱۹	شکل خار دمبرگ	مستقیم کوتاه (۰)، خمیده کوچک (۱)، خمیده درشت (۲)، سوزنی (۳)، خمیده سوزنی (٤)	۰/۹٥۳٤	۱/۱۸٥٤
۲۰	نوع کرک دمبرگ	اکثرا تار (۰)، اکثرا ستاره (۱)، نمدی (۳)، تار و ستاره (٤)	۰/٥٤۱٤	۲/٤۲۷٤
۲۱	وضعیت دمبرگچه جانبی	بدون دمبرگچه (۰)، دارای دمبرگچه با طول بیش از ۲mm (۱)، دمبرگچه خیلی کوتاه با طول کمتر از ۲mm(۲)	۰/۹٤٦۹	۱/۱۲۰۹
۲۲	نوع گل‌آذین	دیهیم (۰)، خوشه ساده (۱)، خوشه گرزن مرکب (۲)، خوشه مرکب (۳)	۰/۹۳۹۳	۱/٦۱۲۹
۲۳	حضور غده در گل‌آذین	غده‌دار (۰)، بدون غده (۱)	۰/٤۰۷۰	۰/۷۹۰۳
۲٤	شکل کاسبرگ	تخم‌مرغی (۰)، مستطیلی (۱)	۰/۹۸۱۲	۰/۸٥٤۸
۲٥	نوک کاسبرگ	بدون نوک تا کمی (۰)، تیز (۱)، سیخکی (۲)	۰/۹٥۰۷	۰/۸۲۲٥
۲٦	نوع کرک کاسبرگ	نمدی (۰)، تار و ستاره (۱)	۰/٤۲۳۲	۰/۲۳۳۸
۲۷	غده پایک‌دار در کاسبرگ	غده‌دار (۰)، بدون غده (۱)	۰/٤۲۳۲	۰/۷٦٦۱
۲۸	شکل گلبرگ	بیضوی کشیده (۰)، بیضوی واژگون (۱)	۱/۰۷۳۱	۰/٥٤۰۳
۲۹	رنگ گلبرگ	صورتی (۰)، سفید (۱)، سفید- صورتی (۲)	۰/٥۱۳٦	۰/٤٥۱٦
۳۰	رنگ میله	صورتی (۰)، سفید (۱)	۰/٤۷۳۲	۰/٦٦۱۲
۳۱	کرک بساک	بدون کرک (۰)، کرک‌دار (۱)	۰/٤۷٥۹	۰/۳٤٦۷
۳۲	شکل گوشوارک	خطی (۰)، خنجری- نیزه‌ای با عرض کمتر از ۲mm (۱)، نیزه‌ای پهن با عرض بیش از ۲mm (۲)	۰/٦۹۷۸	۰/٦٤٥۱
۳۳	تعداد برگچه	یک (۰)، دو الی بیست (۱)، بیست الی چهل(۲)، چهل الی شصت (۳)	۰/۷٤۹۹	۱/۷٤۱۹

جدول ۳- لیست ویژگی‌های کمی سرده *Rubus* و کد گذاری آنها در مطالعه ریخت‌شناسی [۱۸]

ردیف	ویژگی‌ها	انحراف معیار	واریانس
۱	طول دمبرگ	۱/۷۸٤٤	٤/۱۲٤٦
۲	طول برگچه انتهایی	۲/۱٥٤٤	٥/۹۸٥
۳	طول برگچه جانبی	۲/۰۰۹	٥/۱۲۹۳
٤	عرض برگچه انتهایی	۱/۳۷٤۱	٤/۰۹۲۱
٥	عرض برگچه جانبی	۱/۰٥٤٥	۳/۱٥۰٥
٦	طول دمبرگچه برگچه انتهایی	۹/۳٥٤۷	۱٤/٦۲٥۹
۷	طول دمبرگچه برگچه جانبی	۲/۳۸٦۱	۲/٥۱۹٥
۸	طول کاسبرگ	۱/۱٤۹٥	٦/۱۸۹۰

