پیش بینی وضعیت تحصیلی دانشجویان
موسسه آموزش عالی صفاهان با استفاده از داده کاوی به کمک نرم افزار RapidMiner

مریم وفا[۱] و نجمه محمدی اسفرجانی[۲]
[۱] عضو هیات علمی موسسه آموزش عالی صفاهان، گروه کامپیوتر، mis_vafa@yahoo.com
[۲] دانشجوی مهندسی کامپیوتر- نرم افزار، موسسه آموزش عالی صفاهان،n_mohammadi5581@yahoo.com

چکیده_داده کاوی آموزشی یکی از کاربردهای مهم داده کاوی است. در این مقاله به پیش بینی وضعیت تحصیلی دانشجویان موسسه آموزش عالی صفاهان به کمک نرم افزار *Rapid miner* پرداخته شده است. ابتدا مراحل موردنیاز آماده سازی روی داده های جمع آوری شده انجام شده است. سپس جهت ساخت مدل از روش دسته بندی مبتنی بر الگوریتم درخت تصمیم استفاده شده است. پس از ساخت مدل به روش *k-fold cross validation* به ارزیابی مدل حاصل پرداختیم که دقت ۷۱.۳۳٪ به دست آمد. مهمترین نتیجه حاصل از این پروژه آن است که معدل ترم دوم دانشجویان نقش مهمی در پیش بینی وضعیت تحصیلی دانشجویان و تشخیص دانشجویان خاص (ممتاز یا مشروطی) و برنامه ریزی آموزشی مناسب برای ایشان ایفا می کند.

کلید واژه- Data Mining، داده کاوی آموزشی، دسته بندی مبتنی بر درخت تصمیم.

۱- مقدمه

داده کاوی در واقع کشف و استخراج و تحلیل مقادیر زیادی از داده ها برای کشف الگوها و قواعد معنادار است. در واقع هدف داده کاوی کشف دانش است. فرآیند داده کاوی شامل ۳ مرحله است : آماده سازی داده، یادگیری مدل، ارزیابی و تفسیر مدل. داده کاوی کاربردهای متنوعی دارد. یکی از پرکاربردترین این کاربردها داده کاوی آموزشی است. سیستم های آموزشی حاوی داده های غنی و پر محتوا در مورد رفتار دانشجویان هستند. داده کاوی این داده های آموزشی منجر به غنی کردن و توسعه دادن محیط های آموزشی می شود.

سیستم های آموزش عالی از طریق داده کاوی آموزشی قادرند که اثربخشی سیستم های آموزشی را حداکثر کنند، پذیرش و مدیریت ثبت نام را بهبود دهند، نرخ حذف دانشجویان را حداقل کنند، نرخ گذر دانشجویان را ارتقا دهند و موفقیت دانشجویان را افزایش داده و هزینه سیستم را کاهش دهند. یک موسسه آموزشی از طریق داده کاوی قادر خواهد بود که مزیت رقابتی خود را افزایش داده و به استانداردهای بالاتری در سطح دانشگاهی برسد.

در این پروژه به کمک نرم افزار RapidMiner از طریق روش دسته بندی مبتنی بر درخت تصمیم به پیش بینی وضعیت تحصیلی دانشجویان پرداخته شده است. جهت انجام این پروژه از داده های آموزشی ۱۰۰۰ نفر از دانشجویان موسسه آموزش عالی صفاهان استفاده شده است. مراحل انجام داده کاوی در این پروژه در بخش های آتی توضیح داده شده است.

۲- آماده سازی داده

اولین و مهم ترین مرحله در فرآیند داده کاوی آماده سازی داده است. هدف در این مرحله تامین ورودی مناسب برای مرحله حیاتی یادگیری مدل است. در این مرحله داده پردازش نشده از کل منابع داده ای موجود (که ممکن است توزیع شده نیز باشد) استخراج شده، سپس در مرحله ای مستقل مورد پردازش اولیه قرار می گیرد. خروجی در مرحله آماده سازی داده عبارت است از داده پیش پردازش شده که امکان یادگیری مدل از روی آن وجود دارد. مجموعه عملیات متنوعی جهت آماده سازی و پیش پردازش داده ها انجام می گیرد که در ادامه به آن می پردازیم.

۲-۱- پاک سازی داده

یکی از مشکلات شایع داده پایین بودن کیفیت آن است. به عملیاتی که به برطرف شدن مشکل کیفیت داد ها می انجامد پاک سازی داده گفته می شود. مشکلاتی که کیفیت داده را به

مخاطره می اندازد عبارتند از : نویز، نمونه های پرت، مقادیر از دست رفته و داد ه های دو نسخه ای یا تکراری.

