بسمه تعالی

"بهبود دقت پیش بینی الگوریتم C5.0 با استفاده از متد Boosting مطالعه موردی دانشگاه پیام نور استان قم"

نویسندگان: محمد تاری

دانشگاه پیام نور، دانشجوی کارشناسی ارشد گروه علمی مهندسی کامپیوتر، تهران،۳۶۹۷-۱۹۳۹۵، ج.ا.ایران
M_taheri16@yahoo.com

چکیده

یکی از چالشهای جدی در مدیریت امور آموزشی دانشگاهها پیش‌بینی وضعیت تحصیلی دانشجویان در نیمسالهای آینده به منظور شناسایی دانشجویانی است که دچار افت تحصیلی شده و ادامه تحصیل آنها با مشکل روبرو خواهد شد .در این تحقیق با استفاده از تکنیکهای داده‌کاوی وضعیت تحصیلی آتی دانشجویان شامل معدل نیمسال آینده، معدل کل در زمان فارغ التحصیلی، و وضعیت فارغ التحصیلی پیش‌بینی شده است .برای ساخت مدل‌های مورد نظر از الگوریتم ***C5.0*** استفاده شده است ، برای افزایش دقت پیش بینی از متد ***Boostig*** کمک گرفتیم، برای ارزیابی نتایج مدل از ***Confussion Matrix*** ، ***Precision*** ، ***Recall*** استفاده نمودیم.

لگوریتم ***C5.0*** از ***boosting*** در درختان تصمیم گیری پشتیبانی می‌کند . ***Boosting*** یک تکنیک برای تولید و ترکیب کلاسه بندی کننده‌های چندگانه بمنظور بهبود نرخ دقت است .

بانک اطلاعاتی مورد استفاده برای پیاده سازی الگوریتم و ساخت مدل ها وضعیت تحصیلی دانشجویان دانشگاه پیام نور استان قم از سال ۱۳۷۹ تا ۱۳۹۰ می باشد. مراحل انجام تحقیق با استفاده از متدولوژی ***CRISP-DM*** انجام گردیده و ساخت مدل ها با استفاده از نرم افزار ***Clementine 12.0*** الگوریتم های پیاده سازی شده اند.

واژه‌های کلیدی: Boosting ، CRISP-DM،C5.0 ، داده کاوی ، دانشگاه پیام نور استان قم

۱- مقدمه

این یکی از چالشهای جدی در مدیریت امور آموزشی دانشگاهها پیش‌بینی وضعیت تحصیلی دانشجویان در نیمسالهای آینده به منظور شناسایی دانشجویانی است که دچار افت تحصیلی شده و ادامه تحصیل آنها با مشکل روبرو خواهد شد .در این تحقیق با استفاده از تکنیکهای داده‌کاوی وضعیت تحصیلی آتی دانشجویان شامل معدل نیمسال آینده، معدل کل در زمان فارغ التحصیلی، و وضعیت فارغ التحصیلی پیش‌بینی شده است.

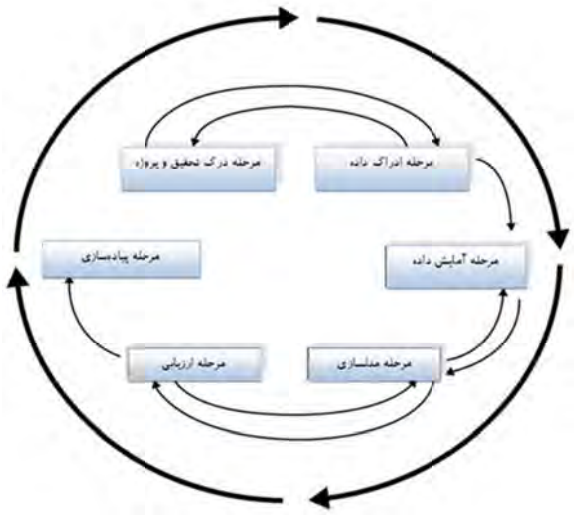
برای ساخت مدل‌های مورد نظر از تکنیک‌های مختلفی

نظیر شبکه‌های عصبی، درخت‌های تصمیم استفاده شده است .این مدلها برای داده‌های سیستم آموزشی دانشگاه پیام نور استان قم پیاده‌سازی شده اند .عملکرد هر یک از مدلها، مورد بررسی قرار گرفته و نتایج به دست آمده با یکدیگر مقایسه گردیده اند .اعتبار سنجی انجام شده بر روی مدلها اثبات می‌کند که نتایج به دست آمده دقیق و قابل اعتماد بوده اند.

