علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۷، شماره ۱، ۱۳۹۸
صفحات ۴۳-۳۵
مقاله عادی

توسعه ساختار طبقه‌بندی کننده بیز ساده با هدف مدل‌سازی وابستگی متقابل شرطی ویژگی‌ها

نیما شیری هرزویلی[۱]، ساسان حسینعلی زاده[۲*]

[*]نویسنده مسئول، دریافت: ۹۷/۰۲/۱۶، بازنگری: ۹۷/۰۷/۰۹، پذیرش: ۹۷/۱۱/۰۹

[۱]دانش‌آموخته کارشناسی ارشد، مهندسی کامپیوتر ، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه آزاد اسلامی، واحد قزوین، ایران

[۲] استادیار، مهندسی کامپیوتر، پژوهشکده فناوری اطلاعات، پژوهشگاه ارتباطات و فناوری اطلاعات، تهران، ایران

چکیده

طبقه‌بندی کننده بیز ساده به دلیل کارایی بالا در پیش‌بینی و سادگی در ساخت مورد توجه محققین بسیاری قرارگرفته است. بنیان این طبقه‌بندی کننده بر اساس استقلال شرطی متغیرها (ویژگی‌ها) به شرط کلاس است. اگرچه، به دلیل وابستگی متقابل بین ویژگی‌ها این فرض در کاربردهای واقعی این طبقه‌بندی کننده صادق نیست. از این رو، در این مقاله از مفهوم متغیرهای پنهان برای ارائه مدلی تحت عنوان "طبقه‌بندی کننده بیز ساده آمیخته با متغیر پنهان (MLNB)" به منظور کاهش فرض استقلال شرطی و مدل‌سازی ویژگی‌ها ارائه شده است. الگوریتم امید ریاضی- بیشینه (EM) به منظور تخمین پارامترهای مدل استفاده شده است. شبیه سازی‌ها بر روی ۷ مجموعه داده از مخزن یادگیری ماشین دانشگاه کالیفورنیا ایروایین نشانگر این است که روش پیشنهادی عملکرد قابل توجه‌ای بر اساس صحت طبقه‌بندی، ناحیه زیر منحنی ROC و معیار F-measure در مقایسه با توسعه‌های اخیر بیز ساده دارد.

کلمات کلیدی: طبقه‌بندی کننده بیز ساده، استقلال شرطی، متغیرهای پنهان، شبکه‌های بیزین.

۱- مقدمه

طبقه‌بندی کننده بیز ساده[۱] (NB) به دلیل کارایی پیش‌بینی بالا و سادگی در ساخت مورد توجه محققیق بسیاری قرار گرفته است [۱]. این طبقه‌بندی کننده مبتنی بر فرض استقلال شرطی ویژگی‌ها به شرط کلاس است. به دلیل وجود وابستگی بالا میان ویژگی‌ها، این فرض در دنیای واقعی صادق نیست و موجب افزایش بایاس [۲] در پیش‌بینی می‌شود که به منظور از بین بردن آن فرض استقلال شرطی متغیرها به شرط کلاس باید کاهش پیدا کند.

روش‌های بسیاری به این منظورارائه شده است [۳-۵] . با توجه به تحقیقات انجام شده در [۶, ۷]، کاربرد بیز ساده تقویت شده درختی[۲] (TAN) [۳] می‌تواند به عنوان آخرین تحقیقات انجام شده در این زمینه قلمداد شود. در TAN، متغیر کلاس و یک ویژگی تصادفی در شبکه به عنوان والد دیگر ویژگی‌ها در نظر گرفته می شوند (توجه شود که هیچ حلقه جهت داری شکل نمی گیرد). از آنجایی که این الگوریتم نیازمند مشخص کردن جهت لینک‌ها در شبکه است، نمی توان از یادگیری ساختار در این الگوریتم صرف نظر کرد. علیرغم زمانی که به منظور یادگیری ساختار TAN مورد نیاز است، این طبقه‌بندی کننده از مشکل برابری راستی آزمایی نیز رنج می برد [۸]. در رویکرد TAN، لینک‌های جهت دار جدید به منظور مدل‌سازی وابستگی ویژگی‌ها به ساختار شبکه بیز ساده اضافه می‌شوند. این در حالی است که ساختار آن به ساختار درختی محدود می شود. این لینک‌های جهت دار ممکن است به عنوان رابطه علیت بین ویژگی‌های مرتبط تفسیر شوند. بنابراین ارائه مدلی که از پیچیدگی یادگیری ساختار صرف نظر کرده و در عین حال توانایی مدل‌سازی ویژگی‌ها (کاهش فرض استقلال شرطی) را داشته باشد انگیزه اصلی این مقاله می باشد.

از این رو در این مقاله از یک متغیر پنهان به منظور ارائه مدلی برای کاهش فرض استقلال شرطی استفاده شده است. در این مدل، متغیر پنهان در امتداد متغیر کلاس به عنوان والد ویژگی‌ها در ساختار شبکه قرار دارد. توجه شود که متغیر پنهان والدی ندارد به طوری که هم راستا با متغیر کلاس در شبکه قرار دارد. از الگوریتم امید ریاضی – بیشینه[۳] (EM) [۹] به منظور تخمین پارامترهای مدل استفاده شده است. به ارزیابی مدل ارائه شده از ۷ مجموعه داده عمومی یادگیری ماشین اخذ شده از مخزن یادگیری ماشین دانشگاه کالیفورنیا ایروین (UCI) استفاده شده است. از طبقه‌بندی کننده‌های NB [۱۰]، TAN [۳]، و بیز ساده پنهان[۴] (HNB) [۱۱]، و تجمیع تخمین زننده های تک وابستگی[۵] (AODE) به منظور مقایسه با روش پیشنهادی استفاده شده است. نتایج از آزمایشات انجام شده نشانگر این است که مدل ارائه شده کارایی قابل توجهی در مقایسه‌ای با توسعه‌های بیز ساده بر اساس صحت طبقه‌بندی، ناحیه زیر منحنی ROC، و F-measure دارد.

۲- روش‌های پیشین

تلاش‌های بسیاری برای کاهش فرض استقلال شرطی بیز ساده صورت گرفته است. به طور کلی این رویکردها به چهار دسته تقسیم بندی می شوند: ۱) رویکردهای انتخاب ویژگی ۲) رویکرد های وزندهی ویژگی ها ۳) رویکرد های یادگیری محلی ۴) رویکردهای توسعه ساختار.

۲-۱- انتخاب ویژگی

رویکرد انتخاب ویژگی با حذف ویژگی‌های اضافی یا نامناسب از مجموعه داده‌های آموزشی و تنها با انتخاب ویژگی‌هایی که در فاز یادگیری اطلاعات بیشتری را دارند، اقدام به بهبود بیز ساده می کند. در واقع هر نوع از الگوریتم های بهبود آن نوعی از بیز ساده هستند که تنها از زیر مجموعه‌ای از ویژگی‌های معین در فرآیند پیش‌بینی را انتخاب می کند. واضع است که چالش اصلی این است که چطور زیر مجموعه‌ای کارا از ویژگی‌ها را انتخاب کنیم. به منظور دست یافتن به این مهم، الگوریتم‌های انتخاب ویژگی متعددی ارائه شده اند و بهبود قابل توجه‌ای در برابر بیز ساده از خود نشان داده اند.