ردیف	ویژگی‌ها	انحراف معیار	واریانس
۹	نسبت طول به عرض کاسبرگ	۰/٤۲۲۳	۲/۱۲۲۰
۱۰	نسبت طول به عرض گلبرگ	۷/٦۳۲٥	۱۲/۳۲٤٥
۱۱	تعداد گل در گل آذین	۱۱/۱۸۳۹	۱۹/٥۰۳۰
۱۲	طول گل آذین	٥/۷۲۰٦	۱۱/٦٤۹۲
۱۳	طول دمگل	۱۰/٤٥۰۷	۱۸/۲٦٦۸

همانگونه که در مقدمه اشاره شد، مساله انتخاب ویژگی، در بسیاری از کاربردها مانند دسته‌بندی (Classification) اهمیت بسیار زیادی دارد، زیرا در این موارد تعداد زیادی ویژگی وجود دارد، که بسیاری از آنها بار اطلاعاتی چندانی ندارند و یا حتی دقت الگوریتم دسته‌بندی را کاهش می‌دهند. بنابراین حذف ویژگی‌های اضافی نه تنها هزینه جمع‌آوری داده‌های کیفی و کمی را کاهش می‌دهد و بار محاسباتی الگوریتم دسته‌بندی را پایین می‌آورد، بلکه ممکن است موجب افزایش دقت دسته‌بندی شود. جهت کاهش ویژگی‌های اضافی و دسته‌بندی آن‌ها، نرم‌افزاری به زبان جاوا پیاده‌سازی شد که از کتابخانه Weka (نسخه ۳/۸) بهره می‌برد که دارای امکانات بسیار گسترده جهت داده‌کاوی (همانند انواع الگوریتم‌های رگرسیون، کاوش قواعد انجمنی، انتخاب ویژگی، خوشه‌بندی و دسته‌بندی داده‌ها)، امکان مقایسه خروجی روشهای مختلف با هم، راهنمای خوب، واسط گرافیگی کارا و سازگاری با سایر برنامه‌ها است [۱۷]. این سیستم به زبان جاوا نوشته شده و بر اساس لیسانس عمومی و فراگیر در پایگاه [۲۳] انتشار یافته است. در نرم‌افزار پیاده‌سازی شده، داده‌ها با استفاده از تعدادی از الگوریتم‌های انتخاب ویژگی واکاوی شده و هر الگوریتم تعدادی ویژگی را از بین ٤٦ ویژگی اولیه انتخاب می‌کند. سپس برای ارزیابی کارایی ویژگی‌های انتخاب شده، از الگوریتم‌های دسته‌بندی استفاده می‌شود.

الگوریتم‌های انتخاب ویژگی[1]

در عصر ارتباطات و اطلاعات، جمع آوری و ذخیره داده‌ها بسیار آسان و ارزان می‌باشد. از سوی دیگر، با افزایش اطلاعات قابل خواندن توسط ماشین، امکان درک و استفاده از اطلاعات بطور همزمان بسیار مشکل بود. ماشین یادگیری ابزارهایی را فراهم می‌کند که می‌توان داده‌های بزرگ را بصورت خودکار تجزیه و تحلیل نمود. لذا اساس یادگیری ماشین، انتخاب ویژگی است. در انتخاب ویژگی باید از بین مجموعه ویژگی‌های مناسب با استفاده از یک روش تفکیک کننده، برجسته ترین و بهترین ویژگی‌ها برای یادگیری انتخاب می‌شود. در روش‌های مختلف انتخاب ویژگی، زیرمجموعه‌ای با کمترین تعداد ویژگی ممکن(ویژگی-های تاثیرگذار) هدف مورد نظر را پیدا می‌کند. فرآیند انتخاب ویژگی در تمامی روش‌ها به بخش‌های زیر تقسیم بندی می‌شود (شکل ۱):

۱- **روش جستجو(Search Method)**: تابع تولید کننده زیرمجموعه‌های کاندید شده را برای روش مورد نظر پیدا می‌کند.