با بکارگیری این مدلها، مدیران آموزشی می‌توانند مشاوره‌های لازم را برای پیشگیری از رسیدن دانشجویان به وضعیت بحرانی بکار گیرند .همچنین این مدلها می‌توانند به عنوان یک ابزار پشتیبان تصمیم گیری در سیستم‌های آموزشی مورد بهره برداری قرار گرفته و نقش مهمی را در ارتقاء سطح علمی دانشگاهها داشته باشند.

۲- متدولوژی CRISP-DM

در تکنیکهای داده‌کاوی از جمله تکنیکهای نوین علمی هستند که در توصیف، تشریح، پیش‌بینی و کنترل پدیده‌ها به کار می‌روند. این تکنیکها به اندازه‌گیری، تشریح و پیش‌بینی درجه وابستگی میان متغیرها می‌پردازند. روشهای داده‌کاوی نه تنها بر جنبه‌های تحلیلی مطالعات، بلکه در طراحی و ابزارهای جمع آوری داده برای تصمیم‌گیری و حل مسائل نیز تأثیر می‌گذارند. موفق‌ترین پروژههای داده‌کاوی، در چارچوب فرآیند استانداردی اجرا می‌شود که توسط یک تیم کاری در شرکت SPSS در قالب پروژه‌های به نام CRISP-DM ارائه شده است. برطبق CRISP-DM یک پروژه داده‌کاوی معین شامل چرخه حیاط شش مرحله‌ای است که توالی مراحل را نشان می‌دهد. هر مرحله از ترتیب مراحل اغلب نتیجه وابستگی مراحل قبلی را نیز دربر دارد. مهمترین وابستگی بین مراحل نمایش پیکانها است. خاصیت تکراری CRISP-DM حاکی از چرخه بیرونی است که اغلب منجر به راه حلی برای مسئله تحقیقی یا تجاری با سوالات اضافی جالب توجه می‌شود. در زیر مراحل کاری در داده‌کاوی را توضیح می‌دهیم(N.balac 2006)



شکل ۱: مراحل و وظایف متدولوژی CRISP-DM(غضنفری و دیگران ۱۳۸۷)

۲-۱- فازهای متدولوژی CRISP-DM

درک فضای کسب و کار،شناسایی داده ها،آماده سازی داده ها ، مدل سازی، پایش و ارزیابی ،اجرا

متدولوژی دارای ۶ فاز می باشد که یکی از مهمترین فازهای آن فاز آماده سازی داده ها می باشد.

پیشپردازش اهمیت آماده‌سازی داده‌ها به دلیل این واقعیت است که؛ "فقدان داده با کیفیت برابر با فقدان کیفیت در نتایج کاوش است" (ندجیم ۲۰۰۷) ورودی نادرست خروجی نادرست به دنبال دارد در جدول 4-۲ مقایسه‌ای بین اهمیت آماده‌سازی داده‌ها نسبت به سایر گام‌های کشف دانش به کمک داده‌کاوی صورت گرفته است. با این حال، متأسفانه بسیاری اهمیت آماده‌سازی داده‌ها را فراموش کرده

و یا آن را کم اهمیت می‌انگارند. از این رو تلاش‌های بسیاری برای بسط و توسعه آماده‌سازی داده‌ها در داده‌کاوی روی داده است. وظیفه اصلی پیش پردازش داده‌ها؛ سازمان دهی داده‌ها در شکل‌های استاندارد برای داده‌کاوی و یا سایر عملیات مبتنی بر کامپیوتر است؛ که در ادامه مو رد اشاره قرار گرفته است. (St.Kliment Ohridski ۲۰۰۷)