به دلایلی ممکن است بعضی از مقادیر مربوط به برخی ویژگی ها Null باشند. به این گونه مقادیر، مقادیر از دست رفته می گوییم.در این پروژه از بین مشکلات ذکر شده ما با مشکل مقادیر ازدست رفته روبرو بوده ایم . چهار روش برای مدیریت مقادیر از دست رفته وجود دارند که عبارتند از :

حذف کردن : در این روش رکوردهایی که حداقل یکی از ویژگی های آنها Null است حذف می شوند. استفاده از این روش در مواقعی مناسب است که تعداد رکوردهای با مقادیر Null در مقایسه با کل رکوردها کم باشد و یا رکوردهای مشابه با رکوردی که می خواهیم حذف کنیم وجود داشته باشند. این روش شایع ترین روش برخورد با مشکل مقادیر از دست رفته است.

تخمین زدن : در این روش مقادیر Null با استفاده از روش های ابتکاری تخمین زده می شوند. این روش در مواقعی مناسب است که ویژگی ها با یکدیگر همبستگی داشته باشند و یا مقادیر ویژگی مورد تخمین از توزیع و نظم ویژه ای تبعیت کنند.

نادیده گرفتن : در این روش در زمان تحلیل داده ها آن دسته از ویژگی هایی که مقادیرشان Null است را در نظر نمی گیریم. ولی از مقادیر بقیه ویژگی های غیر Null استفاده می کنیم. در مواقعی که انباره داده کوچک و تعداد ویژگی های Null هم زیاد باشد از این روش استفاده می شود. همچنین اگر کاربرد مورد نظر خوشه بندی باشد بهتر است فاصله بین رکوردها بر اساس ویژگی هایی که در بخش اعظم داده معلوم هستند محاسبه شده و سایر ویژگی ها نادیده گرفته شوند.

جایگزین کردن : در این روش مقادیر Null با تمام مقادیر امکان پذیر جایگزین می شوند. این روش همیشه با شرط عدم محدودیت زمان و حافظه بهترین روش است. البته تعداد ویژگی هایی که مقادیرشان Null است و نیز تعداد مقادیر امکان پذیر برای آنها در تصمیم گیری برای انتخاب این روش بسیار تاثیر گذار است. به عنوان مثال اگر فقط یک ویژگی مقادیر Null داشته باشد و آن هم فقط دو مقدار اختیار کند (مانند ویژگی جنسیت) آنگاه استفاده از این روش بسیار مناسب خواهد بود. اما چنان چه تعداد ویژگی های با مقادیر Null زیاد بوده و نیز هر کدام از مقادیر نیز حالت های گوناگونی داشته باشند و یا پیوسته باشند استفاده از این روش اصلا مناسب نخواهد بود. نکته آخر اینکه روش جایگزین کردن در مسائلی که دارای داده های گران بها بوده و تعداد رکوردهای آنها بسیار کم است کاربرد فراوانی دارد.

در این پروژه برای رفع مشکل Missing value (مقادیری که در دسترس ما نیستند) از روش جایگزین کردن و در نرم افزار Rapid miner از عملگر Replace missing استفاده شده است .به کمک این عملگر و و تنظیمات انجام شده، برای جایگزینی مقادیر فیلد "آخرین مدرک تحصیلی " که مقدار Null دارند، مقدار دیپلم قرار داده شد.