۲- **مدل ارزیابی (Attribute Evaluator)**: زیر مجموعه کاندید شده را بر اساس روش جستجو ارزیابی

[1] Feature Selection Algorithms

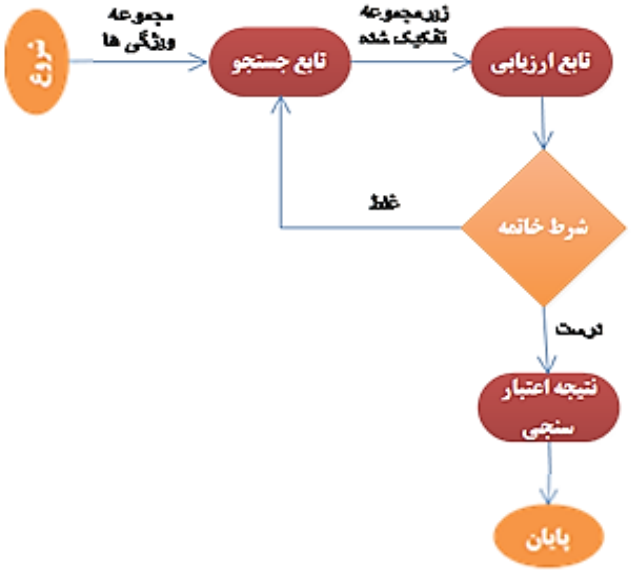
کرده و یک عدد به عنوان میزان مزیت روش باز می‌گرداند. روش‌های مختلف تلاش می‌کنند تا زیرمجموعه‌ای را پیدا کنند که این مقدار را بهینه کند.

۳- شرط خاتمه: برای تصمیم‌گیری در مورد زمان توقف الگوریتم استفاده می‌شود (شکل ۲).

سپس به کمک تابع تعیین اعتبار، اعتبار زیرمجموعه انتخاب شده را تعیین می‌کند. برای این کار می‌توان از داده‌های نمونه برداری شده و یا مجموعه داده‌های شبیه سازی شده استفاده نمود. به بیان دیگر، تابع تعیین اعتبار جزئی از فرآیند انتخاب ویژگی نیست، اما در عمل باید یک زیرمجموعه ویژگی را در شرایط مختلف امتحان نمود تا شرایط مورد نیاز برای حل مساله موردنظر احراز شود [۷].

در بخش انتخاب ویژگی کتابخانه weka سه نوع روش جستجو وجود دارد که شامل Bestfirst، GeridyStepwise و Ranker می‌باشد. این کتابخانه شامل مدل‌های ارزیابی مختلف است که در این تحقیق مطابق (شکل ۱) از مدل‌های CfsSubsetEval، CorrelationAttributeEval، InfoGainAttributeEval، ReliefFAttributeEval و SymetricalAttributeEval استفاده شده است [۲۴،۲۳].

شکل ۱- فرآیند گزینش بهترین الگوریتم انتخاب ویژگی.

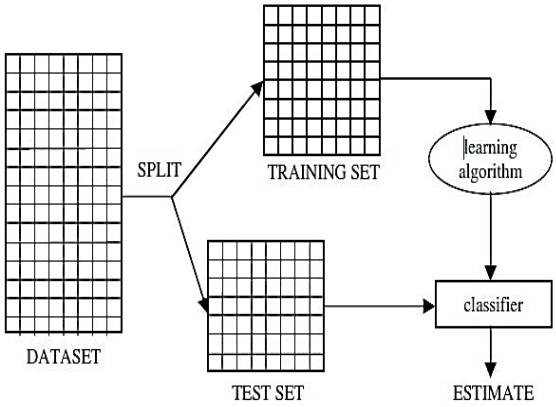


شکل ۲- نمودار فعالیت الگوریتم انتخاب ویژگی در سرده Rubus.

الگوریتم‌های دسته‌بندی[1]

یکی از مهمترین روش‌های ارزیابی کارایی الگوریتم‌های انتخاب ویژگی استفاده از الگوریتم‌های دسته‌بندی است، به طوری که هرچه دقت دسته‌بندی با استفاده از مجموعه‌ای از ویژگی‌ها بیشتر باشد بیانگر این است که آن ویژگی‌ها از کارایی بیشتری برخوردارند.

