ارائه یک روش ترکیبی مبتنی بر بیز ساده به منظور طبقه‌بندی داده‌ها

محمدرضا کیوان‌پور[۱]، فهمیه باعثی[۲]

[۱] عضو هیئت علمی دانشگاه الزهرا (س)
Keyvanpour@Alzahra.ac.ir

[۲] دانشجوی کارشناسی ارشد، دانشگاه الزهرا (س)
Baesi.Fahimeh@Gmail.Com

چکیده

طبقه‌بند بیز ساده یکی از کاراترین الگوریتم‌های طبقه‌بندی است. سادگی و دقت این روش زمانیکه پیش فرض عدم وابستگی مشخصه‌ها نقض شود افت پیدا می‌کند. در این مقاله روشی مبتنی بر درخت تصمیم‌گیری برای رفع مشکل بیز ساده در مجموعه‌هایی با مشخصه‌های وابسته ارائه می‌شود. در روش ارائه شده (Information Gain - Naïve Bayes) مشخصه‌هایی که در سطح اول درخت تصمیم‌گیری دارای کمترین قدرت تفکیک باشند با توجه به تعداد کل مشخصه‌ها حذف شده و مشخصه‌های باقیمانده به بیز ساده داده می‌شود. انجام فرایند پیش پردازش مشخصه‌ها براساس روش IG-NB منجر به بهبود عملکرد الگوریتم بیز ساده در دامنه‌هایی با مشخصه‌های وابسته به هم می‌شود. علاوه براین روش IG-NB در این پژوهش، به دلیل حذف مشخصه های کم اهمیت، سرعت محاسبه الگوریتم را برای مشخصه‌های باقی مانده افزایش می‌دهد. روش IG-NB روی شش مجموعه از مجموعه داده‌های UCI تست شده و نتایج بدست آمده نشانگر کارایی قابل قبول روش می باشد.

کلمات کلیدی

طبقه‌بندی، بیز ساده، درخت تصمیم‌گیری،آنتروپی.

۱- مقدمه

در پژوهش‌های مختلف طبقه‌بندهای گوناگونی استفاده می‌شوند. در این میان دو مورد از کاربردی‌ترین و کاراترین طبقه‌بندها، درخت تصمیم‌گیری ID3 و یادگیری بیز ساده[۱] (NB) [۱] [۲] هستند.
تحقیقات بسیاری روی موضوع ویژگی‌های نامناسب و مزایای انتخاب ویژگی‌های مناسب به منظور افزایش کارایی طبقه‌بند بیز ساده انجام شده است [۲]. پازانی[۲] روشی را برای ترکیب دو (یا تعداد بیشتری) ویژگی وابسته به هم بکار می‌برد. ویژگی مرکب بدست آمده، وابستگی بین ویژگی‌ها را نشان می‌دهد [۲]. روش دیگر برای بهبود طبقه‌بند بیز ساده، تست‌هایی است که با استفاده از مجموعه‌ای از طبقه‌بندها برای توجه بیشتر به نمونه‌هایی که نادرست طبقه‌بندی شده‌اند، انجام می‌شود [۱]. اما در [۳] نشان داده شده است که به طور متوسط در دامنه‌های طبیعی روش ارائه شده در [۱] با مشکل مواجه می‌شود.

در[۴] مارک هال[۳] و ابی فرانک[۴] به منظور بهبود عملکرد بیز ساده و جدول تصمیم‌گیری از ترکیب این دو روش استفاده کرده و روش semi-naïve Baysian ranking را پیشنهاد می‌دهند.

بیز ساده نسبت به ویژگی‌های زائد و نامناسب بسیار حساس است. وابستگی بسیار زیاد دو یا تعداد بیشتری ویژگی باعث می‌شود آن ویژگی‌ها وزن بسیار زیادی در تصمیم‌گیری درمورد تعلق یک نمونه به کلاس خاصی پیدا کنند و این خاصیت موجب کاهش دقت در دامنه-هایی با ویژگی‌های وابسته به هم می‌شود [۲]. برخلاف بیز ساده، ویژگی‌های وابسته به هم مشکلی برای ID3 بوجود نمی‌آورند، زیرا استفاده از هر دو ویژگی برای تقسیم نمونه‌های آموزشی در ID3 امکان‌پذیر نیست و این یکی از دلایل اصلی کارایی بالاتر ID3 در دامنه‌هایی با مشخصه‌های وابسته است. به طورکلی مشاهده می‌شود که کاربرد NB ساده ولی بهبود آن با مشکلات فراوانی همراه است[۲].