۲-۲- انتخاب زیر مجموعه ویژگی ها

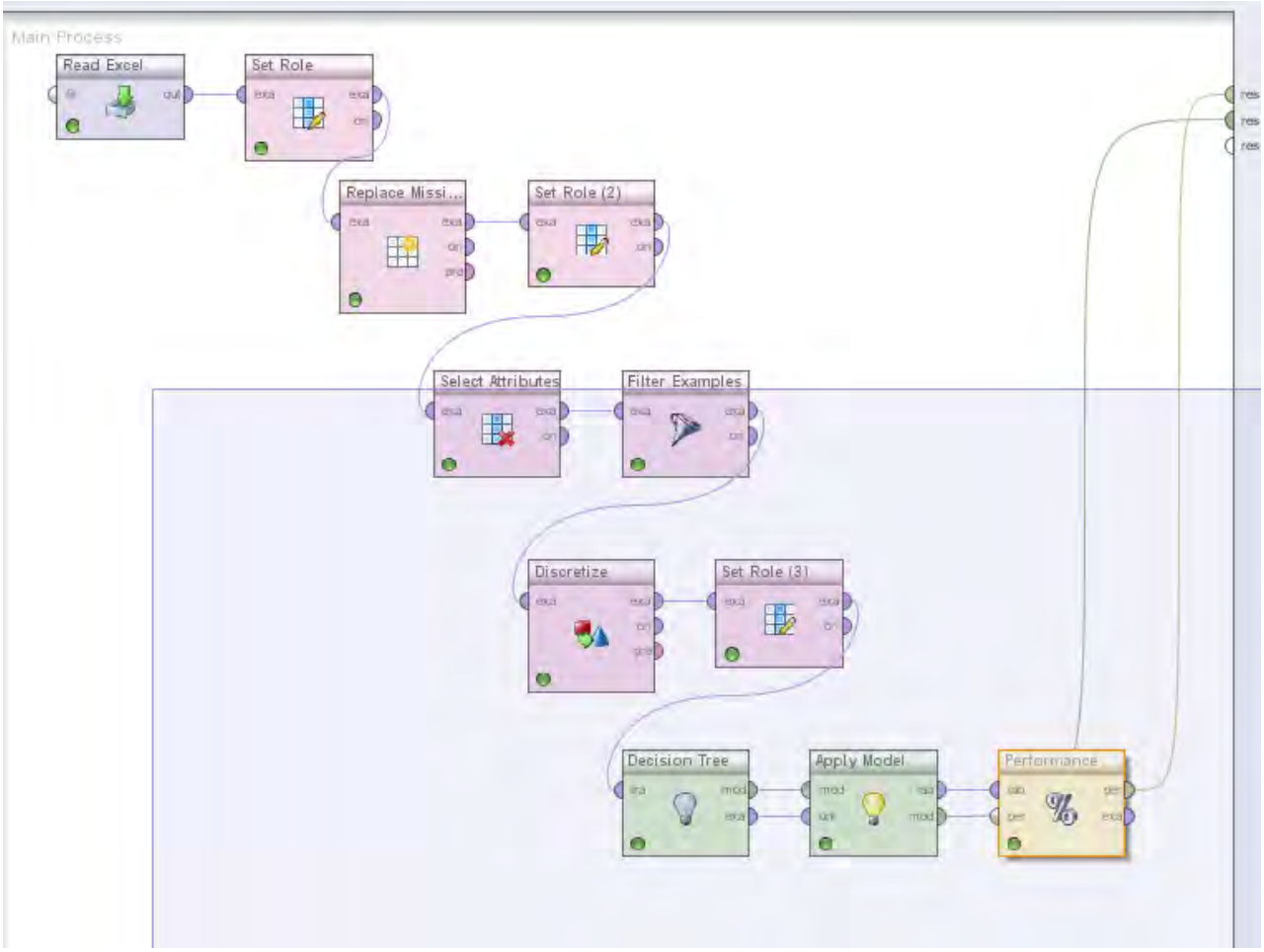
یکی از عملیات های کاهش ابعاد، انتخاب زیر مجموعه ای از ویژگی هاست. در این روش ویژگی های افزونه و غیر مرتبط حذف خواهند شد. ویژگی های افزونه به ویژگی هایی گفته می شود که با توجه به سایر ویژگی ها قابل محاسبه هستند. وجود ویژگی های افزونه باعث خواهد شد فضای الگوریتم بی جهت بزرگ شود. بنابراین انتخاب یک زیر مجموعه از ویژگی ها می تواند بسیار مفید باشد. ویژگی های غیر مرتبط ویژگی هایی هستند که هیچ ارزش اطلاعاتی برای مساله نداشته باشند. روش های انتخاب زیر مجموعه ای از ویژگی ها عبارتند از :

روش نا آگاهانه: در این روش تمام زیر مجموعه های امکان پذیر از ویژگی ها به الگوریتم داده کاوی اعمال خواهند شد.در واقع در این روش الگوریتم داده کاوی فقط یادگیری مدل را با تمامی زیر مجموعه های امکان پذیر از ویژگی ها به عهده دارد و کوششی برای شناسایی هوشمندانه ویژگی های موثر نمی کند.

روش توکار: در این روش الگوریتم داده کاوی هم یادگیری مدل و هم انتخاب ویژگی را به صورت توامان انجام می دهد. در واقع هنگامی که الگوریتم در حال یادگیری مدل است، ویژگی های مهم را نیز انتخاب می کند.

روش فیلتری: در این روش قبل از اجرای الگوریتم داده کاوی، انتخاب ویژگی انجام می شود.سپس ویژگی های انتخاب شده در اختیار الگوریتم داده کاوی قرار می گیرد و از الگوریتم فقط برای ساخت مدل استفاده می شود.