جیانگ [۱۲] الگوریتمی با نام بیز ساده تکاملی[۶] (ENB) ارائه کردند که از الگوریتم ژنتیک به منظور انتخاب یک زیر مجموعه از فضای کل ویژگی‌ها استفاده می کند. این روش از صحت طبقه‌بندی بیز ساده به منظور اعمال یک تابع برازش برای ارزیابی زیر مجموعه های جایگزین ویژگی‌ها استفاده کرده و یک فرضیه فردی را با صحت طبقه‌بندی بیشینه بعد از تعداد ثابتی از نسل‌ها انتخاب می کند.

۲-۲- وزن دهی ویژگی ها

بر خلاف رویکرد های انتخاب ویژگی که ویژگی های اضافی یا نامناسب را به کلی حذف می کنند، این رویکرد بصورت مجزا و با توجه به سهم آن ویژگی در طبقه‌بندی اقدام به وزن دهی ویژگی ها می کند. مدلی که حاصل می شود بیز ساده وزن دار نام دارد[۷] (WNB) [۱، ۱۳، ۱۴] که به منظور طبقه‌بندی نمونه های تست استفاده می شود. واضح است که چگونگی یادگیری وزن ها یک مسئله حیاتی بوده و مورد توجه محققین بسیاری قرار گرفته است. برای مثال ژانگ و شنگ [۱۵] روش های وزن دهی ویژگی بسیاری رو کشف کردند: روش های نسبت دست یابی، تپه نوردی، زنجیزه مارکوف مونه کارلو، روش تپه نوردی آمیخته با روش نسبت دستیابی و ترکیب روش زنجیه مارکوف مونته کارلو و روش نسبت دستیابی. هال [۱۶] یک الگوریتم مبتنی بر درخت تصمیم برای وزن دهی ویژگی ها ارائه کرد. در این روش فرضیه این است که وزن منتسب شده به یک ویژگی برای پیش بینی باید بصورت معکوس به درجه وابستگی با ویژگی های درگیر مرتبط شود. روش ارائه شده آن ها درجه وابستگی ویژگی به وسیله ساخت درخت تصمیم هرس نشده و نگاه به عمق درخت را تخمین می زند، یک روند بگینگ به منظور پایدار سازی تخمین ها استفاده می شود و وزن یک ویژگی صفر در نظر گرفته شد اگر در درخت تصمیم ظاهر نشود. اگرچه وزن دهی ویژگی ها رویکرد معتبری برای بهبود بیز ساده به نظر می رسد، جستجو موثر و کارا برای یک وزن خوب حیاتی است.

۲-۳- یادگیری محلی

ایده اصلی این رویکرد ساخت بیز ساده بر اساس زیر مجموعه ای از مجموعه داده محلی بجای ساخت مدل بر اساس کل مجموعه داده آموزشی است. اگرچه استقلال شرطی ویژگی ها به شرط کلاس همیشه کل مجموعه داده را تحت تاثیر قرار داده است. می توان انتظار داشت که وابستگی درون مجموعه داده محلی ضعیف تر از کل مجموعه داده آموزشی باشد. بنابراین بیز ساده ساخته شده در زیر مجموعه داده آموزشی محلی کارایی بهتری دارد. به علاوه مشاهده شده است که کارایی بیز ساده در مجموعه داده های بزرگ افزایش نیافته است. این قابلیت باعث می شود تا یک مدل برازش شده محلی به داخل یک مدل دیگر مثل یک درخت تصمیم، یا یک نزدیک ترین همسایه ادغام شود.

در یادگیری محلی، الگوریتم نزدیک ترین همسایه[۸] خوش سازماندهی شده ترین الگوریتم می باشد. ایده تلفیق نزدیک ترین همسایه با بیز ساده بسیار ساده است. همانند تمام روش های یادگیری تنبل، داده آموزشی به سادگی ذخیره شده و یادگیری تا زمان طبقه بندی به تعویق می افتد. از زمانی که یک نمونه تست طبقه بندی شد، بیز ساده محلی با استفاده از نزدیک ترین همسایه نمونه تست آموزش داده می شود، سپس نمونه تست طبقه بندی می شود.

در سال های اخیر، محققین بسیاری نزدیک ترین همسایه را با بیز ساده تلفیق کرده اند. برای مثال فرانک [۱۷] یک الگوریتم با نام بیز ساده محلی وزن دارد[۹] ارائه کردند. در این الگوریتم در ابتدا نزدیک ترین همسایه نمونه های تست پیدا می شوند و هر کدام از آن ها بر اساس فاصله خود با مثال تست وزن دهی می شوند. سپس یک بیز ساده محلی از طریق آن وزن ها ساخته می شود.

۲-۴- توسعه ساختار

همانطور که در مقدمه اشاره شد، فرض استقلال شرطی بیز ساده در دنیای واقعی صادق نیست. بنابراین با توسعه ساختار بیز ساده و با مدل‌سازی پیچیده تر ویژگی-ها می توان تا حد قابل توجهی این فرض را کاهش داد. مدلی که در نهایت استخراج خواهد شد یک شبکه بیزین خواهد بود. از این رو نمی توان از یادگیری ساختار شبکه‌های بیزین اجتناب کرد [۱۱]. نیاز است تا محدودیت‌هایی روی یادگیری ساختار شبکه بیزین اعمال شود. طبقه‌بندی کننده TAN نمونه بارز طبقه‌بندی کننده‌ای است که با اعمال محدودیت روی ساختار شبکه بیزین به صورت قابل توجه‌ای پیچیدگی یادگیری ساختار را کاهش داده است و بهبود قابل توجه‌ای نسبت به بیز ساده دارد [۳].

مشکل اجتناب ناپذیر در TAN این است که علاوه بر اعمال محدودیت روی شبکه بیزین، نیازمند مکانیزمی برای مشخص کردن والد دوم در کنار متغیر کلاس است که خود نیازمند عملیات جستجو می باشد [۸]. بنابراین مدلی که نیاز به یادگیری ساختار نداشته باشد، و در عین حال توانایی مدل سازی وابستگی بین ویژگی‌ها را داشته باشد، بسیار حائز اهمیت است.

از این رو جیانگ همدلی تحت عنوان بیز ساده پنهان (HNB) ارائه کردند. [۱۱]. این طبقه‌بندی از مفهوم متغیرهای پنهان به منظور مدل‌سازی ارتباط علیت بین ویژگی‌ها استفاده می کند. این مدل بدون نیاز به یادگیری ساختار اقدام به بهبود ساختار بیز ساده می کند. در ساختار این روش، به غیر از متغیر کلاس که والد همه ویژگی‌ها است، هر ویژگی یک والد پنهان مختص به خود را در اختیار دارد.

تمرکز اصلی این مقاله بر رویکردهای توسعه ساختار می باشد. از این رو روش-های انتخاب شده جهت مقایسه با روش پیشنهادی از دسته روش‌های توسعه ساختار می باشند. در این مقاله، مدلی تحت عنوان "طبقه‌بندی کننده بیز ساده آمیخته با متغیر پنهان[۱۰] MLNB" ارائه شده است که از پیچیدگی‌های یادگیری ساختار اجتناب می کند و در عین حال رابطه بین ویژگی‌ها را با اضافه کردن یک متغیر پنهان به ساختار بیز ساده مدل‌سازی می کند.