ما در این مقاله، روشی را برای افزایش کارایی و بهبود سرعت اجرای طبقه‌بند بیز ساده ارائه می‌دهیم. روش IG-NB با حذف تعدادی ویژگی نامطلوب که توسط روش محاسباتی درخت تصمیم-گیری بدست می‌آید، منجر به بهبود کارایی بیز ساده می‌شود.

روش IG-NB روی شش مجموعه داده‌ی انتخابی از UCI تست شده و نتایج نشان می‌دهد که این روش حذف ویژگی‌ها منجر به بهبود کارایی طبقه‌بند بیز ساده در این مجموعه داده‌ها می‌شود. شکل(۱)، ساختار روش IG-NB برای بهبود الگوریتم یادگیری بیز ساده را نشان می‌دهد.

در ادامه، ابتدا در بخش۲ مروری بر روش یادگیری بیز ساده، ویژگی‌ها و نقاط ضعف آن داریم؛ سپس در بخش۳ به معرفی درخت - تصمیم‌گیری ID3 پرداخته، در بخش ۴ روش پیشنهادی IG-NB و در بخش‌های ۵ و ۶ به ترتیب نتایج تست الگوریتم و نتیجه‌گیری را ارائه خواهیم کرد.

۲- طبقه‌بند بیز ساده

۲-۱- توصیف الگوریتم

طبقه‌بند بیز ساده یکی از کاراترین و کاربردی‌ترین الگوریتم‌های یادگیری استنتاجی برای یادگیری ماشین و داده‌کاوی است. این طبقه-بند ساده بوده و روشی برای یادگیری نظارتی است [۵]. در برخی پژوهش‌ها نشان داده شده است که کارایی این طبقه‌بند قابل مقایسه با روش‌هایی مانند شبکه عصبی و درخت تصمیم‌گیری است.

به عنوان مثال Michine (۱۹۹۴) مقایسه کاملی بین این الگوریتم و سایر الگوریتم‌ها مانند درخت تصمیم و شبکه عصبی انجام داده است [۶]. همچنین اثبات شده است که بیز ساده سریعترین الگوریتم یادگیری است که تمام نمونه‌های ورودی آموزش را تست کرده و به کارایی بسیارخوبی در مسائل مختلف می‌رسد [۲].

یک مجموعه نمونه‌های آموزشی با برچسب کلاس و یک نمونه تست E با n مقدار مشخصه $(a_1,a_2,\ldots,a_n|c)$ را درنظر می‌گیریم و طبقه‌بند بیزین را برای طبقه‌بندی E به صورت زیر تعریف می‌کنیم:

$$c(E)=\arg_{c\in C}\max P(c)P(a_1,a_2,..a_n\,|c) \qquad (۱)$$

فرض پایه طبقه‌بندی بیز ساده این است که در هرکلاس مقادیر مشخصه‌ها از یکدیگر مستقل هستند. بنابراین با استفاده از قانون احتمالی استقلال داریم:

$$P(a_1,a_2,..a_n\,|c)=P(a_1\,|c)P(a_2\,|c)...P(a_n\,|c)$$
$$P(a_1,a_2,...,a_n\,|c)=\prod_{i=1}^{n}p(a_i\,|c) \qquad (۲)$$

با جایگذاری فرمول (۲) در (۱)، طبقه‌بند نتیجه که طبقه‌بند بیز ساده نامیده می‌شود به صورت زیر نشان داده می‌شود:

$$c_{NB}(E)=\arg_{c\in C}\max P(c)\prod_{i=1}^{n}P(a_i\,|c) \qquad (۳)$$

در این رابطه، $c_{NB}(E)$ طبقه‌بند بیز ساده روی نمونه تست E است.

شکل(۱): فرآیند توسعه در IG-NB

با استفاده از فرمول (۳) تمامی احتمالات می‌توانند مستقیما از روی داده‌های آموزشی تعیین شوند. در زمان آموزش بیز ساده جدولی یک بعدی از تخمین‌های احتمالی کلاس و جدولی دوبعدی از تخمین‌های احتمالی مقدار مشخصه شرطی تولید می‌شود. جدول یک بعدی تولید شده با کلاس‌ها و جدول دوبعدی با کلاس‌ها و مقادیر مشخصه شاخص‌گذاری می‌شوند.