روش انحصاری: در این روش از الگوریتم داده کاوی فقط برای انتخاب ویژگی استفاده می شود و الگوریتم ساخت مدل را دراختیار ندارد.



شکل ۱: نمای کلی عملگرهای استفاده شده از نرم افزار Rapidminer در این پروژه

در این پروژه از روش فیلتری و برای تمرکز بر بخش خاصی از ویژگی های مجموعه داده در نرم افزار Rapid miner از عملگر Select attribute استفاده شده است.بر طبق این روش، قبل از اجرای الگوریتم داده کاوی، انتخاب ویژگی صورت گرفت. این ویژگی ها عبارتند از: شماره دانشجویی، رشته تحصیلی، گرایش، سال ورود به دانشگاه، نیمسال ورود به دانشگاه، جنسیت، استان محل سکونت، وضعیت تاهل، معدل هر ترم به صورت جداگانه، معدل کل، وضعیت فارغ التحصیلی و تاریخ فارغ التحصیلی.

۲-۳-فیلترینگ نمونه ها

به عملیات انتخاب زیر مجموعه ای از رکوردها برای کاوش فیلترینگ نمونه ها می گوییم .از آنجایی که مجموعه داده شامل داده های همه دانشجویان اعم از فارغ التحصیلان و یا دانشجویان فعال در ترم های مختلف هستند و از آنجایی که در این پروژه پیش بینی وضعیت تحصیلی دانشجو از طریق روش دسته بندی مد نظر است، بنابراین می بایست که الگوریتم یادگیری را بر روی داده های فارغ التحصیلان که تقریبا تمامی فیلدهای آنان دارای مقداراست، اعمال نماییم .جهت تحقق این امر در نرم افزار Rapid miner از عملگر Filter example استفاده شده است. طبق تنظیمات انجام شده در این عملگر رکوردهایی که وضعیت فارغ التحصیلی انان دارای مقدار ۱ است برای کاوش داده انتخاب می شوند.

۲-۴-گسسته سازی

در عملیات گسسته سازی داده هدف آن است که نوع ویژگی های بازه ای و نرخی به نوع اسمی تبدیل شوند. این کار به این منظور صورت می پذیرد که در مسایل پیچیده داده کاوی (که در آنها داده های با انواع ویژگی های گوناگون و همچنین تعداد بالای رکوردها و ویژگی ها وجود دارند) با گسسته سازی داده، سختی مساله را کاهش داده و زمینه لازم را برای عملکرد موثر الگوریتم های یادگیری مدل فراهم آوریم. انواع روش های گسسته سازی داده عبارتند از :

گسسته سازی بسامدی: در این روش گسسته سازی بر

اساس بسامد رخداد رکوردها در بازه ها صورت می گیرد.

گسسته سازی اندازه ای: در این روش اندازه پارامتری به نام اندازه بازه تعریف می شود.

گسسته سازی بخشی: در این روش پارامتری به نام تعداد بخش ها تعریف می شود.

گسسته سازی اطلاعاتی : در این روش هدف کمینه کردن آنتروپی بازه های تولید شده است.

گسسته سازی انتخابی: در این روش تعداد بازه های تولیدی و همچنین محدوده هر بازه به صورت پارامتر به الگوریتم گسسته سازی اعلام می شود.در واقع در این روش گسسته سازی همه عملیات در اختیار شخص کاوشگر است و وجود دانش پس زمینه از داده مورد کاوش اهمیت بالایی دارد.

در این پروژه برای انجام عملیات گسسته سازی از روش گسسته سازی انتخابی و در نرم افزار Rapid miner از عملگر Discretize by user specification استفاده شده است و طبق تنظیمات انجام شده ، ویژگی "معدل کل" به چهار دسته خیلی خوب، خوب، نرمال و ضعیف تفکیک شده است.

۲-۵-تبدیل داده کاوش انبوهش داده، خلق ویژگی

در عملیات تبدیل داده از یک تابع استفاده می شود که مجموعه کل مقادیر یک ویژگی مفروض را به یک مجموعه جدیدی از مقادیر نگاشت می کند. در این پروژه تبدیل داده ضرورتی نداشته است.

در این پروژه از یکی از عملیات های کاهش ابعاد (انتخاب زیر مجموعه ای از ویژگی ها) استفاده شده است.

انبوهش داده عبارت است از ترکیب دو یا چند ویژگی (یا رکورد) و ایجاد یک ویژگی (یا رکورد) جدید. در این پروژه انبوهش داده ضرورتی نداشته است.

