عنوان: **ارزیابی و تفسیر مدل در داده کاوی**
نویسنده: محمد رجبی
تاریخ: ۱۳۹۳/۰۹/۱۱ ۱۸:۰
آدرس: www.dotnettips.info
گروه‌ها: Analysis Services, data mining, Microsoft SQL Server

مقدمه

دانشی که در مرحله یادگیری مدل تولید می‌شود، می‌بایست در مرحله ارزیابی مورد تحلیل قرار گیرد تا بتوان ارزش آن را تعیین نمود و در پی آن کارائی الگوریتم یادگیرنده مدل را نیز مشخص کرد. این معیارها را می‌توان هم برای مجموعه داده‌های آموزشی در مرحله یادگیری و هم برای مجموعه رکوردهای آزمایشی در مرحله ارزیابی محاسبه نمود. همچنین لازمه موفقیت در بهره مندی از علم داده کاوی تفسیر دانش تولید و ارزیابی شده است.

ارزیابی در الگوریتم‌های دسته بندی

برای سادگی معیارهای ارزیابی الگوریتم‌های دسته بندی، آنها را برای یک مسئله با دو دسته ارائه خواهیم نمود. در ابتدا با مفهوم ماتریس درهم ریختگی (Classification Matrix) آشنا می‌شویم. این ماتریس چگونگی عملکرد الگوریتم دسته بندی را با توجه به مجموعه داده ورودی به تفکیک انواع دسته‌های مساله دسته بندی، نمایش می‌دهد.

		Predicted	
		Positive	Negative
Actual	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

هر یک از عناصر ماتریس به شرح ذیل می‌باشد:

TN: بیانگر تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته بندی نیز دسته آنها را بدرستی منفی تشخیص داده است.

TP: بیانگر تعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته بندی نیز دسته آنها را بدرستی مثبت تشخیص داده است.

FP: بیانگر تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته بندی دسته آنها را به اشتباه مثبت تشخیص داده است.

FN: بیانگر تعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته بندی دسته آنها را به اشتباه منفی تشخیص داده است.

مهمترین معیار برای تعین کارایی یک الگوریتم دسته بندی دقت یا نرخ دسته بندی (Classification Accuracy - Rate) است که این معیار دقت کل یک دسته بند را محاسبه می‌کند. در واقع این معیار مشهورترین و عمومی‌ترین معیار محاسبه کارایی الگوریتم‌های دسته بندی است که نشان می‌دهد، دسته بند طراحی شده چند درصد از کل مجموعه رکوردهای آزمایشی را بدرستی دسته بندی کرده است.

دقت دسته بندی با استفاده از **رابطه I** بدست می‌آید که بیان می‌کند دو مقدار TP و TN مهمترین مقادیری هستند که در یک مسئله دودسته ای باید بیشینه شوند. (در مسائل چند دسته ای مقادیر قرار گرفته روی قطر اصلی این ماتریس - که در صورت کسر محاسبه CA قرار می‌گیرند - باید بیشینه باشند.)

معیار خطای دسته بندی (Error Rate) دقیقاً برعکس معیار دقت دسته بندی است که با استفاده از **رابطه II** بدست می‌آید. کمترین مقدار آن برابر صفر است زمانی که بهترین کارایی را داریم و بطور مشابه بیشترین مقدار آن برابر یک است زمانی که کمترین

کارائی را داریم.
ذکر این نکته ضروری است که در مسائل واقعی، معیار دقت دسته بندی به هیچ عنوان معیار مناسبی برای ارزیابی کارایی الگوریتم‌های دسته بندی نمی‌باشد، به این دلیل که در رابطه دقت دسته بندی، ارزش رکوردهای دسته‌های مختلف یکسان در نظر گرفته می‌شوند. بنابراین در مسائلی که با دسته‌های نامتعادل سروکار داریم، به بیان دیگر در مسائلی که ارزش دسته ای در مقایسه با دسته دیگر متفاوت است، از معیارهای دیگری استفاده می‌شود.
همچنین در مسائل واقعی معیارهای دیگری نظیر DR و FAR که به ترتیب از **روابط III و IV** بدست می‌آیند، اهمیت ویژه ای دارند. این معیارها که توجه بیشتری به دسته بند مثبت نشان می‌دهند، توانایی دسته بند را در تشخیص دسته مثبت و بطور مشابه تاوان این توانایی تشخیص را تبیین می‌کنند. معیار DR نشان می‌دهد که دقت تشخیص دسته مثبت چه مقدار است و معیار FAR نرخ هشدار غلط را با توجه به دسته منفی بیان می‌کند.