جهت ساخت یک دسته‌بندی از روش آموزش و تست (Train and Test Method) مطابق (شکل ۳) استفاده می‌شود که به موجب آن مجموعه داده‌های در دسترس به مجموعه‌های آموزشی (Training Set) و مجموعه‌های تست (Testing Set) تقسیم می‌شود [۱۹،٤]. در این تحقیق مطابق (شکل ۱) از الگوریتم‌های دسته‌بندی MLP[2]، Naive Bayes، C4.5، SMO[3]، KNN[4] و Random Forest استفاده شده است [۱٤،۲۱،۲۵،۲٦،۲۸،۳۰].



شکل ۳- عملکرد الگوریتم‌های دسته‌بندی [۱۹،٤].

نتایج و بحث

در این پژوهش نرم‌افزاری با استفاده از کتابخانه Weka و به زبان جاوا نوشته شد که داده‌ها را از فایل ورودی دریافت کرده و عملیات پردازشی (شامل انتخاب ویژگی و دسته‌بندی) را روی داده‌ها اعمال می‌کند، سپس گزارشی (شامل ویژگی‌های انتخاب شده و دقت و خطای دسته‌بندی با استفاده از آن ویژگی‌ها) را به صورت یک فایل اکسل بر می‌گرداند. جهت کاهش تعداد ویژگی‌ها، داده‌ها با استفاده از تعدادی الگوریتم‌های انتخاب ویژگی واکاوی شده و

[1] Classification Algorithms
[2] Multilayer Perceptron
[3] Sequential Minimal Optimization
[4] K-Nearest Neighbors

هر الگوریتم تعدادی ویژگی را از بین ٤٦ ویژگی اولیه انتخاب نمود. روش جستجوی Ranker به هر ویژگی یک امتیاز اختصاص می‌دهد که هرچه رتبه یک ویژگی بیشتر باشد بدین معنی است که آن ویژگی از اهمیت بیشتری برخوردار می‌باشد. بنابراین جهت انتخاب ویژگی‌های برتر، یک حد آستانه(Threshold) برای رتبه تعیین شده است. به این صورت که اگر امتیاز یک ویژگی از حد آستانه بیشتر شد، آن ویژگی انتخاب می‌شود. حد آستانه نیز بدین شکل بدست آمده است که ویژگی‌های انتخاب شده توسط آن حد آستانه بیشترین دقت را در دسته‌بندی داشته باشند.

تمامی ویژگی‌های انتخاب شده در تجزیه واریانس در سطح یک درصد معنی‌دار می‌باشند. در الگوریتم‌های دسته‌بندی، از ۷۰ درصد داده‌ها (٨٦ نمونه) جهت آموزش و ۳۰ درصد (۳۷ نمونه) باقی مانده برای حالت آزمایش استفاده شده است. به منظور افزایش دقت اعتبارسنجی نتایج، روش Repeated Random Subsampling با ۲۰۰ تکرار به کار گرفته شده است [۱۵]. این روش اعتبار سنجی استفاده شده است زیرا تمامی حالت‌های ممکن برای آزمایش وآموزش الگوریتم دسته‌بندی را در نظر می‌گیرد.

(جدول ٤) میزان دقت و خطای الگوریتم‌های مختلف را نشان می‌دهد. برای ارزیابی دسته‌بندی‌های ذکر شده و انتخاب بهترین دسته‌بندی، از ضریب کاپای کوهن [۵] و ریشه خطای مربع نسبی [۸] استفاده می‌شود. ضریب کاپای کوهن یک معیار آماری توافق درون ارزیاب بین دو اندازه‌گیری برای دسته-بندی موارد مشابه است. بنابراین هرچه ضریب کاپای کوهن یک الگوریتم بیشتر باشد، میزان دقت آن الگوریتم بالاتر است. همچنین، معیار ریشه خطای مربع نسبی، خطا بر حسب درصدی از میانگین واقعی دقت یک الگوریتم است که هرچه این معیار کمتر باشد میزان دقت آن الگوریتم بیشتر است [٨،٦].