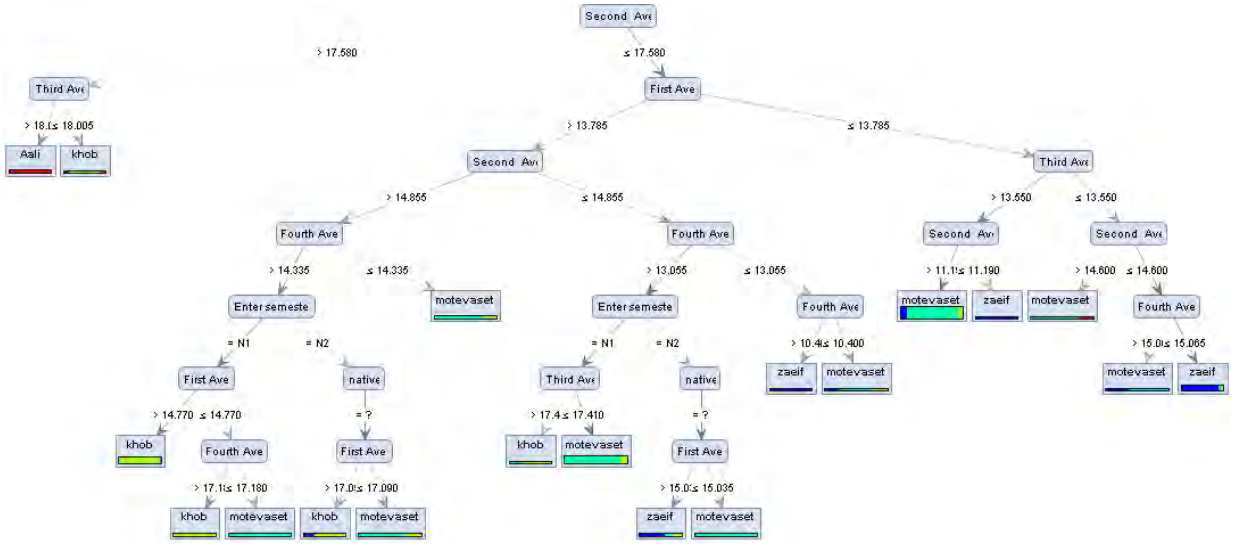
خلق ویژگی عبارت است از خلق ویژگی های جدیدی که بتوانند در کنار سایر ویژگی های پیشین اطلاعات مهم موجود در یک مجموعه داده را موثر تر و کامل تر از ویژگی های اولیه نمایش دهند.در این پروژه نیازی به ایجاد ویژگی جدید وجود نداشت. نمای کلی عملگرهای استفاده شده از نرم افزار Rapidminer در این پروژه در شکل شماره ۱ آمده است.

۳- یادگیری مدل

روش های مختلف کاوش داده را می توان در دو گروه روش های پیش بینی و روش های توصیفی طبقه بندی نمود.

روش های پیش بینی از مقادیر بعضی از ویژگی ها برای پیش بینی کردن مقادیر یک ویژگی مشخص استفاده می کنند. در متون علمی مختلف روش های پیش بینی با نام روش های با ناظر نیز شناخته می شوند. روش های **دسته بندی، رگرسیون، و تشخیص انحراف** سه روش یادگیری مدل در داده کاوی با ماهیت پیش بینی هستند. روش های توصیفی الگوهای قابل توصیفی را پیدا می کنند که روابط حاکم بر داده ها را بدون در نظر گرفتن هر گونه برچسب و یا متغیرخروجی تبیین نمایند. در متون علمی مختلف روش های توصیفی با نام روش های بدون ناظر نیز شناخته می شوند. روش های **خوشه بندی، کاوش قوانین انجمنی و کشف الگوهای ترتیبی** سه روش یادگیری مدل در داده کاوی با ماهیت توصیفی هستند. در این پروژه از روش دسته بندی که جز روش های پیش بینی است، جهت پیش بینی وضعیت تحصیلی دانشجویان استفاده شده است.