۲-۲- خصوصیات بیز ساده

طبقه‌بند بیز ساده، ساختاری ساده و کارایی محاسباتی بالایی دارد. تمامی احتمالات مورد نیاز برای ساخت یک طبقه‌بند بیز ساده می تواند با یکبار بررسی پیدا شوند. همچنین این مدل به سادگی بروز می-شود، بنابراین آموزش در تعداد نمونه‌ها و مشخصه‌ها خطی بوده و این یکی از نقاط قوت بیز ساده است [۷, ۸][۵].

در مقایسه با دیگر طبقه‌بندها، بیز ساده به داده‌های آموزشی کمتری نیاز دارد. این طبقه‌بند بسیار سریع آموزش می‌بیند و به فضای ذخیره‌سازی کمی در زمان آموزش و طبقه‌بندی نیاز دارد. طبقه‌بند بیز ساده شفاف است، به سادگی اجرا می‌شود و همانند شبکه‌های عصبی و ماشین‌های بردار پشتیبان به پارامترهای زیادی نیاز ندارد [۵, ۹, ۱۰]. بیز ساده به طورطبیعی نسبت به از دست دادن مقادیر که در محاسبه احتمالات نادیده گرفته شده قوی بوده و تصمیم نهایی را تغییر نمی-دهد [۹, ۱۱].

۲-۳- نقاط ضعف الگوریتم بیز ساده

فرضیه اصلی در طبقه‌بند بیز ساده این است که برای عضویت در یک کلاس خاص، مقدار احتمال مشخصه‌های خاص مستقل از هم هستند. ولی در عمل غالبا این پیش فرض نقض می‌شود. برای مثال، در داده-های آمارگیری افراد، بسیاری از مشخصه‌ها مانند سن و درآمد به یکدیگر وابسته‌اند[۲].

بیز ساده عملا کارایی بالایی دارد و بهبود آن دشوار است. همچنین این روش طبقه‌بندی نسبت به مشخصه‌های زائد و نامرتبط حساس است. اگر دو یا تعداد بیشتری مشخصه بسیار وابسته به هم باشند، آن مشخصه‌ها وزن بسیار زیادی در تصمیم‌گیری نهایی تعلق یک نمونه به یک کلاس می‌گیرند و این منجر به کاهش صحت پیشگویی در دامنه-هایی که ویژگی‌های وابسته به هم دارند، می‌شود. بدلیل اینکه ممکن است بیز ساده وزن بالایی تحت تاثیر دو مشخصه گرفته و باقیمانده مشخصه‌ها وزن بسیار کمی بگیرند، نتیجه نهایی طبقه‌بندی ممکن است بایاس باشد[۲, ۱۲, ۱۳].

۳- درخت تصمیم‌گیری

درخت تصمیم‌گیری، درختی است که گره‌های میانی آن نشان‌دهنده انتخابی بین تعدادی پیشنهاد و هر برگ نمایانگر یک تصمیم است [۱۴].

یک درخت تصمیم‌گیری با گره ریشه شروع کرده و برطبق الگوریتم یادگیری به صورت بازگشتی در هر گره تقسیم می‌شود. نتیجه نهایی یک درخت تصمیم‌گیری را نشان می‌دهد که هر شاخه نمایش دهنده سناریوی ممکن تصمیم‌گیری و خروجی آن است.

الگوریتم یادگیری درخت تصمیم‌گیری روشی برای تقریب توابع هدف با مقادیر گسسته است. در نهایت تابع یادگیری شده با یک درخت تصمیم‌گیری نمایش داده می‌شود [۱۵] .

الگوریتم‌های یادگیریِ درخت تصمیم‌گیری مختلفی وجود دارد. ما در ادامه این پژوهش از درخت تصمیم‌گیری ID3 استفاده کرده‌ایم.

۳-۱- الگوریتم یادگیری درخت تصمیم‌گیری ID3

ID3 الگوریتمی یادگیری درخت تصمیم‌گیریِ ساده‌ای است که بوسیله روس کینلن[۵] توسعه یافت. هدف اصلی الگوریتم ID3 ساخت درخت تصمیمی با استفاده از جستجوی‌ای حریصانه و بالا به پائین در میان مجموعه‌های داده برای تست هر مشخصه در هرگره درخت است. به منظور انتخاب مشخصه‌ای که برای طبقه‌بندی مجموعه داده‌ها مفیدترین است، به معیار Information Gain نیاز است.