۳- طبقه‌بندی کننده بیز ساده

طبقه بندی کننده بیز ساده [۱۰] یک رویکرد طبقه بندی بیزین است که بر اساس فرض استقلال شرطی ویژگی ها به شرط کلاس ایجاد می شود. در این رویکرد، با فرض داشتن یک مجموعه ویژگی $a_1,...,a_n$ و متغیر کلاس C هدف محاسبه بیشینه کردن احتمال پسین متغیر کلاس به شرط ویژگی ها (مشاهدات) است به طوری که:

$$\underset{c\in C}{argmax}\ P(C\mid a_1,...,a_n) \tag{۱}$$

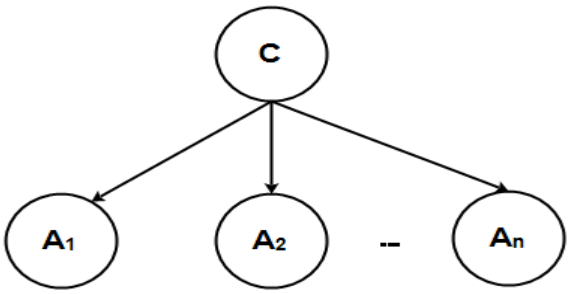
در این رویکرد احتمال پسین بر اساس قانون بیز محاسبه می شود:

$$P(C\mid a_1,...,a_n)=\frac{\prod_{i=1}^{n}P(a_1,...,a_n\mid C)P(C)}{P(a_1,...,a_n)} \tag{۲}$$

در معادله (۲) حذف مخرج تاثیر در نتیجه طبقه‌بندی ندارد. بنابراین مساله به شکل زیر ساده می شود:

$$\underset{c\in C}{argmax}\ P(C)\prod_{i=1}^{n}P(a_i\mid C) \tag{۳}$$

ساختار طبقه بندی کننده بیز ساده در شکل ۱ به تصویر کشیده شده است:



شکل ۱- ساختار طبقه بندی کننده بیز ساده

۴- طبقه‌بندی کننده بیز ساده با متغیر پنهان

مدل ارائه شده در این مقاله به گونه‌ای است که یک متغیر پنهان به ساختار بیز ساده اضافه می کند. همانطور که از شکل (۲) پیداست، هر جفت از ویژگی‌ها از طریق دو مسیر ایزوله یکی از مسیر متغیر کلاس و دیگری از مسیر متغیر پنهان به یکدیگر متصل شده اند. اگرچه مسیر متغیر کلاس در صورت مشاهده شده این متغیر مسدود می‌شود، مسیر متغیر پنهان اتصال را نگه می‌دارد. این اتصال به منظور مدل‌سازی وابستگی بین ویژگی‌ها مورد استفاده قرار می‌گیرد. از این رو مدل پیشنهادی را می‌توان به عنوان توسعه‌ای کم هزینه برای بیز ساده قلمداد کرد. بنابراین احتمال پسین متغیر کلاس را به شکل زیر می توان بازنویسی کرد:

$$\underset{c\in C}{argmax}\ P(C\mid a_1,...,a_n)= \underset{c\in C}{argmax}\ \sum_{h\in R_H}P(C,H\mid a_1,...,a_n) \tag{۴}$$

عبارت داخل سیگما را می توان با استفاده از تئوری بیز به شکل زیر بازنویسی کرد:

$$P(C,H\mid a_1,...,a_n)=\frac{P(a_1,...,a_n\mid C,H)P(C,H)}{\sum_{C,H}P(a_1,...,a_n\mid C,H)} \tag{۵}$$

مقادیر مختلفی که به متغیر H منتسب می شود با R_H نشان داده می شود. با دانستن اینکه مخرج برای تمام مقادیر کلاس ثابت است می توان از آن صرف نظر کرد. بنابراین داریم:

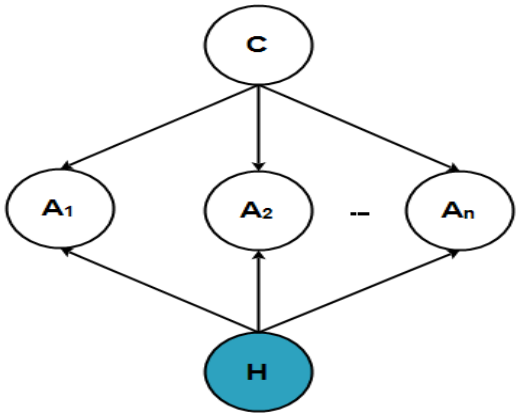
$$\underset{c\in C}{argmax}\ P(C\mid a_1,...,a_n)\propto \underset{c\in C}{argmax}\ \sum_{h\in R_H}P(a_i,...,a_n\mid C,H)P(C,H) \tag{۶}$$

با فرضیات شبکه های بیزین درمورد اتصلات واگرا[۱۱] [۱۸]، رابطه (۶) را می توان به شکل رابطه (۷) بازنویسی کرد:

$$P(a_i,...,a_n\mid C,H)=\prod_{i=1}^{n}P(a_i\mid C,H) \tag{۷}$$

به علاوه $P(C,H)$ به دلیل ارتباط d-separated به صورت حاصلضرب $P(C)$ و $P(H)$ نوشته می‌شود. در نهایت مساله طبقه‌بندی در مدل ارائه شده به شکل رابطه (۸) تعریف می شود:

$$\underset{c\in C}{argmax}\ \sum_{h\in R_H}\prod_{i=1}^{n}P(a_i\mid C,H)P(C)P(H) \tag{۸}$$



شکل ۲- ساختار مدل پیشنهادی

۴-۱- تخمین پارامتر های مدل

به منظور تخمین پارامتر های مدل، از الگوریتم امید EM ارائه شده توسط دمپستر [۹] استفاده شده است. این الگوریتم یک الگوریتم مبتنی بر تکرار است به گونه ای که تکرار ها از θ^0 در مرحله اول آغاز می شوند و تا همگرایی الگوریتم ادامه پیدا می کنند. در مرحله اول امید ریاضی داده های کامل به شرط اطلاعات در مورد پارامتر های مرحله قبل به علاوه داده های مشاهده شده محاسبه می شود. در قسمت دوم، پارامتر های تعیین می شوند به طوری که امید ریاضی بیشینه می شود.

۴-۱-۱- مرحله امید ریاضی

در مرحله امید ریاضی، امید ریاضی لگاریتم راستی آزمایی داده کامل به شرط مقادیر پارامتر ها در مرحله قبل و مجموعه داده محاسبه می شود. ما از یک نمایش فشرده برای توزیع احتمال N_X به شرط اطلاعات کلاس و متغیر پنهان به شکل زیر استفاده کرده ایم:

$$P(N_{A_i}|C,H)=\prod_{c=c_1}^{c_j}\prod_{H=h_1}^{h_k} Mc,h^{I(C=c,H=h)} \quad (۹)$$

که در آن I تابع نشانگر می باشد. بنابراین داریم:

$$P(CD)=\prod_{c=1}^{C}\prod_{h=1}^{H}\{P(n_{a_1}^{A_i},n_{a_2}^{A_i},...,n_{v_m}^{A_i}|C=c,H=h) \quad (۱۰)$$

$$P(C)P(H)\}^{I(C=c,H=h)}$$

در نهایت، با استفاده از رابطه زیر احتمال متغیر پنهان از طریق رابطه زیر محاسبه می شود:

$$P(H^{(i)}=h|\theta^{n-1},D)=\frac{P(D_j,H^{(i)}=h|\theta^{n-1})}{\sum_{h\in R_H}P(D_j,H^{(i)}=h|\theta^{n-1})} \quad (۱۱)$$

۴-۱-۲- مرحله بیشینه سازی

در این مرحله، مقادیر جدید برای پارامتر ها محاسبه می شوند به طوری که امید ریاضی بیشینه می شود. این امر با معادل قرار دادن دیفرانسیل امید های مشتق شده از قسمت قبل با عدد صفر با توجه به پارامتر ها انجام می شود. نیاز است تا ریشه معادله را بیابیم. مقادیر جدید هر توزیع چند جمله ای مرتبط با خصیصه های X_i با یافتن ریشه معادله های زیر به دست می آید:

$$j\in\{1,...,m\}\frac{\partial E\,(ll\,(CD)|\theta^{n-1},D)}{\partial\Gamma_{X_i}}\times\frac{\partial\Gamma_{X_i}}{\partial p_{v_j,c,h}}=0 \quad (۱۲)$$