جدول ۱: مقایسه اهمیت گام آماده‌سازی داده‌ها با سایر گام‌های داده‌کاوی

گام داده‌کاوی	درصد زمان صرف شده از کل کار	درصد اهمیت در موفقیت نهایی کار
آماده‌سازی داده	۷۵	۷۵
بررسی داده	۲۰	۱۵
مدل‌سازی داده	۵	۱۰

۳- نرم افزار Clementine 12.0

ابزار داده‌کاوی مورد استفاده برای اعمال الگوریتم‌ها بر روی بانک اطلاعاتی دانشجویان نرم افزار Spss Clementine 12.0 است. تفاوت این نرم افزار با نرم افزارهای دیگر در پردازش داده از طریق به کارگیری تعدادی گره [1] است که در قالب یک رویه با یکدیگر ارتباط دارند. علاوه بر موارد فوق امکان بصری‌سازی [2] نتایج الگوریتم‌ها با استفاده از روشهای متعدد دارا است، گره‌های موجود در این نرم افزار به ۶ گروه اصلی تقسیم می‌شود که عبارتند از :

منابع [3]، عملیات بر روی رکورد[4] ، عملیات بر روی فیلد[5] ، گرافها [6]، مدل‌ها [7]،خروجی [8]

۴- الگوریتم C5.0

یکی از الگوریتم‌های درختان تصمیم گیری است که در تحقیقمان بمنظور کشف دانش و قوانین با کیفیت تر مورد استفاده قرار گرفت . الگوریتم c5.0 یک نوع درخت تصمیم گیری تک متغیره و بهبود یافته الگوریتم c4.5 است که توسط محقق استرالیایی J.ROSS Quinlan در سال ۱۹۹۳ طراحی شد . این الگوریتم مشابه با الگوریتم CART، ابتدا درختی تقریبا پر ایجاد می‌کند . ولی استراتژی هرس آن کاملا متفاوت است . این الگوریتم، کلاسه بندی را با تقسیم داده‌ها به زیر مجموعه‌هایی که شامل رکوردهای همکن تر از والد خود هستند، انجام می‌دهد . در C5.0 تقسیم کردن نمونه‌ها بر اساس فیلدی که بیشترین بهره اطلاعات را دارد، صورت می‌گیرد.هر زیر نمونه توسط اولین انشعاب تعیین می‌شود.سپس معمولا بر اساس فیلدی دیگر مجددا تقسیم بندی انجام می‌گیرد، و این فرایند به دفعات تکرار می‌شود تا اینکه زیر نمونه‌ها قابلیت تقسیم شدن را نداشته باشند. سرانجام، انشعاب‌های پایین ترین سطح از نو آزموده می‌شوند، و آن انشعاب‌های که ارزش چشمگیری ندارند از مدل حذف می‌شوند . لازم به ذکر است که C5.0 تنها فیلد خروجی از نوع categorical را می‌پذیرد، اما

---

[1] Node
[2] *Visualization*
[3] Resources
[4] Record Ops
[5] Field Ops
[6] Graphs
[7] Models
[8] Output

فیلدهای دیگر می‌توانند از هر نوعی باشند.

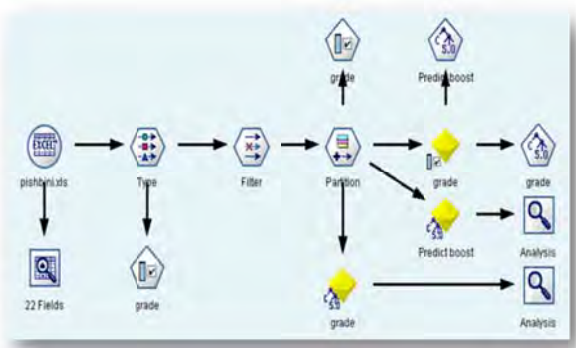
استنتاج قانون از طریق الگوریتم C5.0 بر اساس درخت تصمیم گیری است، این احتمال وجود دارد که بیش از یک قانون به ازای هر رکورد خاص صدق کند و یا هیچ قانونی به کار نرود . اگر چندین قانون برای یک رکورد مناسب باشد، هر قانون مبتنی بر اطمینان مربوط به آن قانون، وزنی تحت عنوان "Vote " می‌گیرد . در اینصورت بر اساس ترکیب وزن همه قوانین مناسب برای رکورد، پیش‌بینی نهایی تعیین می‌شود، و اگر هیچ قانونی مناسب نباشد، یک پیشگویی پیش فرض به آن رکورد نسبت داده می‌شود.(Bakır, B; Batmaz 2006)