$$\text{I:} \qquad CA = \frac{TN+TP}{TN+FN+TP+FP}$$

$$\text{II:} \qquad ER = \frac{FN+FP}{TN+FN+TP+FP} = 1 - CA$$

$$\text{III:} \qquad DR = \frac{TP}{FN+TP}$$

$$\text{IV:} \qquad FAR = \frac{FP}{TN+FP}$$

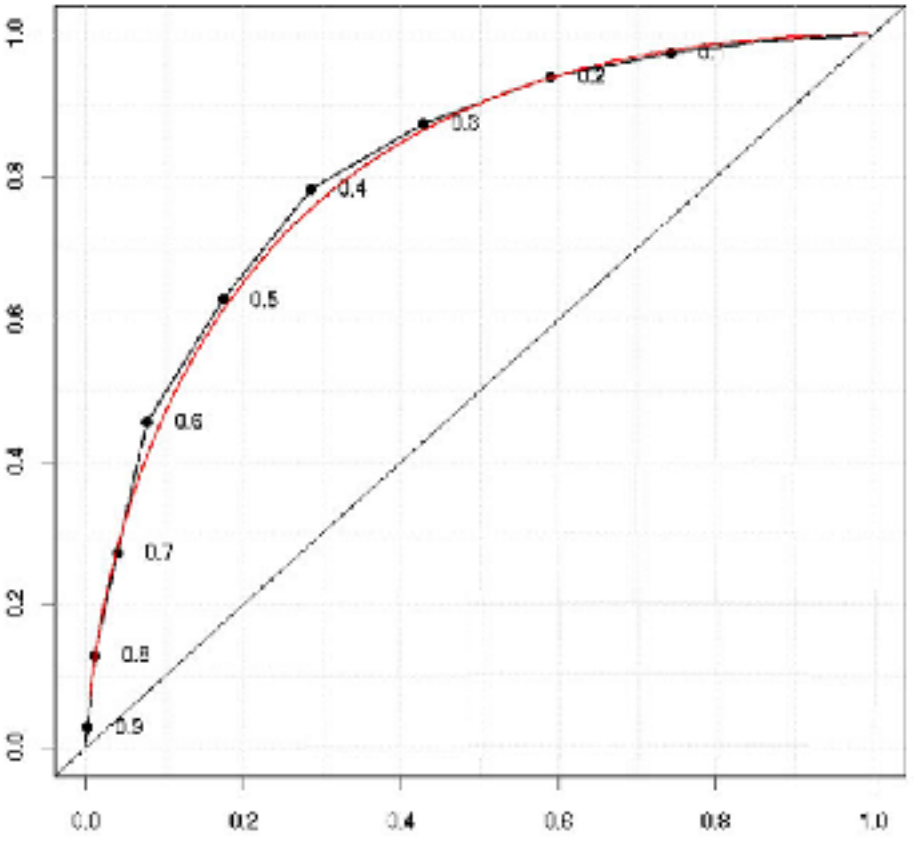
معیار مهم دیگری که برای تعیین میزان کارایی یک دسته بند استفاده می‌شود معیار (Area Under Curve) **AUC** است.

AUC نشان دهنده سطح زیر نمودار (Receiver Operating Characteristic) **ROC** می‌باشد که هر چه مقدار این عدد مربوط به یک دسته بند بزرگتر باشد کارایی نهایی دسته بند مطلوب‌تر ارزیابی می‌شود. نمودار ROC روشی برای بررسی کارایی دسته بندها می‌باشد. در واقع منحنی‌های ROC منحنی‌های دو بعدی هستند که در آنها DR یا همان نرخ تشخیص صحیح دسته مثبت (True Positive Rate - TPR) روی محور Y و بطور مشابه FAR یا همان نرخ تشخیص غلط دسته منفی (False Positive Rate - FPR) روی محور X رسم می‌شوند. به بیان دیگر یک منحنی ROC مصالحه نسبی میان سودها و هزینه‌ها را نشان می‌دهد.