با یافتن ریشه های رابطه (۱۲) داریم:

$$p_{a_j,c,h}=\frac{\sum_{i=1}^{q}\sum_{h\in R_H}\sum_{c\in R_C}n_{a_j}^{A_i}P(H^{(h)}=h|D,\theta^{n-1})}{\sum_{a_j\in R_{A_i}}\sum_{i=1}^{q}\sum_{h\in R_H}\sum_{c\in R_C}n_{a_j}^{A_i}P(H^{(h)}=h|D,\theta^{n-1})} \quad (۱۳)$$

برای یافتن احتمالات پیشین متغیر کلاس $P(C=c)$:

$$\frac{\partial\partial E(ll(CD)|\theta^{n-1},D)}{\partial\alpha}\times\frac{\partial\alpha}{\partial\alpha_j}=0 \quad (۱۴)$$

برای محاسبه احتمال متغیر کلاس از رابطه (۱۵) محاسبه می شود. بنابراین داریم:

$$\forall j\in\{1,2\}\ \alpha_j=\frac{\sum_{i=1}^{q}\sum_{h\in R_H}P(C^{(i)}=j)P(H^{(i)}=h^l|D^{(i)},\theta^{n-1})}{\sum_{i=1}^{q}\sum_{c\in R_C}\sum_{h}P(C^{(i)}=c)P(H^{(i)}=h|D^{(i)},\theta^{n-1})} \quad (۱۵)$$

در نهایت برای محاسبه احتمالات پیشین متغیر پنهان، ریشه معادله را زیر باید پیدا کنیم:

$$\forall j\in\{1,...,m\}\quad\frac{\partial\partial E(ll(CD)|\theta^{n-1},D)}{\partial\beta}\times\frac{\partial\beta}{\partial\beta_j}=0 \quad (۱۶)$$

که حاصل آن به شکل زیر محاسبه می شود:

$$\beta_j=\frac{\sum_{i=1}^{q}\sum_{C\in R_C}P(C^{(i)}=c^{(i)})P(H^{(i)}=j\,|\,D^{(i)},\theta^{n-1})}{\sum_{i}\sum_{c}\sum_{h\in R_H}P(C^{(i)}=c^{(i)})P(H^{(i)}=h|\,D^{(i)},\theta^{n-1})} \quad (۱۷)$$

۵- آزمایشات

در این قسمت آزمایشات و نتایج مورد بحث و بررسی قرار می گیرند. در بخش اول، مجموعه داده‌های استفاده شده ارائه شده است. در ادامه معیارهای ارزیابی استفاده شده در این مقاله بررسی شده‌اند. و سپس نتایج مورد حاصل از آزمایشات مورد بررسی قرار گرفته است. در انتها نتیجه گیری و کارهای آتی ارائه شده است

۵-۱- مجموعه داده‌ها

به منظور ارزیابی مدل ارائه شده با توسعه‌های طبقه بندی کننده بیز ساده، از ۷ مجموعه داده استخراج شده از مخزن یادگیری ماشین دانشگاه کالیفورنیا ایروین [۱۹] استفاده شده است. مجموعه داده های انتخاب شده دارای نمونه های با تعداد کم و زیاد می باشد تا طبقه بندی کننده ارائه شده به خوبی ارزیابی شود.

جدول ۱- مشخصات مجموعه داده‌های استفاده شده در این مقاله

شماره	مجموعه داده	تعداد نمونه‌ها	تعداد ویژگی‌ها	داده گم شده	مقدار عددی
۱	vote	۴۳۵	۱۷	دارد	ندارد
۲	monk's	۴۳۷	۷	ندارد	دارد
۳	adult	۴۸۸۴۲	۱۴	دارد	دارد
۴	wine	۱۷۸	۱۳	ندارد	دارد
۵	breast-w	۶۹۹	۱۰	دارد	ندارد
۶	heart	۲۷۰	۱۴	ندارد	ندارد
۷	diabetes	۷۶۸	۱۳	ندارد	دارد

۵-۲- پیش پردازش داده ها

به منظور آماده سازی داده ها برای مدلسازی و اعتبارسنجی از ۳ عملیات پیش پردازش استفاده شده است:

- گسسته سازی: ویژگی های پیوسته با ابزار گسسته سازی وکا به ۱۰ بازه تقسیم بندی شده اند.
- مقادیر گم شده: داده های گم شده با ابزار مدیریت داده های گم شده وکا مدیریت شده اند.
- نمونه برداری: به منظور صرفه جویی در زمان، مجموعه داده هایی که بیش از ۵۰۰۰ نمونه دارند را با ابزار نمونه برداری نرم افزار وکا نمونه برداری کرده ایم. به طوری که از کل مجموعه داده ۲۰ درصد نمونه ها نمونه برداری شده است.

۵-۳- معیارهای ارزیابی

۵-۳-۱- معیارهای انتخاب مدل

در این مقاله از معیارهای لگاریتم راستی آزمایی، معیار اطلاعاتی بیزین یا شوآرز (BIC) و معیار اطلاعاتی آکایکه (AIC) برای عملیات انتخاب مدل استفاده شده است. در علم آمار، تابع راستی آزمایی تابعی است از پارامترهای مدل به شرط دانستن مجموعه‌ای از مشاهدات [۲۰]. فرض می‌کنیم $D_i<a_1,...,a_m>,C$ ویژگی‌ها و متغیر کلاس در مجموعه داده D باشند. لگاریتم راستی آزمایی بیز ساده به شکل رابطه (۱۰) تعریف می‌شود:

$$P(D_i)=P(D_i<a_1,...,a_m>,C)=\prod_{j=1}^{m}P(a_j|C)P(C) \quad (۱۸)$$

آنگاه، لگاریتم راستی آزمایی مدل پیشنهادی به شکل رابطه (۱۹) تعریف می-شود:

$$P(D_i) = P(D_i < a_1,...,a_m >, C) = \sum_{H} \prod_{i=1}^{n} P(a_i | C, H) P(C) P(H) \quad (۱۹)$$

معیار AIC [۲۱] معیاری است برای انتخاب مدل که کیفیت هر مدل را به شرط مجموعه ای از مشاهدات و پارامترها ارزیابی می کند. با فرض داشتن یک مدل آماری و مقدار تابع راستی آزمایی L و تعداد پارامترهای متناظر K، این معیار به شکل زیر تعریف می شود:

$$AIC = 2*L + 2*n \quad (۲۰)$$

معیار BIC [۲۲] سنجشی است از یک مدل بهینه از میان مجموعه‌ای از مدل‌های آماری. مدل با پایین ترین مقدار BIC به عنوان مدل برتر شناخته می شود. زمانی که در حال برازش یک مدل به مجموعه از داده‌ها هستیم، ممکن است با اضافه کردن پارامترها موجب پدیده بیش برازشی بشویم. BIC تعداد پارامترها را با یک تابع جریمه می کند و از بیش برازشی جلوگیری می کند که به شکل زیر تعریف می شود:

$$BIC = -2 * L + k * \ln(n) \quad (۲۱)$$

جایی که L مقدار تابع راستی آزمایی، n تعداد نمونه‌ها در مجموعه داده و k تعداد پارامترهای آزاد تخمین زده شده هستند.