۴-۱ Boosting

الگوریتم C5.0 از boosting در درختان تصمیم گیری پشتیبانی می‌کند . Boosting یک تکنیک برای تولید و ترکیب کلاسه بندی کننده‌های چندگانه بمنظور بهبود نرخ دقت است . Boostingیا ساختار چندین مدله در یک توالی اجرا می‌شود . اولین مدل معمولا از این روش ساخته شده و سپس، دومین مدل با تمرکز بر رکوردهایی که در مدل اولیه کلاسه بندی شدند ساخته می‌شود . سپس مدل سوم، بر اساس خطاهای مدل دوم ایجاد می‌شود و این روند ادامه می‌یابد . در نهایت آن نمونه‌های کلاسه بندی شده از طریق هر یک از این مدل ها، توسط زیر روال رای گیری وزن دار، پیشگویی‌های مجزا را تحت یک پیشگویی‌های مجزا را تحت یک پیشگویی واحد ترکیب می‌کند . Boosting بطور چشمگیری دقت یک مدل C5.0 را بهبود می‌بخشد . اما در اینصورت زمان آموزش طولانی تر می‌شود . علاوه بر این، غربال‌سازی خودکار Attributeها در C5.0 نهایتا منجر به کلاسه بندی‌های کوچکتر و با دقت پیشگویی بالاتر می‌شود.

۵- ایجاد مدل C5.0

برای ایجاد مدل با الگوریتم C5.0، مجموعه داده‌ها بطور تصادفی به دو بخش آموزش و تست، با تناسبی به ترتیب معادل با 80٪ , 20% تقسیم شده اند . در طی مرحله آموزش به جهت بالا بردن دقت مدل از متد boosting با trials معادل ۱۰ استفاده نمودیم . بعد از ساخت درخت، رویه هرس سراسری بمنظور جلوگیری از برازش بیش از حد با شدت هرس ۸۰٪ اجرا شده است . حداقل تعداد برگ‌ها در هر گره برگ با قدار ۲ تنظیم شده و همچنین از گزینه symbolic group بمنظور ترکیب کردن مقادیر symbolic ایی که الگوهای مشابهی را در ارتباط با فیلد



شکل ۱: مدل ایجاد شده توسط الگوریتم C5.0 درحالت عادی و Boosted

خروجی ارائه می‌دهند، استفاده شده است . این مراحل بر روی هر دو مجموعه اطلاعات اجرا شد.

پس از اجرای مدل‌ها خروجی آنها را مورد ارزیابی قرار دادیم که نتایج زیر به دست آمد:

جدول ۲ :مقایسه ارزیابی مدل‌ها درحالت عادی و Boosted

ردیف	داده آموزشی	داده تست

ارزیـــــابی قبـــــل از Boosting	88.64%	88.7%
ارزیـــــابی بعـــــد از Boosting	91.66%	91.54%
مقدار بهود ارزیابی	٪۳.۰۲	٪۲.۸۴

جدول ۳ :مقایسه ارزیابی مدل‌ها درحالت عادی و Boosted

ردیف	داده آموزشی		داده تست	
تعداد درست قبـــــل از Boosting	صحیح	نادرست	صحیح	نادرست
	۳۸۶۳	۴۹۲	۱۱۳۱	۱۴۵
تعـــــداد نادرست بعد از Boosting	صحیح	نادرست	صحیح	نادرست
	۳۹۹۲	۳۶۳	۱۱۶۸	۱۰۸

همانگونه که در جداول ۲ و ۳ مشاهده می‌کنید میزان دقت ارزیابی و موارد صحیح پیش بینی شده به صورت قابل توجهی افزایش یافته است. به طور مثال بعد از اجرای Boosting تعداد موارد صحیح پیش بینی در داده‌های آموزشی داشته ایم.

