عنوان: **یادگیری مدل در داده کاوی**
نویسنده: محمد رجبی
تاریخ: ۱۳۹۳/۰۹/۰۱ ۱۳:۴۶
آدرس: www.dotnettips.info
گروه‌ها: Analysis Services, data mining, Microsoft SQL Server

مقدمه هدف اصلی داده کاوی کشف دانش است، که این دانش نظمی که در داده‌ها وجود دارد را نمایان می‌سازد. پس از کشف دانش ممکن است با دو وضعیت مواجه شویم:
حالت اول هنگامی است که افراد خبره در دامنه داده مورد کاوش، آگاه به دانش استخراج شده باشند که در این صورت آن دانش به عنوان یک قانون صحیح تلقی خواهد شد.
در حالت دوم ممکن است دانش کشف شده، یک دانش جدید بوده و در بین افراد خبره در آن حوزه شناخته شده نباشد، در این صورت این دانش بررسی شده و در صورت منطقی بودن تبدیل به فرضیه شده و در نهایت درست یا غلط بودن این فرضیه با آزمایشات و بررسی‌های متعدد اثبات می‌شود و در صورت درست بودن فرضیه تبدیل به قانون خواهد شد.

روش‌های یادگیری مدل در داده کاوی پیشتر به معرفی مراحل کاری در داده کاوی که مشتمل بر سه مرحله اساسی: **آماده سازی داده** ، **یادگیری مدل** و در نهایت **ارزیابی و تفسیر مدل** می‌باشد، پرداختیم.
در مرحله یادگیری مدل با استفاده از الگوریتم‌های متنوع و با در نظر گرفتن ماهیت داده، نظم‌های مختلف موجود در داده‌ها شناسائی می‌شود. بطور کلی روش‌های مختلف کاوش داده را به دو گروه روش‌های پیش بینی و روش‌های توصیفی طبقه بندی می‌کنند.

در **روش‌های پیش بینی** از مقادیر بعضی ویژگی‌ها برای پیش بینی کردن مقدار یک ویژگی مشخص استفاده می‌کنند. این روش‌ها در متون علمی با نام روش‌های با ناظر (Supervised Methods) نیز شناخته می‌شوند. الگوریتم‌های با ناظر از دو مرحله با عنوان مرحله آموزش (یادگیری) و مرحله ارزیابی تشکیل شده اند.
در مرحله آموزش؛ با استفاده از مجموعه داده‌های آموزشی مدل ساخته می‌شود. شکل مدل ساخته شده به نوع الگوریتم یادگیرنده بستگی دارد.
در مرحله ارزیابی؛ از مجموعه داده‌های آزمایشی برای اعتبارسنجی و محاسبه دقت مدل ساخته شده استفاده می‌شود، در واقع از داده هایی که در مرحله آموزش و ساخت مدل؛ الگوریتم این مجموعه داده‌ها را ندیده است (Previously Unseen Data) استفاده می‌شود.
برای نمونه روش‌های **دسته بندی** (Classification)، **رگرسیون** (Regression) و **تشخیص انحراف** (Anomaly Detection) سه روش یادگیری مدل در داده کاوی با ماهیت پیش بینی هستند.
در **روش‌های توصیفی** همانطور که انتظار داریم الگوهای قابل توصیف از روابط حاکم بر داده‌ها بدون در نظر گرفتن هر گونه برچسب و یا متغیر خروجی بدست می‌آید. این روش‌ها در متون علمی با نام روش‌های بدون ناظر (Unsupervised Methods) نیز شناخته می‌شوند. برای نمونه روش‌های **خوشه بندی** (Clustering)، **کاوش قوانین انجمنی** (Association Rules Mining) و **کشف الگوهای ترتیبی** (Sequential Pattern Discovery) سه روش یادگیری مدل در داده کاوی با ماهیت توصیفی هستند.

در ادامه به معرفی هر کدام از این روش‌ها می‌پردازیم:

دسته بندی: در الگوریتم‌های دسته بندی مجموعه داده اولیه به دو مجموعه داده با عنوان مجموعه داده‌های آموزشی (Train Dataset) و مجموعه داده‌های آزمایشی (Test Dataset) تقسیم می‌شود. می‌دانیم هر Case شامل مجموعه ای از Attribute هاست، که یکی از این ویژگی‌ها **ویژگی دسته** نامیده می‌شود.
در مرحله آموزش؛ مجموعه داده‌های آموزشی به یکی از الگوریتم‌های دسته بندی داده می‌شود تا بر اساس سایر ویژگی‌ها برای مقادیر ویژگی دسته، مدل ساخته شود.
پس از ساخت مدل، در مرحله ارزیابی؛ دقت مدل ساخته شده به کمک مجموعه داده‌های آزمایشی ارزیابی خواهد شد. در الگوریتم‌های دسته بندی از آنجا که ویژگی دسته مربوط به هر Case مشخص است به صورت الگوریتم‌های با ناظر محسوب می‌شوند. بدیهی است که تشخیص بر اساس دسته هایی است که مدل در مرحله آموزش با آنها روبرو شده است؛ بنابراین امکان تشخیص دسته جدید در کاربرد دسته بندی وجود نخواهد داشت.