۵-۳-۲- معیارهای ارزیابی طبقه‌بندی کننده‌ها

معیارهای بسیاری برای سنجش کارایی پیش بینی طبقه‌بندی کننده‌ها ارائه شده است. در این مقاله از صحت طبقه‌بندی، ناحیه زیر منحنی ROC و معیار F-measure به منظور ارزیابی روش پیشنهادی و توسعه‌های بیز ساده استفاده شده است. صحت یک طبقه‌بندی کننده به صورت زیر تعریف می شود:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (۲۱)$$

هر یک از پارامترهای صحت طبقه‌بندی به شکل زیر تعریف می شوند:

TP: تعداد نمونه‌هایی که به درستی مثبت تشخیص داده شده اند.

FP: تعداد نمونه‌هایی که به اشتباه مثبت تشخیص داده شده اند.

TN: تعداد نمونه‌هایی که به درستی منفی تشخیص داده شده اند.

FN: تعداد نمونه‌هایی که به اشتباه منفی تشخیص داده شده اند.

با فرض اینکه $f_1(s)$ تابع چگالی احتمال امتیازات s برای کلاس های $l \in \{0,1\}$ باشد، و $F_1(s)$ تابع توزیع تجمعی مرتبط باشد. آنگاه، معیار ناحیه زیر منحنی ROC نیز به شکل زیر تعریف می شود:

$$AUC = \int_{-\infty}^{+\infty} F_0(s) f_1(s) ds \quad (۲۲)$$

از معیار F-measure [۲۳] نیز در کنار صحت طبقه بندی و ناحیه زیر منحنی ROC نیز استفاده شده است. این معیار به شکل زیر تعریف می شود:

$$F - measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (۲۳)$$

این معیار میانگین هارمونیک دقت[۱۲] و بازخوانی[۱۳] می باشد. معیار دقت به شکل زیر تعریف می شود:

$$Precision = \frac{TP}{TP + FN} \quad (۲۴)$$

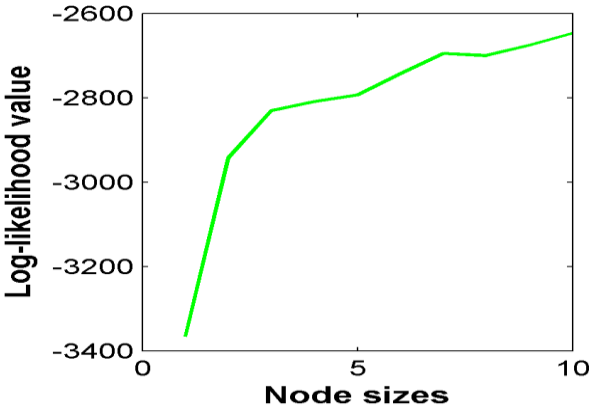
همینطور معیار بازخوانی نیز به شکل زیر تعریف می شود:

$$Recall = \frac{TP}{TP + FN} \quad (۲۵)$$

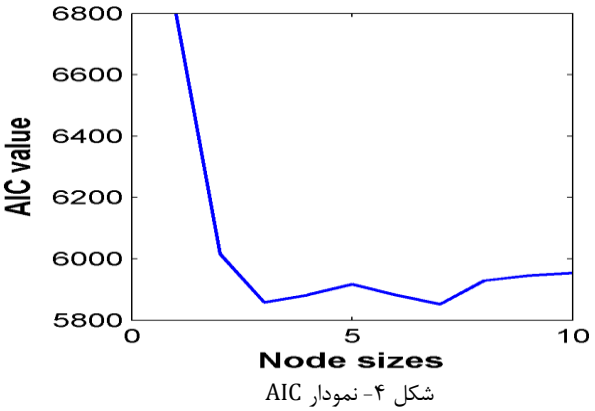
۶- نتایج و مباحثه

۶-۱- نتایج انتخاب مدل

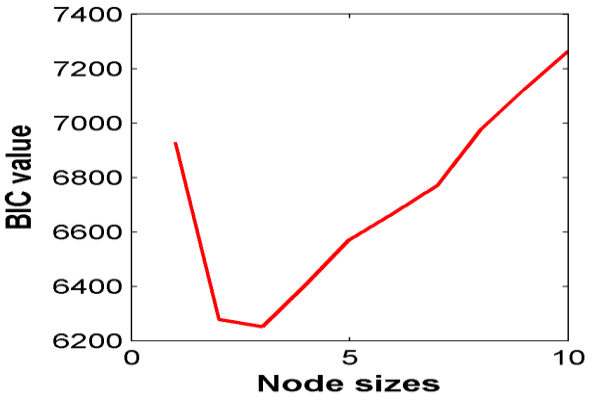
هدف عملیات انتخاب مدل جلوگیری از بیش برازشی مدل ارائه شده در این مقاله است. بنابراین از معیارهای BIC و AIC استفاده شده است تا از بیش برازشی جلوگیری شود. نتایج عملیات انتخاب مدل در شکل (۳) تا (۵) به تصویر کشیده شده است. توجه شود به منظور صرفه جویی از فضای نوشتاری مقاله، این عملیات تنها بر روی مجموعه داده vote اعمال شده است. عملیات انتخاب مدل در مجموعه داده‌های دیگر در این مقاله رفتار مشابهی دارند. همانطور که از شکل (۳) پیداست، مقدار لگاریتم راستی آزمایی با اضافه شدن مقدار به متغیر پنهان در حال همگرایی بوده و به صفر نزدیک می شود. اما با توجه به اینکه این معیار پارامترهای مدل را جریمه نمی کند، با افزایش اندازه متغیر پنهان پارامترهای مدل افزایش پیدا کرده و مدل مستعد پدیده بیش برازشی می شود. نمودار BIC نشان می دهد که مدل پیشنهادی بعد از اندازه نود ۳ تعداد پارامترهای آن در حال افزایش است. در انتخاب مدل باید توجه داشته باشیم که تنها مدلی را انتخاب کنیم که بین بازه اندازه نود ۲ تا ۳ است. این بازه یک محدوده ایمن برای طبقه‌بندی کننده پیشنهادی به شمار می رود. اما از آنجایی که معیار AIC رویکرد سخت گیرانه‌ای نسبت به BIC نیست، محدوده خطر بیش برازشی را بعد از اندازه نود ۸ نشان می دهد.