جدول ٤- نتیجه دسته‌بندی الگوریتم‌ها

الگوریتم دسته‌بندی	الگوریتم انتخاب ویژگی		کاپای کوهن	ریشه مربع خطای نسبی(٪)	تعداد ویژگی‌های انتخاب شده	دقت دسته‌بندی	
	روش جستجو	مدل ارزیابی				درصد درست	درصد غلط
C4.5	Best-First	Cfs-Subset-Eval	۰/۸۳۳	۵۴/۵۷	۱۷	۸۷/۲۹	۱۲/۷۱
	Greedy-Stepwise	Cfs-Subset-Eval	۰/۸۵۴	۵۱/۱۴	۲۱	۸۸/۹۱	۱۱/۰۹
	Ranker	Info-Gain-Attribute-Eval	۰/۸۱۸	۵۴/۹۹	۶	۸۵/۹۴	۱۴/۰۶
		Gain-Ratio-Attribute-Eval	۰/۸۳۶	۵۴/۱۸	۸	۸۷/۵۶	۱۲/۴۴
		ReliefF-Attribute-Eval	۰/۷٦۷	٦۱/۵۸	۷	۸۲/٤۳	۱۷/۵۷
		Correlation-Attribute-Eval	۰/۷٦٤	۵۷/۹۲	۵	۸۱/۸۹	۱۸/۱۱
		Symmetrical-Attribute-Eval	۰/۷۵۸	٦۳/۱٦	۱۲	۸۱/۳۵	۱۸/٦۵
	–	–	۰/۸۰۳	۵۸/۲۸	٤٦	۸٤/۵۹	۱۵/٤۱
SMO	**Best-First**	**Cfs-Subset-Eval**	**۰/۸۷۵**	**۹۲/۰۵**	**۱۷**	**۹۰/۵٤**	۹/٤٦
	Greedy-Stepwise	Cfs-Subset-Eval	۰/۸۸۵	۹۱/۹٦	۲۱	۹۱/۳۵	۸/٦۵
	Ranker	Info-Gain-Attribute-Eval	۰/۸٦۷	۹۲/۰٤	۲۳	۸۹/۹۹	۱۰/۰۱