در الگوریتم های دسته بندی مجموعه داده اولیه به دو مجموعه داده با عنوان مجموعه داده های آموزشی و مجموعه داده آزمایشی تقسیم می شود، با استفاده از مجموعه داده های آموزشی مدل ساخته می شود و از مجموعه داده آزمایشی برای اعتبار سنجی و محاسبه دقت مدل ساخته شده استفاده می شود. هر رکورد شامل یک مجموعه از ویژگی هاست. یکی از ویژگی ها، **ویژگی دسته** نامیده می شود. در الگوریتم های دسته بندی چون ویژگی دسته مربوط به هر رکورد مشخص است بنابراین جزء الگوریتم های باناظر محسوب می شوند. الگوریتم های با ناظر شامل دو مرحله با عنوان مرحله آموزش (یادگیری) و مرحله ارزیابی هستند. در مرحله آموزش مجموعه داده های آموزشی به یکی از الگوریتم های دسته بندی داده می شود تا بر اساس مقادیر سایر ویژگی ها برای مقادیر ویژگی دسته، مدل ساخته شود. شکل مدل ساخته شده به نوع الگوریتم یادگیرنده بستگی دارد. به عنوان مثال اگر الگوریتم یادگیرنده الگوریتم درخت تصمیم باشد مدل ساخته شده یک درخت تصمیم خواهد بود. اگر الگوریتم یادگیرنده یک دسته بندی بر قانون باشد مدل ساخته شده یک مجموعه قانون خواهد بود. در هر صورت با توجه به الگوریتم یادگیرنده مورد استفاده در مرحله آموزش، مدل ساخته می شود. پس از ساخت مدل در مرحله ارزیابی دقت مدل ساخته شده به کمک مجموعه داده های آزمایشی که مدل ساخته شده در مرحله آموزش این مجموعه داده ها را ندیده است ارزیابی خواهد شد. از مجموعه داده های آزمایشی در مرحله آموزش و ساخت مدل استفاده نمی شود.



شکل ۲: مدل حاصل از اعمال الگوریتم دسته بندی مبتنی بر درخت تصمیم

در روش دسته بندی یک زیر مجـــــموعه داده آموزشـی و یک مجموعه داده آزمایـــشی موجود است . مجموعـه داده هـای آموزشی به بخش یادگیری مدل دسته بنـدی ارسـال مـی شـود. درادامه مدل ساخته شده بر اساس مجموعه داده هـای آزمایشـی مورد سنجش قرار می گیرد. الگوریتم دسته بندی کننده، ویژگـی دسته هر رکـورد را در مجموعـه داده هـای آزمایشـی در فرآینـد پیش بینی دسته مورد استفاده قرار نمی دهد.

انواع گوناگون الگوریتم های دسته بندی عبارتند از:

روشهای مبتنی بر درخت تصمیم

روشهای مبتنی بر قانون

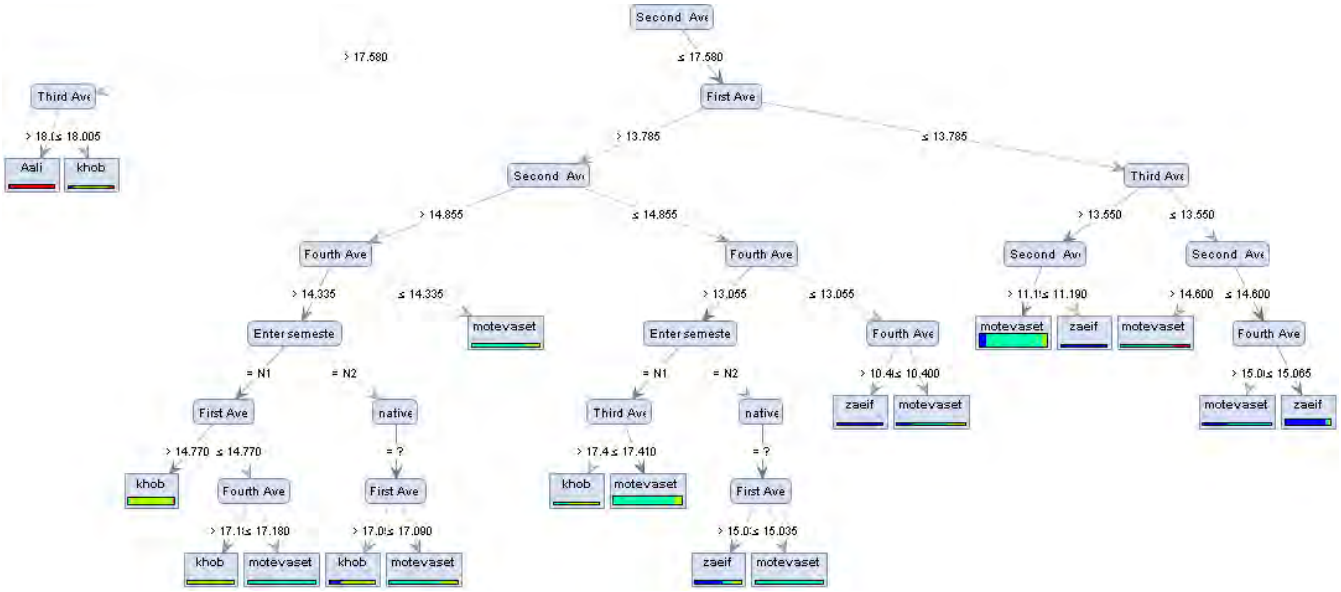
روشهای مبتنی بر حافظه

شبکه های عصبی

روشهای مبتنی بر بیز

ماشین های بردار پشتیبان

در این پروژه از الگوریتم درخت تصمیم بـرای دسـته بنـدی استفاده شـده اسـت. درخـت تصـمیم یکـی از ابزارهـای قـوی و متداول برای دسته بندی و پیش بینی می باشد. درخـت تصـمیم برخلاف شبکه های عصبی به تولید قاعده می پـردازد.در سـاختار