شکل ۳- نمودار تابع لگاریتم راستی آزمایی



شکل ۴- نمودار AIC



شکل ۵- نمودار BIC

۶-۲- نتایج کارایی طبقه‌بندی کننده‌ها

در این بخش نتایج کسب شده از آزمایشات به منظور ارزیابی کارایی پیش بینی طبقه بندی کننده پیشنهادی به همراه توسعه های بیز ساده فراهم شده است. این طبقه بندی کننده ها به وسیله نرم افزار وکا [۲۴] پیاده سازی شده اند. به منظور جلوگیری از توزیع نامتعادل برچسب متغیر کلاس در مجموعه داده ها از معیارهای F-measure و ناحیه زیر منحنی ROC [۲۵] نیز در کنار معیار صحت طبقه بندی استفاده شده است. همچنین به منظور جلوگیری از بایاس در پیش بینی ها، از روش اعتبار سنجی k فولد با مقدار k=10 استفاده شده است. تست دو دنباله ای تی با بازه اطمینان ۹۵ درصد به منظور مقایسه دو به دو طبقه بندی کننده ها استفاده شده است. طبقه‌بندی، ناحیه زیر منحنی ROC ، و F-measure برای هر طبقه‌بندی کننده را در هر مجموعه داده به همراه انحراف از معیار با علامت ± و میانگین امتیازات در پایین جدول نشان می دهند. درکنار میانگین نتایج، فرمت w=t=l نیز برای هر الگوریتم بر اساس مجموعه داده ها نشان داده شده است. تفسیر این فرمت به این شکل است که یک طبقه بندی کننده در مجموعه داده‌ها دارای w برد، t مساوی و l باخت است. توجه شود طبقه بندی پیشنهادی در این جداول به عنوان طبقه بندی کننده پایه در نظر گرفته شده است. جدول (۳)، جدول (۵)، و جدول (۷) خلاصه نتایج تست دو دنباله ای تی الگوریتم های مختلف در مجموعه داده ها را نشان می ده. هر ورودی در این جداول به شکل $i\,(j)$ نشان داده شده است. در این قالب، i تعداد مجموعه داده هایی است که الگوریتم در ستون مقدار صحت طبقه بندی، ناحیه زیر منحنی ROC، و F-measure بالاتری نسبت به الگوریتم در سطر متناظر خود کسب کرده است، و j تعداد مجموعه داده هایی است که الگوریتم در ستون نسبت به الگوریتم در سطر تعداد برد هایی بیشتری کسب کرده است. همانطور که از جداول پیداست، کارایی طبقبندی کننده پیشنهادی بصورت قابل توجه‌ای بالا بوده و توسعه های بیز ساده را شکست داده است. بر اساس صحت طبقه‌بندی (جدول ۲)، NB متحمل ۵ شکست و ۲ مساوی در مقایسه با طبقه بندی کننده پیشنهادی شده است. HNB در مقایسه با طبقه بندی کننده پیشنهادی، متحمل ۵ شکست شده است. این در حالی است که ۲ مساوی نیز کسب کرده است. طبقه بندی کننده AODE در مقابل، کمترین شکست را در مقایسه با طبقه بندی کننده پیشنهادی متحمل شده است.

برخی از مهمترین نتایج مقایسه‌ای طبقه‌بندی کننده‌ها بر اساس صحت طبقه بندی، ناحیه زیر منحنی ROC ، و معیار F-measure به این شرط می‌باشد:

- بر اساس معیار ناحیه زیر منحنی ROC ، در مقایسه با MLNB، TAN متحمل ۴ شکست شده است.
- بر اساس معیار F-measure، HNB متحمل ۲ شکست در مقایسه با MLNB شده است.
- بر اساس معیار F-measure، MLNB در مقایسه با AODE متحمل ۴ شکست شده است.
- بر اساس معیار صحت طبقه بندی، TAN متحمل ۵ شکست در مقایسه با طبقه MLNB شده است.
- بر اساس معیار صحت طبقه‌بندی، MLNB به طور معنا داری تمام توسعه‌های بیز ساده را شکست داده است.
- بر اساس معیار ناحیه زیر منحنی ROC، HNB در مقایسه با MLNB متحمل ۴ شکست شده است.

۶-۳- مباحثه

همانطور که در شکل ۲ نشان داده شده است، یک متغیر پنهان به ساختار شبکه بیز ساده اضافه شده است به گونه‌ای که در کنار متغیر کلاس والد تمام ویژگی‌ها می باشد. تفسیر وجود این متغیر پنهان به این شکل است که می گوییم ویژگی‌ها به شرط دانستن اطلاعات در مورد متغیر کلاس و پنهان از یکدیگر مستقل هستند. متغیر پنهان به عنوان یک فاکتور عمل می‌کند به گونه‌ای که وابستگی بین ویژگی‌ها را مدل‌سازی می کند. از آنجایی که متغیر پنهان مقادیر مختلفی به منظور افزایش لگاریتم راستی آزمایی اتخاذ می‌کند، می توانیم تعداد پارامترهای مدل را با توجه به پیچیدگی مجموعه داده و تعداد نمونه‌ها کنترل کنیم. مدل ارائه شده در مقایسه با TAN، از آنجایی که ساختار شبکه آن مشخص است نیازی به یادگیری ساختار با این پیچیدگی را ندارد. به علاوه، TAN از مشکل برابری راستی آزمایی [۸] رنج می برد. این در حالی است که MLNB فاقد این مشکل می باشد زیرا ساختار آن مشخص است و نیازی به جستجو برای بهترین لگاریتم راستی آزمایی را ندارد. در مقایسه با HNB، هر دو روش از متغیرهای پنهان به منظور معرفی ساختارهای پیچیده استفاده می کنند. در HNB، لایه‌ای از متغیرهای پنهان وجود دارد که در آن هر ویژگی یک والد پنهان دارد. اگرچه تنها یک متغیر پنهان در MLNB والد همه ویژگی‌ها در کنار متغیر کلاس است. MLNB از توزیع چند جمله‌ای به منظور مدل‌سازی استفاده می کند. این در حالی است که HNB هیچ فرضی روی مجموعه داده‌ها ندارد. MLNB قابلیت کنترل تعداد پارامترها را با تغییر اندازه متغیر پنهان دارد. در مواردی که تعداد ویژگی‌ها در یک مجموعه داده زیاد ولی تعداد نمونه‌ها اندک باشد، مدل ارائه شده قادر است با تنظیم پارامترها مدلی را انتخاب کند که از بیش برازشی جلوگیری می کند.

۷- نتیجه گیری و کارهای آتی

در این مقاله به منظور مدل‌سازی وابستگی بین ویژگی‌ها برای فرض استقلال شرطی مدلی ارائه شده که تعمیمی از طبقه‌بندی کننده بیز ساده می باشد. مدل ارائه شده یک ساختار تمام متصل است، به گونه‌ای که یک متغیر پنهان به ساختار شبکه بیز ساده اضافه می‌شود که هم سطح با متغیر کلاس می باشد. به منظور تخمین پارامترهای مدل از الگوریتم EM استفاده شد. نتایج از شبیه سازی‌ها نشان داد که روش پیشنهادی عملکرد قابل توجه ای در مقایسه با TAN، HNB، و AODEاز خود نشان داده است. همچنین نتایج عملیات انتخاب مدل نشان می دهد که روش پیشنهادی از بیش برازشی جلوگیری می کند. طبقه‌بندی کننده ارائه شده در کاربردهایی که یادگیری ساختار برای متخصصین دامنه هزینه بر است، بسیار پرکاربرد می باشد.

به عنوان مسیر تحقیقاتی آینده، هدف ما تلفیق متغیرهای پنهان بیشتر به ساختار این مدل است به طوری که وابستگی پیچیده‌تر از ویژگی‌ها را نمایش دهد. آزمایشات مبتنی بر مسئله طبقه‌بندی کننده دودویی هستند. به عنوان کارهای آتی دیگر، می توان طبقه‌بندی کننده ارائه شده را در مجموعه داده های بیشتری مورد ارزیابی قرار داد.