		Gain-Ratio-Attribute-Eval	۰/۸۹٦	۹۱/۹۱	۲۹	۹۲/۱٦	۷/۸٤
		ReliefF-Attribute-Eval	۰/۸٤۳	۹۲/۲۹	۱۸	۸۸/۱۰	۱۱/۹
		Correlation-Attribute-Eval	۰/۹۰۰	۹۱/۹٤	٤۳	۹۲/٤۳	۷/۵۷
		Symmetrical-Attribute-Eval	۰/۸۸۵	۹۱/۹۵	۲٦	۹۱/۳۵	۸/٦۵
	–	–	۰/۸۹۵	۹۱/۲۸	٤٦	۹۱/۸۹	۸/۱۱
Nave Bayes	Best-First	Cfs-Subset-Eval	۰/۹۰۷	٤۱/۰۹	۱۷	۹۲/۹۷	۷/۰۳
	Greedy-Stepwise	Cfs-Subset-Eval	۰/۹۰۰	٤۲/۳۲	۲۱	۹۲/٤۳	۷/۵۷
	Ranker	Info-Gain-Attribute-Eval	۰/۸۲۳	۵۵/۸۱	۱۳	۸٦/٤۹	۱۳/۵۱
		Gain-Ratio-Attribute-Eval	۰/۸۲۸	۵۵/۱۳	۲۳	۸٦/۷۵	۱۳/۲۵
		ReliefF-Attribute-Eval	۰/۸۳۸	۵۲/۸۱	۱۳	۸۷/۵٦	۱۲/٤٤
		Correlation-Attribute-Eval	۰/۸٤۸	۵۲/٦۲	۲٤	۸۸/۳۷	۱۱/٦۳
		Symmetrical-Attribute-Eval	۰/۸٤۳	۵۲/۳۹	۱۹	۸۸/۱۰	۱۱/۹
	–	–	۰/۸۳۱	۵٤/۸٤	٤٦	۸۷/۰۲	۱۲/۹۸
Multilayer Perceptron	**Best-First**	**Cfs-Subset-Eval**	**۰/۸۳٤**	**٤۷/۹۳**	**۱۷**	**۸۷/۲۹**	۱۲/۷۱
	Greedy-Stepwise	Cfs-Subset-Eval	۰/۸٦۵	٤٤/۸۱	۲۱	۸۹/۷۲	۱۰/۲۸
	Ranker	Info-Gain-Attribute-Eval	۰/۷٤۲	٦۱/۹٦	٦	۸۰/۲۷	۱۹/۷۳
		Gain-Ratio-Attribute-Eval	۰/۸۱۷	۵۵/۰۳	۸	۸۵/۹٤	۱٤/۰٦
		ReliefF-Attribute-Eval	۰/۸۷۳	٤٤/۲۰	٤۱	۹۰/۲۷	۹/۷۳
		Correlation-Attribute-Eval	۰/۷۷۰	۵۷/۷٤	۵	۸۲/٤۳	۱۷/۵۷
		Symmetrical-Attribute-Eval	۰/۷۹۵	۵٦/۸۱	٤	۸٤/۳۲	۱۵/٦۸
	–	–	۰/۸۸۱	٤۲/۰۹	٤٦	۹۰/۸۱	۹/۱۹
K-nearest neighbors	Best-First	Cfs-Subset-Eval	۰/۸۸۲	٤٤/۳۰	۱۷	۹۱/۰۸	۸/۹۲
	Greedy-Stepwise	Cfs-Subset-Eval	۰/۸۷۹	٤٤/۰۱	۲۱	۹۰/۸۱	۹/۱۹
	Ranker	Info-Gain-Attribute-Eval	۰/۷۸۵	۵٦/۳۳	۱۵	۸۳/۵۱	۱٦/٤۹
		Gain-Ratio-Attribute-Eval	۰/۸۸۱	٤۸/۲۸	۱۳	۹۱/۰۸	۸/۹۲
		ReliefF-Attribute-Eval	۰/۸۷۱	٤۷/۹۳	۱۳	۹۰/۲۷	۹/۷۳
		Correlation-Attribute-Eval	۰/۸۷۵	٤۵/۲٦	۲٤	۹۰/۵٤	۹/٤٦
		Symmetrical-Attribute-Eval	۰/۸۷۱	٤٦/۸۵	۲٦	۹۰/۲۷	۹/۷۳
	–	–	۰/۸۷۷	٤٤/٦۸	٤٦	۹۰/۵٤	۹/٤٦
Random Forest	**Best-First**	**Cfs-Subset-Eval**	**۰/۹۰۰**	**٤۳/٦۵**	**۱۷**	**۹۲/٤۳**	۷/۷۵
	Greedy-Stepwise	Cfs-Subset-Eval	۰/۹۰۳	٤٤/۳٦	۲۱	۹۲/۷۰	۷/۳۰
	Ranker	Info-Gain-Attribute-Eval	۰/۹۲۱	٤٤/۲۷	۲۸	۹٤/۰۵	۵/۹۵
		Gain-Ratio-Attribute-Eval	۰/۹۱٤	٤٤/۸٦	٤۰	۹۳/۵۱	٦/٤۹
		ReliefF-Attribute-Eval	۰/۹۱۸	٤٤/۷۰	۳۰	۹۳/۷۸	٦/۲۲
		Correlation-Attribute-Eval	۰/۸۹۷	٤۷/۰٤	۲٤	۹۲/۱٦	۷/۸٤
		Symmetrical-Attribute-Eval	۰/۹۱۱	٤۵/۱۲	۳۹	۹۳/۲٤	٦/۷٦
	–	–	۰/۹۰۸	٤۷/۳۲	٤٦	۹۲/۹۷	۷/۰۳

در مجموع با توجه به (جدول ٤) بیشترین دقت در دسته‌بندی با الگوریتم Random Forest حاصل شد، بطوری که الگوریتم فوق به دقت ۹٤/۰۵ درصد در دسته‌بندی با استفاده از ۲۸ ویژگی انتخاب شده توسط روش جستجوی Ranker و مدل ارزیابی Info Gain Attribute Eval رسید. همچنین با تمامی ٤٦ ویژگی و بدون استفاده از هیچ یک از الگوریتم‌های انتخاب ویژگی، به دقت ۹۲/۹۷ درصد در دسته‌بندی رسید. نکته قابل توجه دیگر این است که دسته‌بندی با الگوریتم MLP با استفاده از چهار ویژگی انتخاب شده توسط روش جستجوی Ranker و مدل ارزیابی Symetrical Attribute Eval به دقت ۸٤/۳۲ درصد و دسته‌بندی با الگوریتم C4.5 با استفاده از شش ویژگی انتخاب شده توسط روش جستجوی Ranker و مدل ارزیابی Info Gain Attribute Eval به دقت ۸۵/۹٤ درصد رسید.