PerformanceVector (Performance (2)) | PerformanceVector (Performance) | Tree (Decision Tree)

ext View | Annotations

Multiclass Classification Performance | Annotations

Table View | Plot View

accuracy: 71.33% +/- 5.74% (mikro: 71.32%)

	true bad	true normal	true good	true very good	class precision
pred. bad	167	43	7	1	76.61%
pred. normal	16	30	14	0	50.00%
pred. good	7	18	75	5	71.43%
pred. very good	1	0	3	14	77.78%
class recall	87.43%	32.97%	75.76%	70.00%	

شکل ۳: نتایج ارزیابی حاصل از اعمال روش K-fold cross validation

درخت تصمیم، پیش بینی به دست آمده از درخت در قالب یکسری قواعد توضیح داده می شود.

از آنجا که در این پروژه از روش دسته بندی استفاده می شود ، باید یک ویژگی به عنوان ویژگی دسته تعریف شود و دارای برچسب label باشد. بنابراین خروجی عملگر گسسته سازی به ورودی عملگر Set role داده شده است و در آنجا ویژگی "معدل کل" به عنوان ویژگی دسته و برچسب تعریف شود.سپس خروجی آن را به عملگر Disicion Tree داده شده است تا این عملگر مدل مربوطه را بر اساس ورودی که به عملگرداده شده است، بسازد. این عملگر با قابلیت یادگیری از انواع داده ورودی چه اسمی باشند و چه عددی یک درخت تصمیم را به عنوان مدل خروجی میسازد. این درختان برای دسته بندی رکوردهای اعمال شده مدل را از بالا به پایین پیمایش می کنند. در این پروژه عملگر یادگیرنده درخت تصمیم از الگوریتم C4.5 استفاده می کند. هرگره در یک درخت تصمیم با یک ویژگی برچسب گذاری می شود. مقدار ویژگی متناظر رکورد برای این ویژگی شاخه خروجی را تعیین می کند. برای هر مقدار ممکن است از ویژگی های اسمی یک شاخه خروجی خواهیم داشت. همچنین برای ویژگی های عددی شاخه های خروجی با بازه گسسته برچسب گذاری خواهند شد. این الگوریتم به این صورت است که وقتی یک گره جدید در یک مرحله معین ایجاد شد یک ویژگی را برای آن گره انتخاب می کند که قدرت تفکیک آن گره را با توجه به رکوردهایی که به یک زیر درخت خاص تعلق پیدا می کنند حداکثر کند. این قدرت تفکیک با معیارهای مختلفی از قبیل بهره اطلاعات، نرخ بهره، شاخص جینی و... اندازه گیری می شود. در این پروژه حداکثر عمق درخت(maximal depth) برابر ۲۰ و حداقل بهره ای(minimal gain) که باید حاصل شود ۰.۰۱ و درجه اطمینان(confidence) برابر ۰.۲۵ تعریف شده است. مدل به دست آمده در شکل شماره ۲ آمده است.

سپس برای اعمال مدل ساخته شده بر روی یک مجموعه داده از عملگر Apply Model استفاده شده است . طی این عملیات هر رکورد از ورودی مورد بررسی قرار می گیرد و یک برچسب به هر رکورد نسبت داده می شود.

۴- ارزیابی و تفسیر مدل

در این مرحله دانش تولید شده در مرحله قبل ارزیابی شده و مورد تفسیر قرار می گیرد. منظور از ارزیابی دانش آن است که می بایست میزان صحت دانش تولید شده مشخص شود تا بتوان به آن اعتماد نمود و به صورت عملی از آن استفاده کرد. روش های مختلفی برای ارزیابی دانش تولید شده وجود دارند که رابطه تنگاتنگی با روش یادگیری مدل دارند. نفسیر مدل به معنای آن است که دانش تولید شده را مورد بررسی قرار داده و توجیهی معنایی جهت تبیین منطق آن ارائه نماییم.