۶- Confussion Matrix

اما مساله ما در اینجا دارای ۳ کلاس A، B و C می‌باشد در این حالت Confution Matrix یک ماتریس ۳*۳ می‌شود که دیگر به تنهایی نشان دهنده همان مقادیر بیان شده در ماتریس۲*۲ نیستند بلکه برای به دست آوردن آن مقادیر باید بر روی مقادیر حاضر یک سری عملیات جبری صورت داد.

در یک بررسی کلی جمع کلیه عناصری که بر روی قطر اصلی قرار دارند نشان دهنده مقدار Accuracy و جمع کلیه عناصری که بر روی قطر اصلی قرار ندارند نشان دهنده مقدار Error می‌باشد.

به عنوان نمونه ما در اینجا با استفاده از نرم افزار کلمنتاین ابتدا Confusion Matrix مربوط به بهترین الگوریتم پیش بینی اجراشده را که C5.0 می‌باشد، به دست می‌آوریم. سپس مقادیر مربوط به Accuracy و Error را محاسبه خواهیم کرد.

جدول ۴ : نمونه یک Confussion Matrix

کلاس	A	B	C
A	82	1633	991
B	21	4547	380
C	8	413	2898

بر اساس جدول 6-۱۶ آمده است حاصل جمع عناصر قطر اصلی تقسیم بر کل مقادیر نمونه مقدار Accuracy را به دست می‌دهد:

$$Accuracy = \frac{82 + 4547 + 2898}{82 + 1633 + 991 + 21 + 4547 + 380 + 8 + 413 + 2898} = \frac{7527}{10973} = 68/59\% \quad (۱)$$

که دقیقا برابر است با مقداری که در جدول ۱۰-۵ برای درصد درستی پیش بینی الگوریتم C5.0 به دست آمده است. اما با توجه به اینکه مقدار Error برابر است با 1- Accuracy پس داریم:

Error = 100-68/59=31/41

مقدار Error از روی جدول به صورت ذیل به دست می‌آید:

که این مقادیر مربوط به کلاس A هستند به همین ترتیب مقادی کلاس‌های B و C به دست می‌آیند.

$$Error = \frac{1633 + 991 + 21 + 380 + 8 + 413}{82 + 1633 + 991 + 21 + 4547 + 380 + 8 + 413 + 2898} = \frac{3446}{10973} = 31/41\%$$

(۲)

اما مقادیر Recall و Precision چگونه به دست می‌آیند؟

اگر Confusion Matrix را به صورت ذیل در نظر بگیریم:

Confusion Matrix با ابعاد ۳*۳

		بینی مقادیر پیش		
		A	B	C
مقادیر واقعی	A	tpA	eAB	eAC
	B	eBA	tpB	eBC
	C	eCA	eCB	tpC

آنگاه مقادیر Recall و Precision با فرمول‌های زیر به دست می‌آیند:

$$Precision_A = tp_A/(tp_A + e_{BA} + e_{CA})$$

$$Recall_A = tp_A/(tp_A + e_{AB} + e_{AC})$$

۱. ارزیابی مدل C5.0

جدول ۵ :نتایج دقت ارزیابی مدل C.50 قبل از اعمال الگوریتم Boosting

قسمت	آموزشی	درصد	تست	درصد
صحیح	3863	88.7%	1131	٪88.64
نادرست	492	٪11.3	145	11.36%
کل	4355		1276	

در جدول ۵ نتایج دقت ارزیابی مدل C5.0 محاسبه گردیده که تعداد ۳۸۶۳ رکورد برای آموزش انتخاب شده که ۸۸.۷٪ به صورت درست پیش بینی شده است . و تعداد ۴۹۲ رکورد به صورت نادرست پیش بینی شده است که ۱۱.۳٪ می باشد. تعداد کل داده‌های آموزشی ۴۳۵۵ رکورد می‌باشد.