ویژگی‌های بهینه انتخاب شده توسط برخی از الگوریتم‌های مذکور در (جدول ۵) بیان شده است. از این ویژگی‌ها می‌توان به جای تمامی ٤٦ ویژگی برای شناسایی گونه‌ها با کارایی بیشتر استفاده کرد، زیرا به دلیل تعداد ویژگی کمتر و دقت بیشتر در دسته‌بندی (بجز دسته‌بندی MLP) با هزینه بسیار کمتر می‌توان گونه‌ها را شناسایی کرد. سه ویژگی رنگ گلبرگ، تعداد برگچه و نوع کرک در بین الگوریتم‌ها مشترک هستند و ویژگی شکل خار شاخه گل دهنده توسط اکثر الگوریتم‌ها انتخاب شده است. بنابراین چهار ویژگی فوق از پتانسیل بالایی برای انتخاب شدن به عنوان کلید شناسایی بین گونه‌های مختلف تمشک برخوردارند. تمامی این ویژگی‌ها کیفی هستند و جهت بدست آوردن آنها نیازی به هزینه اندازه‌گیری آزمایشگاهی نمی‌باشد.

جدول ۵- ویژگی‌های بهینه انتخاب شده

الگوریتم دسته‌بندی	الگوریتم انتخاب ویژگی		دقت دسته‌بندی (٪)	تعداد ویژگی انتخاب شده	ویژگی‌ها	
	روش جستجو	مدل ارزیابی			کیفی	کمی
Random Forest	Ranker	Info Gain Attribute Eval	۹٤/۰۵	۲۸	شکل خار شاخه گل دهنده، رنگ گلبرگ، تعداد برگچه، نوع کرک، فرم ساقه، نوک کاسبرگ، رنگ سطح تحتانی برگچه، کرک بساک، پوشش کرک در سطح فوقانی برگچه، رنگ میله، شکل گوشوارک، نوع کرک کاسبرگ، نوع کرک دمبرگ، نوع گل آذین، نوع کرک سطح فوقانی، نوک برگچه انتهایی، فرم رویشی، حالت شاخه گل دهنده، پوشش کرک در سطح تحتانی، غده کاسبرگ، شکل خار دمبرگ، نوک برگچه جانبی، غده در گل آذین، حضور غده درساقه	طول دمبرگ، طول برگچه جانبی، طول دمبرگچه انتهایی، طول برگچه انتهایی
K-nearest neighbors	Ranker	Gain Ratio Attribute Eval	۹۱/۰۸	۱۳	نوع گل آذین، فرم رویشی، حالت شاخه گل دهنده، رنگ گلبرگ، نوع کرک کاسبرگ، حضور غده در ساقه، نوع کرک، نوک کاسبرگ، کرک بساک، تعداد برگچه، رنگ میله، غده کاسبرگ، غده در گل آذین	-
C4.5	Ranker	Info Gain Attribute Eval	۸۵/۹٤	٦	رنگ گلبرگ، تعداد برگچه، نوع کرک، شکل خار	طول دمبرگ

	شاخه گل دهنده، فرم ساقه					
–	رنگ گلبرگ، تعداد برگچه، نوع کرک، شکل خار شاخه گل دهنده	٤	٨٤/٣٢	Symetrical Attribute Eval	Ranker	MLP

References

[1] Aalders, L. E. and Hall I. V. 1966. A Cytotaxaonomic survey of the native blackberries of Nova Scotia.Canadian Journal of Genetics and Cytology 8: 528-532.

[2] Ali, A.S.O., Malik, A.S. and Aziz, A. 2013. A geometrical approach for age-invariant face recognition. International Visual Informatics Conference. – Springer 81-96.