رگرسیون: رگرسیون در علوم آمار و شبکه‌های عصبی بطور وسیعی مورد بررسی و مطالعه قرار می‌گیرد. پیش بینی مقدار یک متغیر پیوسته بر اساس مقادیر سایر متغیرها بر مبنای یک مدل وابستگی خطی یا غیر خطی رگرسیون نامیده می‌شود. یک نوع خاصی از رگرسیون، **پیش بینی سری‌های زمانی** (Time Series Prediction) است؛ برای مثال تغییرات قیمت سهام شرکتی را به صورت نمودار داریم؛ می‌خواهیم ادامه روند این نمودار را برای مدتی مشخص پیش بینی کنیم. در مسائل سری‌های زمانی یکی از متغیرهای اصلی زمان می‌باشد. بدیهی است که رگرسیون لزوماً سری زمانی نیست و همانند دسته بندی کاربرد رگرسیون نیز از نوع پیش بینی با ناظر است و بطور مشابه در رگرسیون هم دو مرحله آموزش و ارزیابی نیز وجود دارد. مثال هایی از رگرسیون می‌تواند شامل موارد زیر باشد: پیش بینی میزان فروش یک محصول جدید، براساس میزان فروش محصولات گذشته و یا براساس میزان تبلیغات انجام شده و ... همچنین مسائل مربوط به پیش بینی سری‌های زمانی از قبیل بورس و ... .

تشخیص انحراف: از کاربردهای متداول تشخیص انحراف، می‌توان به **کشف کلاهبرداری** کارت‌های اعتباری (Credit Card Fraud Detection) اشاره کرد. در مواقعی از این کاربرد استفاده می‌شود که تنها نمونه هایی با یک برچسب یکسان که معمولاً وضعیت نرمال را نشان می‌دهند در دسترس می‌باشند و امکان مالکیت بر داده‌ها با تمامی برچسب‌های موجود به دلایل مختلف وجود ندارد. بنابراین چون فقط نمونه‌های دسته نرمال در اختیار است، الگوریتم برای وضعیت نرمال و با توجه به یک آستانه (Threshold) مشخص مدل را می‌سازد و هر گونه تخطی از آن آستانه را؛ بعنوان وضعیت غیرنرمال در نظر می‌گیرد. توجه شود روش‌های دسته بندی تنها قادر به شناسائی دسته هایی هستند که در مرحله آموزش، نمونه ای از آنها به الگوریتم ارائه شده است، بنابراین امکان تشخیص هیچ گونه کلاهبرداری توسط روش‌های دسته بندی وجود ندارد.

خوشه بندی: در این مسائل از آنجا که بر خلاف دسته بندی هیچ گونه دسته خاصی وجود ندارد، بنابراین براساس معیار شباهت داده‌ها گروه بندی و خوشه بندی صورت می‌گیرد. بدین ترتیب Case هایی که بیشترین شباهت را به یکدیگر دارند در یک خوشه قرار می‌گیرند، به بیان دیگر Caseهای موجود در خوشه‌های متفاوت کمترین شباهت را به یکدیگر خواهند داشت. بدیهی است که خوشه بندی براساس ویژگی ورودی نمونه‌ها انجام می‌گیرد و از آنجائی که برای این الگوریتم‌ها ویژگی دسته تعریف نمی‌شود و Caseها برچسب خاصی ندارند، جزء الگوریتم‌های بدون ناظر محسوب می‌شوند. در واقع هدف در تمامی الگوریتم‌های خوشه بندی **کمینه کردن فاصله درون خوشه ای** (Intra-Cluster Density) و **بیشینه نمودن فاصله بین خوشه ای** (Inter-Cluster Density) است و عملکرد خوب یک الگوریتم خوشه بندی زمانی محرز می‌شود که تا حد امکان خوشه‌ها را از یکدیگر دورتر کند و در ضمن Caseهای موجود در یک خوشه بیشترین شباهت را به یکدیگر داشته باشند.

کشف قوانین انجمنی: قوانین وابستگی (انجمنی) اتفاق و وقوع یک شیء را براساس وقوع سایر اشیاء توصیف می‌کنند، برای مثال در یک سوپر مارکت هدف در کاوش قوانین انجمنی؛ یافتن نظم حاکم بر سبد خرید می‌باشد، در این کاربرد به ازای هر سبد؛ یک قانون پیدا می‌شود و بررسی خواهد شد که این قانون در چه تعداد از سبدها صدق می‌کند و در نهایت یک مجموعه قوانین که در بیشترین تعداد از سبدها صدق می‌کند به عنوان مجموعه قوانین انجمنی خروجی ارائه می‌شود. به بیان دیگر در این کاربرد به دنبال پیدا کردن یک مجموعه از قوانین وابستگی هستیم تا براساس آن قوانین بتوانیم نتیجه گیری کنیم وجود کدامیک از مجموعه اشیاء (Item Set) بر وجود چه مجموعه اشیاء دیگری تاثیر گذار است.

کشف الگوهای ترتیبی: در این کاربرد به دنبال کشف الگوهایی هستیم که وابستگی‌های ترتیبی محکمی را در میان وقایع مختلف نشان می‌دهند. این کاربرد مشابه کاوش قوانین انجمنی می‌باشد با این تفاوت که در کاوش قوانین انجمنی زمان و ترتیب زمانی مطرح نیست، اما در کشف الگوهای ترتیبی زمان و ترتیب اهمیت ویژه ای دارند برای مثال می‌توان به دنباله‌های تراکنش‌های فروش اشاره نمود.

منبع: با اندکی تغییر و تلخیص "داده کاوی کاربردی در RapidMiner، انتشارات نیاز دانش"