۳-۱-۱- آنتروپی [۵]

فرض می‌کنیم درخت تصمیم‌گیری نتیجه نمونه‌ها را به دو گروه تقسیم می‌کند، که ما آنها را P(+) و N(-) می‌نامیم.

مجموعه S را که حاوی هدف‌های مثبت و منفی است را درنظر می‌گیریم، آنتروپی S وابسته به این طبقه‌بند بولی به صورت زیر است:

$$\text{Entropy}(S) = -P(+)\log\text{-}2P(+) - P(-)\log 2P(-) \qquad (۴)$$

P(+): سهم نمونه‌های مثبت در S و P(-): سهم نمونه‌های منفی در S.

۳-۱-۲- Information Gain [۵]

برای مینیمم‌کردن عمق ID3، نیاز به انتخاب مشخصه بهینه برای جداسازی گره درخت داریم، بنابراین مشخصه‌ای با بیشترین کاهش آنتروپی بهترین انتخاب است.

برای کاهش آنتروپی وابسته به مشخصه‌ای خاص معیار Information Gain را برای جداسازی یک گره درخت تعریف می‌کنیم.

$$\text{Gain}(S,A) = \text{Entropy}(S) - \text{Sum for v from 1 to n of } \left(|S_v|/|S|\right) * \text{Entropy}(S_v) \qquad (۵)$$

می‌توان از مفهوم Gain برای رتبه‌بندی مشخصه‌ها و ساخت درخت‌های تصمیم‌گیری استفاده کرد. در هر گره مشخصه‌ای با بیشترین مقدار Gain از میان مشخصه‌ها که در مسیر ریشه و تا این گره هنوز بکار برده نشده است انتخاب می‌شود.

۴- روش پیشنهادی (IG-NB)

در بیز ساده، اگر دو یا چند مشخصه بسیار وابسته به هم باشند، این مشخصه‌ها وزن زیادی در تصمیم‌گیری نهایی درمورد تعلق یک نمونه به کلاس خاصی می‌گیرند که این امر باعث کاهش صحت پیشگویی در دامنه‌هایی با مشخصه‌های وابسته به هم می‌شود. اما روش یادگیری ID3 این مشکل را ندارد، زیرا اگر دو مشخصه وابسته به هم باشند، امکان ندارد که از هر دوی آنها برای جداکردن مجموعه آموزشی استفاده شود.

از طرف دیگر، بیز ساده نسبت به پارامترهای اولیه بسیار حساس است و باید یک مرحله یادگیری را پشت سر بگذارد که این مرحله یادگیری توسط انسان یا الگوریتمی دیگر انجام می‌شود و بسیار زمانبر است.

با توجه به این نکات، ما از معیارهایی که ID3 برای ساخت درخت تصمیم‌گیری بکار می‌برد، استفاده می‌کنیم. بدین صورت که مقدار Information Gain را در سطح اول برای تمامی مشخصه‌ها محاسبه کرده، سپس با توجه به این نکته که بیز ساده بسیار وابسته به مقادیر مشخصه‌ها است و مقادیر آنها در نتیجه نهایی بسیار موثر است مشخصه‌هایی را که مقدار Information Gain آنها کمترین مقدار را در بین دیگر مشخصه‌ها دارند حذف کرده و سپس مشخصه‌های باقیمانده را برای طبقه‌بندی به بیز ساده می‌دهیم.

به منظور افزایش کارایی در روش پیشنهادی ما به نام Information Gain- Naïve Bayesian (IG-NB)، تعداد مشخصه‌هایی که قرار است در یک مجموعه داده حذف شود، می‌تواند با توجه به تعداد مشخصه‌ها، اهمیت مشخصه‌ها و وابستگی بین مشخصه‌ها در مجموعه نمونه‌های انتخابی متغییر باشد.

۵- شبیه سازی و ارائه نتایج تجربی

در جدول (۱) نتایج تست الگوریتم‌های یادگیری بیز ساده و IG-NB روی ۶ مجموعه انتخابی از UCI آمده است. مشاهده می‌شود که در ۴ مورد بهبود با استفاده از روش ترکیبی حاصل شده است. در جدول (۲) و شکل (۴) از سه معیار FN، FP و Accuracy برای مقایسه کارایی الگوریتم‌ها روی سه مجموعه دو کلاسه monk1، monk2 و monk3 استفاده کرده‌ایم.