در الگوریتم های دسته بندی اینکه چه بخشی از مجموعه داده اولیه برای آموزش و چه بخشی برای آزمایش به کار رود، بستگی کاملی به روش ارزیابی مورد استفاده دارد. روش های مشهور ارزیابی الگوریتم های دسته بندی عبارتند از:

روش Holdout: در این روش چگونگی نسبت تقسیم مجموعه داده ها بستگی به تشخیص تحلیل گر دارد. معمولا دو نسبت ۵۰-۵۰ و ۷۵-۲۵ بیشتر استفاده می شوند.

روش Random subsampling: این روش، از روش Holdout قابل اعتمادتر است و روش Holdout را چندین مرتبه تکرار می کند.

روش Cross- Validation: در این روش اگر مجموعه داده به دو بخش آموزشی و آزمایشی تقسیم شود.در مرحله اول ابتدا برای ساخت مدل از داده اموزشی و برای ارزیابی مدل از مجموعه داده آزمایشی استفاده می شود و در مرحله دوم از مجموعه داده آموزشی برای ارزیابی زمایشی برای ساخت مدل استفاده می شود.میانگین دقت های محاسبه شده در دو مرحله را به عنوان دقت نهایی معرفی می نماییم. به این روش 2-fold cross validation گفته می شود. اگر به جای ۲ قسمت کردن و ۲ بار انجام دادن این مراحل، مجموعه داده رابهK قسمت تقسیم و K بار مراحل انجام شود روش را K-fold cross validation می گوییم .معمواترین مقدار برای k مقدار ۱۰ است.

روش Bootstrap: در این روش برای انجام فرایند یادگیری مدل از مجموعه داده اولیه به صورت نمونه برداری با جایگذاری انتخاب خواهند شد.سپس مجموعه رکوردهای انتخاب نشده برای ارزیابی دسته بند مورد استفاده قرار می گیرند.

در این پروژه برای ارزیابی مدل ساخته شده از روش K-fold cross validation و عملگر X-validation استفاده شده است مقدار k=10 و نمونه برداری از نوع stratified sampling است که داده ها به صورت متوازن از دسته های تعیین شده انتخاب خواهند شد. همچنین از عملگر performance که از عملگرهای ارزیابی عمومی است نیز استفاده شده است.

مهمترین معیار برای تعیین کارایی یک الگوریتم دسته بندی معیار دقت (Accuracy)است .این معیار دقت کل یک دسته بند را محاسبه می کند . مقدار دقت حاصل از عملگر performance برابر ۸۸.۵۳٪ است و دقت حاصل از عملگر X-validation برابر (mikro: 71.32%) 47.5% +/- ۷۱.۳۳٪ می باشد.

معیار recall دقت دسته بندی دسته x را با توجه به کل رکوردهای با برچسب x نشان می دهد. معیار precision دقت دسته بندی دسته x را با توجه به کل مواردی نشان می دهد که برچسب x برای رکورد مورد بررسی توسط دسته بند پیشنهاد شده است.این معیارها در شکل شماره ۳ آمده است.

۵- نتیجه‌گیری

در این تحقیق با استفاده از رویکردی کاربردی روش دسته بندی مبتنی بر درخت تصمیم بر روی داده های دانشجویان موسسه آموزش عالی صفاهان اعمال گردید و مدل کاوشی به منظور پیش بینی وضعیت تحصیلی دانشجویان بر اساس مشخصه های آموزشی ایشان استخراج گردید.

در درخت تصمیم صفت های مهم تر در گره های بالای درخت قرار می گیرد. در این پروژه معدل ترم دوم در راس درخت قرار دارد که این مسئله دارای توجیه منطقی است . از آنجایی که اغلب دانشجویان در ترم اول به دلایل مختلف دچار افت تحصیل می شوند و از ترم دوم جدیت بیشتری در یادگیری دارند، بنابراین کاملا منطقی است که راس درخت از معدل ترم دوم شروع گردد. بنابراین جهت ارزیابی و پیش بینی آینده آموزشی ایشان در ترم های اول استفاده از معدل ترم دوم اهمیت و اعتبار بالاتری نسبت به معدل ترم اول داراست. این مدل برای پیش بینی وضعیت تحصیلی دانشجویان می تواند مورد استفاده قرار گیرد و براساس این پیش بینی دانشجویان در ۴ دسته عالی، خوب، متوسط و ضعیف رده بندی می گردند که بر این اساس می توان اقدامات موردنیاز را برای حمایتشان و جلوگیری از مشکلات آتی برنامه ریزی و اعمال نمود.

مراجع

[1] Jiawei Han and Micheline Kamber, Data mining: concepts and Techniques, second edition, morgan Kaufmann publishers, pp. 285-306.

[2] [illegible]

[3] [illegible]

[5] [illegible]