جدول ۲ – نتایج آزمایشات بر اساس صحت طبقه بندی

Dataset	MLNB	NB		HNB		TAN		AODE		
adult	۰.۸۳ ± ۰.۰۳	۰.۸۱ ± ۰.۰۳	*	۰.۸۳ ± ۰.۰۳		۰.۸۳ ± ۰.۰۳		۰.۸۳ ± ۰.۰۳		
breast-w	۰.۹۷ ± ۰.۰۲	۰.۹۷ ± ۰.۰۲		۰.۹۶ ± ۰.۰۲	*	۰.۹۵ ± ۰.۰۳	*	۰.۹۷ ± ۰.۰۲		
diabetes	۰.۸۳ ± ۰.۰۴	۰.۷۶ ± ۰.۰۵	*	۰.۷۵ ± ۰.۰۵	*	۰.۷۵ ± ۰.۰۵	*	۰.۷۶ ± ۰.۰۴	*	
heart	۰.۹۰ ± ۰.۰۵	۰.۸۴ ± ۰.۰۵	*	۰.۸۲ ± ۰.۰۶	*	۰.۸۰ ± ۰.۰۷	*	۰.۸۳ ± ۰.۰۶	*	
monk's	۰.۷۴ ± ۰.۰۵	۰.۶۳ ± ۰.۰۳	*	۰.۶۵ ± ۰.۰۵	*	۰.۶۰ ± ۰.۰۵	*	۰.۶۴ ± ۰.۰۴	*	
vote	۰.۹۵ ± ۰.۰۳	۰.۹۰ ± ۰.۰۴	*	۰.۹۴ ± ۰.۰۳	*	۰.۹۵ ± ۰.۰۳		۰.۹۵ ± ۰.۰۳		
wine	۰.۹۷ ± ۰.۰۴	۰.۹۶ ± ۰.۰۴		۰.۹۷ ± ۰.۰۴		۰.۹۲ ± ۰.۰۷	*	۰.۹۷ ± ۰.۰۴		
Average	۰.۸۹	۰.۸۴		۰.۸۴		۰.۸۳		۰.۸۵		
(v/ /*)		(۰/۲/۵)		(۰/۲/۵)		(۰/۲/۵)			(۰/۴/۳)	

جدول ۳ – نتایج آزمون دو دنباله ای تی با بازه اطمینان ۹۵ درصد بر اساس صحت طبقه بندی

a	b	c	d	e	([سطر] << [ستون] تعداد مجموعه داده هایی که در آن)
-	۴ (۳)	۲ (۲)	۵ (۴)	۷ (۵)	a = NB
۳ (۲)	-	۳ (۱)	۶ (۳)	۷ (۵)	b = HNB
۵ (۴)	۴ (۴)	-	۶ (۴)	۶ (۵)	c = TAN
۲ (۲)	۱ (۰)	۱ (۰)	-	۵ (۳)	d = AODE
۰ (۰)	۰ (۰)	۱ (۰)	۲ (۰)	-	e = MLNB

جدول ۴ - نتایج آزمایشات بر اساس ناحیه زیر منحنی ROC

Dataset	MLNB	NB		HNB		TAN		AODE	
adult	۰.۷۹ ± ۰.۰۴	۰.۸۸ ± ۰.۰۳	v	۰.۸۷ ± ۰.۰۴	v	۰.۸۷ ± ۰.۰۳	v	۰.۸۹ ± ۰.۰۳	v
breast-w	۰.۹۹ ± ۰.۰۱	۰.۹۹ ± ۰.۰۱		۰.۹۹ ± ۰.۰۱		۰.۹۹ ± ۰.۰۱	*	۰.۹۹ ± ۰.۰۱	
diabetes	۰.۸۴ ± ۰.۰۳	۰.۸۳ ± ۰.۰۵		۰.۸۲ ± ۰.۰۵	*	۰.۸۱ ± ۰.۰۵	*	۰.۸۳ ± ۰.۰۵	
heart	۰.۹۲ ± ۰.۰۵	۰.۹۱ ± ۰.۰۵		۰.۸۹ ± ۰.۰۵	*	۰.۹۰ ± ۰.۰۴	*	۰.۹۱ ± ۰.۰۵	
monk's	۰.۷۹ ± ۰.۰۶	۰.۵۷ ± ۰.۰۶	*	۰.۶۴ ± ۰.۰۷	*	۰.۶۲ ± ۰.۰۶	*	۰.۶۴ ± ۰.۰۶	*
vote	۰.۹۸ ± ۰.۰۲	۰.۹۷ ± ۰.۰۲	*	۰.۹۹ ± ۰.۰۱		۰.۹۹ ± ۰.۰۱		۰.۹۹ ± ۰.۰۱	
wine	۰.۹۸ ± ۰.۰۲	۱.۰۰ ± ۰.۰۰	v	۱.۰۰ ± ۰.۰۰	v	۱.۰۰ ± ۰.۰۱	v	۱.۰۰ ± ۰.۰۱	v
Average	۰.۹۰	۰.۰۸۸		۰.۸۸		۰.۸۸		۰.۸۹	
(v/ /*)		(۲/۳/۲)		(۲/۲/۳)		(۲/۱/۴)		(۲/۴/۱)	

جدول ۵- نتایج آزمون دو دنباله ای تی با بازه اطمینان ۹۵ درصد بر اساس ناحیه زیر منحنی ROC

a	b	c	d	e	([سطر] << [ستون] تعداد مجموعه داده هایی که در آن)
-	۲ (۲)	۲ (۲)	۵ (۳)	۵ (۲)	a = NB
۵ (۴)	-	۳ (۱)	۵ (۴)	۴ (۳)	b = HNB
۵ (۵)	۴ (۳)	-	۶ (۶)	۴ (۴)	c = TAN
۲ (۰)	۲ (۰)	۱ (۰)	-	۳ (۱)	d = AODE
۲ (۲)	۳ (۲)	۳ (۲)	۴ (۲)	-	e = MLNB

جدول ۶- نتایج آزمایشات بر اساس F-measure

Dataset	MLNB	NB	HNB	TAN	AODE
adult	۰.۷۶ ± ۰.۰۵	۰.۷۵ ± ۰.۰۴	۰.۸۹ ± ۰.۰۲ v	۰.۸۹ ± ۰.۰۲ v	۰.۸۹ ± ۰.۰۲ v
breast-w	۰.۹۶ ± ۰.۰۲	۰.۹۷ ± ۰.۰۲ v	۰.۹۷ ± ۰.۰۲	۰.۹۶ ± ۰.۰۲	۰.۹۸ ± ۰.۰۱ v
diabetes	۰.۶۴ ± ۰.۰۵	۰.۶۸ ± ۰.۰۴ v	۰.۶۱ ± ۰.۰۷ *	۰.۶۴ ± ۰.۰۸	۰.۶۴ ± ۰.۰۷
heart	۰.۷۹ ± ۰.۰۷	۰.۸۱ ± ۰.۰۶	۰.۷۹ ± ۰.۰۷	۰.۷۶ ± ۰.۰۹	۰.۸۰ ± ۰.۰۷
monk's	۰.۵۴ ± ۰.۰۷	۰.۴۳ ± ۰.۰۴ *	۰.۷۵ ± ۰.۰۴ v	۰.۷۲ ± ۰.۰۴ v	۰.۷۶ ± ۰.۰۳ v
vote	۰.۹۵ ± ۰.۰۴	۰.۹۰ ± ۰.۰۴ *	۰.۹۳ ± ۰.۰۴ *	۰.۹۳ ± ۰.۰۴	۰.۹۳ ± ۰.۰۴ *
wine	۰.۹۷ ± ۰.۰۵	۰.۹۷ ± ۰.۰۵	۰.۹۹ ± ۰.۰۴ v	۰.۹۵ ± ۰.۰۷ *	۰.۹۹ ± ۰.۰۳ v
Average	۰.۸۰	۰.۷۹	۰.۸۵	۰.۸۴	۰.۸۶
(v/ /*)		(۲/۳/۲)	(۳/۲/۲)	(۲/۴/۱)	(۴/۲/۱)

جدول ۷- نتایج آزمون دو دنباله ای تی با بازه اطمینان ۹۵ درصد بر اساس F-measure

a	b	c	d	e	([سطر] << [ستون] تعداد مجموعه داده هایی که در آن)
-	۴ (۴)	۳ (۳)	۵ (۴)	۴ (۲)	a = NB
۳ (۱)	-	۳ (۲)	۶ (۳)	۳ (۲)	b = HNB
۴ (۳)	۴ (۴)	-	۵ (۴)	۵ (۱)	c = TAN
۲ (۱)	۰ (۰)	۲ (۰)	-	۱ (۱)	d = AODE
۳ (۲)	۴ (۳)	۲ (۲)	۶ (۴)	-	e = MLNB

مراجع

[1] R. T. Asmono, R. S. Wahono, and A. Syukur, "Absolute Correlation Weighted Naïve Bayes for Software Defect Prediction," *Journal of Software Engineering,* vol.1, no. 1, pp. 38-45, 2015.