[3] Ballington, JR. Luteyn, MM. Thompson, K. Romoleroux, K. and Castillo, R. 1993. *Rubus* and Vacciniaceous germplasm resources in the Andes of Ecuador. Plant Genetic Resources newsletter 93: 9-15.

[4] Bramer, M. 2007. Principle of data mining. Springer.

[5] Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics 22: 249–254.

[6] Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. Educational and psychological measurement, 20: 37-46.

[7] Dash, M. and Liu, H. 1997. Feature selection for classification. Intelligent data analysis 1:131-156.

[8] Diaz, F., & Jones, R. 2004. Using temporal profiles of queries for precision prediction. International ACM SIGIR conference on Research and development in information retrieval 18-24.

[9] Focke, W. O. 1910. Species Ruborum. Monographiae Generic. Rubi Prodromus. Bibliotheca Botanica 17: 1-120.

[10] Focke W. O. 1911. Species Ruborum. Monographic Genesis. Rubi Prodromus. Pars I, Stuttgart.

[11] Focke, W. O. 1914. Species Ruborum. Monographic Genesis. Rubi Prodromus. Pars I-II. Stuttgart.

[12] Gu, Y., C. M. Zhao, W. Jin, and W. L. Li. 1993. *Rubus* resources in Fujan and Hunan provinces. Acta Horticulturae 345: 117-125.

[13] Gustafsson, A. 1942. The origin and properties of the European blackberry flora. Hereditas 28: 249-277.

[14] Gardner, M. W., Dorling, S. R. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627-2636.

[15] Han, J., Pei, J. and Kamber, M. 2011. *Data mining: concepts and techniques*. Elsevier.

[16] Hummer, K. E. 1996. *Rubus* diversity. Hort Science 31: 182-183.

[17] Ian H. Witten and Eibe Frank. 2005. Data Mining Practical Machine Learning Tools and Techniques.

[18] Kasalkhe, R., Jorjani, E., Sabori, H., Sattarian, A., Habibi, M. 2016. Biosystematic study of *Rubus* L. (Rosaceae) in North of Iran. MSc thesis. University of Gonbad-e-Kavous, 276 pp.

[19] Kantardzic, M. 2003.Data Mining: Concepts, models, methods, and algorithms. Wiley-Interscience.

[20] Khatamsaz, M. 1992.Flora of Iran (Rosacea).-Research Institute of forests and Rangelands 6: 20-35.

[21] Larose, D. T. 2005. k-nearest neighbor algorithm. *Discovering Knowledge in Data: An Introduction to Data Mining*, 90-106.

[22] Lu, L. T. 1983. A study on the genus *Rubus* of china. Actaphyto taxonomic sinica 21: 13-25.

[23] Published on the Internet https://weka.waikato.ac.nz/explorer [accessed 13 November 2017].

[24] Published on the Internet http://weka.sourceforge.net/doc.dev/weka/attributeSelection/AttributeEvaluator.html [accessed 13 November 2017].

[25] Platt, J. 1998. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Microsoft Research.

[26] Quinlan, J.R. 2014. C4.5: programs for machine learning. Elsevier 58-60.

[27] Richards, A. J., J. Kirschner, J. Stepanek, and K. Marhold. 1996. Apomixis and taxonomy: an introduction. Folia Geobotanica phytotaxonomica 31: 281-282.

[28] Rish, I. 2001. An empirical study of the naive Bayes classifier. IJCAI Workshop.

[29] Robertson, K. R. 1974. The genera of Rosaceae in the southe astern United States. Journal of the Arnold Arboretum 55: 352-360.

[30] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., Feuston, B. P. 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947-1958.

[31] Thompson, M. M. 1995. Chromosome number of *Rubus* species at the National Clonal Germplasm Repository. Hort Science 30: 1447-1452.

[32] Weber, H. E. 1995. Die Gattung *Rubus* L. im nordwestlichen Europa. Phanerogamarum Monographiae Tomus VII. J. Cramer, Lehre, Germany.

[33] Remagnino, P., Mayo, S., Wilkin, P., Cope, J. and Kirkup, D., 2016. Computational Botany. Springer Berlin Heidelberg:.

[34] Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., Benítez, J.M. and Herrera, F., 2016. Data discretization: taxonomy and big data challenge. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 6(1), pp.5-21.