بسیاری از دسته بندها همانند روش‌های مبتنی بر درخت تصمیم و یا روش‌های مبتنی بر قانون، به گونه ای طراحی شده اند که تنها یک خروجی دودویی (مبنی بر تعلق ورودی به یکی از دو دسته ممکن) تولید می‌کنند. به این نوع دسته بندها که تنها یک خروجی مشخص برای هر ورودی تولید می‌کنند، دسته بندهای گسسته گفته می‌شود که این دسته بندها تنها یک نقطه در فضای ROC تولید می‌کنند.
بطور مشابه دسته بندهای دیگری نظیر دسته بندهای مبتنی بر روش بیز و یا شبکه‌های عصبی نیز وجود دارند که یک احتمال و یا امتیاز برای هر ورودی تولید می‌کنند، که این عدد بیانگر درجه تعلق ورودی به یکی از دو دسته موجود می‌باشد. این دسته بندها پیوسته نامیده می‌شوند و بدلیل خروجی خاص این دسته بندها یک آستانه جهت تعیین خروجی نهایی در نظر گرفته می‌شود.

یک منحنی ROC اجازه مقایسه تصویری مجموعه ای از دسته بندی کننده‌ها را می‌دهد، همچنین نقاط متعددی در فضای ROC قابل توجه است. نقطه پایین سمت چپ (0,0) استراتژی را نشان می‌دهد که در یک دسته بند مثبت تولید نمی‌شود. استراتژی مخالف،

که بدون شرط دسته بندهای مثبت تولید می‌کند، با نقطه بالا سمت راست (1,1) مشخص می‌شود. نقطه (0,1) دسته بندی کامل و بی عیب را نمایش می‌دهد. بطور کلی یک نقطه در فضای ROC بهتر از دیگری است اگر در شمال غربی‌تر این فضا قرار گرفته باشد. همچنین در نظر داشته باشید منحنی‌های ROC رفتار یک دسته بندی کننده را بدون توجه به توزیع دسته‌ها یا هزینه خطا نشان می‌دهند، بنابراین کارایی دسته بندی را از این عوامل جدا می‌کنند. فقط زمانی که یک دسته بند در کل فضای کارایی به وضوح بر دسته دیگری تسلط یابد، می‌توان گفت که بهتر از دیگری است. به همین دلیل معیار AUC که سطح زیر نمودار ROC را نشان می‌دهد می‌تواند نقش تعیین کننده ای در معرفی دسته بند برتر ایفا کند. برای درک بهتر نمودار ROC زیر را مشاهده کنید.



مقدار AUC برای یک دسته بند که بطور تصادفی، دسته نمونه مورد بررسی را تعیین می‌کند برابر 0.5 است. همچنین بیشترین مقدار این معیار برابر یک بوده و برای وضعیتی رخ می‌دهد که دسته بند ایده آل بوده و بتواند کلیه نمونه‌های مثبت را بدون هرگونه هشدار غلطی تشخیص دهد. معیار AUC برخلاف دیگر معیارهای تعیین کارایی دسته بندها مستقل از آستانه تصمیم گیری دسته بند می‌باشد. بنابراین این معیار نشان دهنده میزان قابل اعتماد بودن خروجی یک دسته بند مشخص به ازای مجموعه داده‌های متفاوت است که این مفهوم توسط سایر معیارهای ارزیابی کارایی دسته بندها قابل محاسبه نمی‌باشد. در برخی از مواقع سطح زیر منحنی‌های ROC مربوط به دو دسته بند با یکدیگر برابر است ولی ارزش آنها برای کاربردهای مختلف یکسان نیست که باید در نظر داشت در این گونه مسائل که ارزش دسته‌ها با یکدیگر برابر نیست، استفاده از معیار AUC مطلوب نمی‌باشد. به همین دلیل در این گونه مسائل استفاده از معیار دیگری به جزء هزینه (Cost Matrix) منطقی به نظر نمی‌رسد. در انتها باید توجه نمود در کنار معیارهای بررسی شده که همگی به نوعی دقت دسته بند را محاسبه می‌کردند، در دسته بندهای قابل تفسیر نظیر دسته بندهای مبتنی بر قانون و یا درخت تصمیم، پیچیدگی نهایی و قابل تفسیر بودن مدل یاد گرفته شده نیز از اهمیت بالایی برخوردار است.