۱۲۷۶ رکورد برای تست ارزیابی مدل انتخاب شد که میزان صحیح رکوردهای پیش بینی شده ۱۱۳۱ رکورد که ۸۸.۶۴٪ می باشد و تعداد ۱۴۵ رکورد به صورت نادرست پیش بینی شدند که ۱۱.۳۶٪ می باشد. تمام موارد بالا قبل از اعمال الگوریتم Boosting بر روی مدل C5.0 بود که در ادامه ارزیابی مدل C5.0 بعد از اعمال الگوریتم Boosting را مورد بررسی قرار می دهیم.

عنوان مقاله: " بهبود دقت پیش بینی الگوریتم C5.0 با استفاده از متد Boosting مطالعه موردی دانشگاه پیام نور استان قم

نویسندگان: محمد تاری

جدول ۶ :نتایج دقت ارزیابی مدل C.50 بعد از اعمال الگوریتم Boosting

قسمت	آموزشی	درصد	تست	درصد
صحیح	۳۹۹۲	91.66%	1168	٪91.54
نادرست	۳۶۳	٪11.3	108	8.46%
کل	4355		1276	

همانگونه که در جداول ارزیابی مشاهده می‌کنیم بعد از اعمال الگوریتم Boosting ٪۲.۹۶ افزایش دقت پیش بینی در داده‌های آموزشی و 2.9% افزایش دقت پیش بینی در داده‌های تست داشته ایم ، یعنی حدود ۴۳۵۵ و ۱۲۷۶ رکورد برای آموزش و تست انتخاب شده اند که از ۴۳۵۵ رکورد آموزشی ۳۹۹۲ رکورد به صورت درست پیش بینی شده اند و ۳۶۳ رکورد نادرست پیش بینی شده است.

و بعد از آموزش داده‌ها از ۱۲۷۶ رکورد ۱۱۶۸ به صورت درست پیش بینی و مقدار ۱۰۸ رکورد به صورت نادرست پیش بینی شده است..

۷- نتیجه‌گیری

با توجه به اینکه بیشتر نرم افزار ها مبتنی بر بانک اطلاعاتی بر اساس دستورات ساده پایگاه داده عمل می‌کنند و هیچ گونه پویایی در این نرم افزار ها موجود نمی باشد و بیشتر انتخاب‌ها بر اساس آزمون و خطا می باشد و هیچ گونه وابستگی به داده های قبلی ندارد ، ولی با ایجاد یک موتور داده کاوی داخل این نرم افزارها می‌توانیم سیستم را پویا نموده و در هنگام برنامه ریزی مشاوره و راهنمایی ها لازم به کارشناسان سیستم ارائه گردد که موجب کاهش خطا و افزایش بهره وری سیستم گردد. پس با استفاده از مدل های ایجاد شده در این تحقیق می توان موتور داده کاوی را به سیستم های سنتی اضافه نمود و به صورت پویا پیش بینی های لازم را بر روی داده ها اعمال نمود.

سپاسگزاری

از استاد ارجمند جناب آقای دکتر بهروز مینایی به پاس کوشش‌ها و رهنمودهای بی شائبه شان در طی انجام تحقیقات تشکر و قدردانی می‌نمایم .

مراجع

[1] **J. Luan, PhD** Data Mining Applications In Higher Education 2006
[2] N.Balac Introduction To Data Mining , Elsevier Science 2006
[3] **Han , J. and Kamber , M.** Data Mining : Concepts and Techniques, Second Edition , Morgan Kaufman Publisher , 2006
[4] **michael steinbach vipin kumar** Introduction To DataMining pang-ning , 2006
[5] Spatial data mining implementation Alternatives and performances Nadjim
[۶] مهدی غضنفری ، سمیه علیزاده ، بابک تیمورپور، .داده کاوی و کشف دانش انتشارات دانشگاه علم و صنعت ایران ۱۳۸۷
[۷] اعظم ایرجی،بهروز مینایی ، ونوس شکورنیاز، استخراج قوانین تصمیم با استفاده از الگوریتم درخت تصمیم جهت هدایت تحصیلی دانش آموزان به کمک دسته بندی داده ها ی آموزش و پرورش دومین کنفرانس داده کاوی ایران، ۱۳۸۷