در شکل (۲) و (۳) مقایسه کارایی بیز ساده و IG-NB را روی چهار مجموعه داده نشان داده شده است.

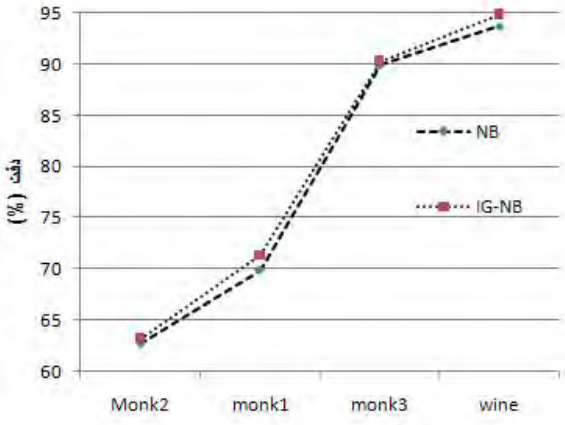
با توجه به نتایج ارائه شده، مشاهده می‌شود که در مجموعه داده‌های متفاوت، روش IG-NB با حذف ویژگی‌های زائد به دقت بالاتری نسبت به بیز ساده می‌رسد.

جدول (۱): نتایج تست بیز ساده و IG-NB برای بهبود بیز ساده

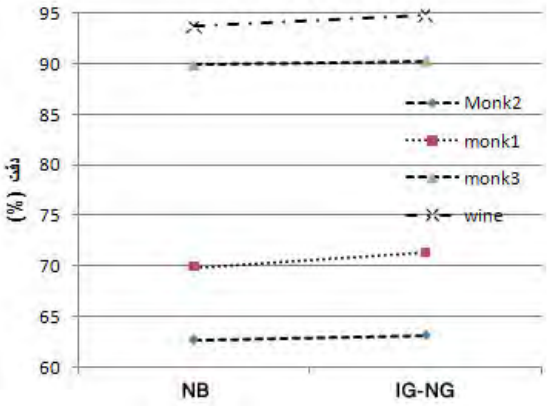
مجموعه داده	دقت NB	دقت IG-NB	تعداد داده‌های آموزشی	تعداد داده-های تست	تعداد ویژگی‌های حذف شده
Monk1	۶۹,۹۱	۷۱,۳۰	۱۲۳	۴۲۹	۲
Monk2	۶۲,۷۳	۶۳,۱۹	۱۶۸	۴۳۰	۲
Monk3	۸۹,۸۱	۹۰,۲۸	۱۲۱	۴۳۱	۲
Wine	۹۳,۷۵	۹۴,۸۲	۵۰	۱۲۸	۳
Abalone	۴۵,۹۹	۴۵,۵۱	۴۰۰	۳۷۷۷	۲
Abalone	۴۵,۹۹	۴۵,۸۸	۴۰۰	۳۷۷۷	۳

جدول(۲): نتایج تست الگوریتم روی مجموعه داده های دو کلاسه با استفاده از دومعیار FP و FN

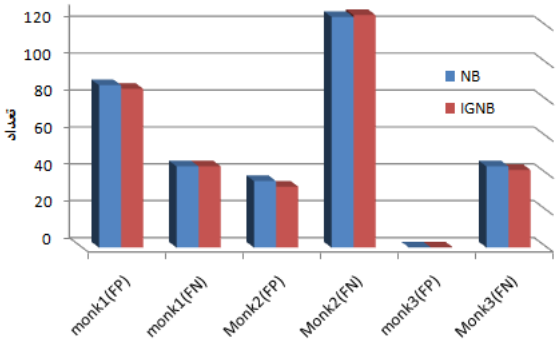
مجموعه داده ها	دقت	FN	FP
Monk1(NB)	۶۹/۹۱	۴۴	۸۸
Monk1(IGNB)	۷۱/۳۰	۴۴	۸۶
Monk2(NB)	۶۲/۷۳	۱۲۵	۳۶
Monk2(IGNB)	۶۳/۱۹	۱۲۶	۳۳
Monk3(NB)	۸۹/۸۱	۴۴	۰
Monk3(IGNB)	۹۰/۲۸	۴۲	۰



شکل(۲): مقایسه NB و IG-NB روی چهار مجموعه داده UCI



شکل (۳): نمایش NB و IG-NB روی چهار مجموعه داده UCI



شکل (۴): ارزیابی معیار FP و FN روی سه مجموعه داده

۶- نتیجه‌گیری

در این مقاله روشی برای بهبود عملکرد الگوریتم بیز ساده برای ویژگی‌های وابسته به هم پیشنهاد شد. در IG-NB پیش از اجرای الگوریتم بیز ساده، پیش پردازشی روی مجموعه داده‌ها صورت می‌گیرد. پیش پردازش مذکور با حذف ویژگی‌های نامناسب از مجموعه داده با استفاده از محاسبه Information Gain انجام می‌شود.