[2] J. H. Friedman, "On bias, variance, 0/1—loss, and the curse-of-dimensionality," *Data mining and knowledge discovery,* vol. 1, no. 1, pp. 55-77, 1997.

[3] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning,* vol. 29, no. (2-3), pp. 131-163, 1997.

[4] G. I. Webb, J. R. Boughton, and Z. Wang, "Not so naive Bayes: aggregating one-dependence estimators," *Machine learning*, vol. 58, no. 1, pp. 5-24, 2005.

[5] L. Jiang and H. Zhang, "Weightily averaged one-dependence estimators," *in Pacific Rim International Conference on Artificial Intelligence*, pp. 970-974,2006.

[6] G. Abaei and A. Selamat, "A survey on software fault detection based on different prediction approaches," *Vietnam Journal of Computer Science*, vol. 1, no. 2, pp. 79-95, 2014.

[7] K. Dejaeger, T. Verbraken, and B. Baesens, "Toward comprehensible software fault prediction models using bayesian network classifiers," *IEEE Transactions on Software Engineering*, vol. 39, no. 2, pp. 237-257, 2013.

[8] C. P. de Campos, G. Corani, M. Scanagatta, M. Cuccu, and M. Zaffalon, "Learning extended tree augmented naive structures," *International Journal of Approximate Reasoning*, vol. 68, pp. 153-163, 2016.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society*. Series B (methodological), pp. 1-38, 1977.

[10] K. P. Murphy, "*Naive bayes classifiers*," University of British Columbia, 2006.

[11] L. Jiang, H. Zhang, and Z. Cai, "A novel Bayes model: Hidden naive Bayes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 10, pp. 1361-1371, 2009.

[12] L. Jiang, H. Zhang, Z. Cai, and J. Su, "Evolutional naive bayes," *in Proceedings of the International Symposium on Intelligent Computation and its Application (ISICA)*, pp. 344-350, 2005.

[13] B. Turhan and A. B. Bener, "Software Defect Prediction: Heuristics for Weighted Naïve Bayes," *in ICSOFT (SE)*, pp. 244-249, 2007.

[14] S. Taheri, J. Yearwood, M. Mammadov, and S. Seifollahi, "Attribute weighted Naive Bayes classifier using a local optimization," *Neural Computing and Applications*, vol. 24, no. 5, pp. 995-1002, 2014.

[15] H. Zhang and S. Sheng, "Learning weighted naive Bayes with accurate ranking," *Fourth IEEE International Conference on Data Mining (ICDM'04),* pp. 567-570, 2004.

[16] M. Hall, "A decision tree-based attribute weighting filter for naive Bayes," *Knowledge-Based Systems*, vol. 20, no. 2, pp. 120-126, 2007.

[17] E. Frank, M. Hall, and B. Pfahringer, "Locally weighted naive bayes," *in Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pp. 249-256, 2002.

[18] F. V. Jensen, *An introduction to Bayesian networks*. UCL press London, 1996.

[19] K. Bache and M. Lichman, "*UCI machine learning repository*," 2013.

[20] C. M. Bishop, *Pattern recognition and Machine Learning*, springer, 2006.

[21] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *in Selected Papers of Hirotugu Akaike*, Springer, pp. 199-213, 1998.

[22] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461-464, 1978.

[23] C. J. V. Rijsbergen, *Information retrieval*, Dept. of Computer Science, University of Glasgow, URL: citeseer.ist.psu.edu/vanrijsbergen79information.html, 1979.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and L. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10-18, 2009.

[25] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine learning*, vol. 45, no. 2, pp. 171-186, 2001.

نیما شیری هرزویلی مدرک کارشناسی مهندسی کامپیوتر گرایش نرم افزار را از دانشگاه علمی کاربردی جهاد دانشگاهی بندر انزلی و مدرک کارشناسی ارشد خود را در رشته مهندسی کامپیوتر گرایش نرم افزار از دانشگاه آزاد اسلامی قزوین اخذ کرده است. زمینه های تحقیقاتی مورد علاقه ایشان یادگیری ماشین، شبکه‌های بیزین، داده کاوی و پیش بینی خطای نرم افزار می باشد. آدرس پست الکترونیکی ایشان عبارت است از:

nimashiri@qiau.ac.ir

ساسان حسینعلی زاده مدرک کارشناسی مهندسی کامپیوتر گرایش سخت افزار را از دانشگاه شیراز و مدرک کارشناسی ارشد خود را در رشته علوم کامپیوتر گرایش سیستم‌های هوشمند از دانشگاه صنعتی امیرکبیر اخذ کرده است. او همچنین مدرک دکتری تخصصی علوم کامپیوتر را از دانشگاه صنعتی امیرکبیر اخذ کرده است و هم اکنون استادیار پژوهشکده فناوری اطلاعات در پژوهشگاه ارتباطات و فناوری اطلاعات می باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان یادگیری ماشین، فرآیندهای تصادفی، شبکه‌های بیزین، شبکه-های پیچیده پویا، سیستم های توصیه‌گر، داده کاوی و مهندسی نرم افزار می باشد. آدرس پست الکترونیکی ایشان عبارت است از:

s.alizadeh@itrc.ac.ir

---

[1]Naïve Bayes
[2]Tree-Augmented Naïve Bayes
[3]Expectation-Maximization
[4]Hidden Naïve Bayes
[5]Aggregating One-dependence Estimators
[6]Evolutional Naïve Bayes
[7]Weighted Naïve Bayes
[8]K-nearest neighbor
[9]Locally Weighted Naïve Bayes
[10]Mixture of Latent Naïve Bayes
[11]Diverging
[12]Precision
[13]Recall

The CSI Journal on Computing Science and Information Technology
Vol. 17, No.1, 2019

Extending Naïve Bayes classifier structure to model conditional mutual dependency among attributes

Nima Shiri Harzevili[1], Sasan H.Alizadeh[2]

[1]Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

[2] Faculty of Information Technology, ICT Research Institute (Iran Telecommunication Research Center), Tehran, Iran

Abstract

Naïve Bayes classifier has been attracted by many researchers due to its simple structure and noticeable classification performance. It isbased on the conditional independence assumption of attributes given the class variable. However, this assumption may not hold to be true in real-world applications of naïve Bayes classifier. Therefore, the concept of latent variables is employed in this paper to propose Mixture of Latent Naïve Bayes (MLNB) classifier to model the conditional mutual dependency among attributes. We have slightly modified the Expectation-Maximization (EM) algorithm to estimate the parameters of the model. Experiments on 7 datasets obtained from the University of California, Irvine (UCI) machine learning repository indicate that MLNB exhibits a significant predictive performance compared to the state-of-the-art extensions of naïve Bayes classifier in terms of classification accuracy, under the ROC curve, and F-measure.

Keywords: Naïve Bayes classifier, conditional independence, latent variables, Bayesian networks.