نتایج محاسبات نشان می‌دهد که IG-NB کارایی قابل قبولی روی مجموعه داده‌های تست بکار برده شده در این پژوهش دارد. این روش نیازی به طراحی کامل درخت تصمیم‌گیری نداشته و در همان مراحل اولیه محاسبه ID3 ویژگی‌های مطلوب و نامطلوب از یکدیگر جدا می‌شوند. بنابراین، IG-NB نسبت به روش‌های ارائه شده پیشین در این زمینه هزینه محاسباتی کمتری داشته و به نتایج بهتری نسبت به بیز ساده دارد.

مطالعات صورت گرفته در این حوزه نشان می‌دهد الگوریتم بیز ساده قابلیت بهبود بوسیله دیگر روش‌های طبقه‌بندی داده‌ها را دارد. در پژوهش‌های آتی به بهره‌گیری از ویژگی‌های مفید دیگر طبقه‌بندها و ترکیبی از آن‌ها برای بهبود عملکرد بیز ساده خواهیم پرداخت.

مراجع

[1] Elkan, C., "BOOSTING AND NAIVE BAYESIAN LEARNING". September 1997.
[2] Gunopulos, C.A.R.a.D., "Scaling up the Naive Bayesian Classifier:Using Decision Trees for Feature Selection". in Proceedings of Workshop on Data Cleaning and Preprocessing (DCAP 2002), IEEE International Conference on Data Mining (ICDM 2002).
[3] Zheng, K.M.T.a.Z., "Improving the Performance of Boosting for Naive Bayesian Classification ", in In Proceedings of the PAKDD-99, pp.296-305, Beijing, China. 1999.
[4] Frank, M.H.a.E., "Combining Naive Bayes and Decision Tables", in Association for the Advancement of Artificial Intelligence. 2008.
[5] Khadija Mohammad Al-Aidaroos, A.A.B.a.Z.O., "Naïve Bayes Variants in Classification Learning". IEEE 2010.
[6] m.mitchell, T., "Machine Learning", ed. 16. March 1,1997: McGraw-Hill Science/Engineering/Math; (March 1, 1997).
[7] E. Frank, M.H., and B. Pfahringer, "Locally Weighted Naive Bayes", in In: Proceedings of the Conference on Uncertainty in Artificial Intelligence, 2003: pp. 249–256. Morgan Kaufmann.
[8] Shang, Y.J.a.L., "RoughTree: A Classifier with Naïve-Bayes and Rough Sets Hybrid in Decision Tree Representation". 2007 IEEE International Conference on Granular Computing, 2007: pp. 221-226.
[9] S.B. Kotsiantis, I.D.Z., and P.E. Pintelas, "Machine Learning: A Review of Classification and Combining Techniques". Artificial Intelligence Review, 2006. 26(3): pp. 159-190.
[10] Liu, B., "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data". Data-Centric Systems and Applications. 2007: Springer.
[11] R. Abraham, J.B.S., and S.S. Iyengar, "Medical Datamining with a New Algorithm for Feature Selection and Naïve Bayesian Classifier". in ICIT. 2007: IEEE Computer Society, pp.44-49.
[12] Sage, P.L.a.S., "Induction of Selective Bayesian Classifiers", in in Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann. 1994: pp. 399–406.
[13] Webb, F.Z.a.G.I., "Efficient Lazy Elimination for Averaged One-Dependence Estimators", in in Proceedings of the 23rd International Conference on Machine Learning, 2006: pp. 1113-1120.
[14] S.Weiss, C.A.a., "Data Mining with Decision Trees and Decision Rules". Future Generation Computer Systems, 1997. 13:197-210.
[15] Mitchell, T.M., Machine Learning. 1997.

زیرنویس‌ها

[1] Naïve Bayes Learning
[2] Pazzani
[3] Mark Hall
[4] Eibe Frank
[5] Ross Quinlan(1993)