از روش‌های ارزیابی الگوریتم‌های دسته بندی (که در این الگوریتم روال کاری بدین صورت است که مدل دسته بندی توسط مجموعه داده آموزشی ساخته شده و بوسیله مجموعه داده آزمایشی مورد ارزیابی قرار می‌گیرد.) می‌توان به روش **Holdout** اشاره کرد که در این روش چگونگی نسبت تقسیم مجموعه داده‌ها (به دو مجموعه داده آموزشی و مجموعه داده آزمایشی) بستگی به تشخیص تحلیلگر دارد که معمولاً دو سوم برای آموزش و یک سوم برای ارزیابی در نظر گرفته می‌شود. مهمترین مزیت این روش

سادگی و سرعت بالای عملیات ارزیابی است ولیکن روش Holdout معایب زیادی دارد از جمله اینکه مجموعه داده‌های آموزشی و آزمایشی به یکدیگر وابسته خواهند شد، در واقع بخشی از مجموعه داده اولیه که برای آزمایش جدا می‌شود، شانسی برای حضور یافتن در مرحله آموزش ندارد و بطور مشابه در صورت انتخاب یک رکورد برای آموزش دیگر شانسی برای استفاده از این رکورد برای ارزیابی مدل ساخته شده وجود نخواهد داشت. همچنین مدل ساخته شده بستگی فراوانی به چگونگی تقسیم مجموعه داده اولیه به مجموعه داده‌های آموزشی و آزمایشی دارد. چنانچه روش Holdout را چندین بار اجرا کنیم و از نتایج حاصل میانگین گیری کنیم از روشی موسوم به **Random Sub-sampling** استفاده نموده ایم. که مهمترین عیب این روش نیز عدم کنترل بر روی تعداد دفعاتی که یک رکورد به عنوان نمونه آموزشی و یا نمونه آزمایشی مورد استفاده قرار می‌گیرد، است. به بیان دیگر در این روش ممکن است برخی رکوردها بیش از سایرین برای یادگیری و یا ارزیابی مورد استفاده قرار گیرند.

چنانچه در روش Random Sub-sampling به شکل هوشمندانه‌تری عمل کنیم به صورتی که هر کدام از رکوردها به تعداد مساوی برای یادگیری و تنها یکبار برای ارزیابی استفاده شوند، روش مزبور در متون علمی با نام **Cross Validation** شناخته می‌شود.

همچنین در روش جامع **k-Fold Cross Validation** کل مجموعه داده‌ها به k قسمت مساوی تقسیم می‌شوند. از k-1 قسمت به عنوان مجموعه داده‌های آموزشی استفاده می‌شود و براساس آن مدل ساخته می‌شود و با یک قسمت باقی مانده عملیات ارزیابی انجام می‌شود. فرآیند مزبور به تعداد k مرتبه تکرار خواهد شد، به گونه ای که از هر کدام از k قسمت تنها یکبار برای ارزیابی استفاده شده و در هر مرتبه یک دقت برای مدل ساخته شده، محاسبه می‌شود. در این روش ارزیابی دقت نهایی دسته بند برابر با میانگین k دقت محاسبه شده خواهد بود. معمول‌ترین مقداری که در متون علمی برای k در نظر گرفته می‌شود برابر با 10 می‌باشد. بدیهی است هر چه مقدار k بزرگتر شود، دقت محاسبه شده برای دسته بند قابل اعتمادتر بوده و دانش حاصل شده جامع‌تر خواهد بود و البته افزایش زمان ارزیابی دسته بند نیز مهمترین مشکل آن می‌باشد. حداکثر مقدار k برابر با تعداد رکوردهای مجموعه داده اولیه است که این روش ارزیابی با نام **Leaving One Out** شناخته می‌شود.

در روش هایی که تاکنون به آن اشاره شده، فرض بر آن است که عملیات انتخاب نمونه‌های آموزشی بدون جایگذاری صورت می‌گیرد. به بیان دیگر یک رکورد تنها یکبار در یک فرآیند آموزشی مورد توجه واقع می‌شود. چنانچه هر رکورد در صورت انتخاب شدن برای شرکت در عملیات یادگیری مدل بتواند مجدداً برای یادگیری مورد استفاده قرار گیرد روش مزبور با نام **Bootstrap** و یا Bootstrap 0.632 شناخته می‌شود. (از آنجا که هر Bootstrap معادل 0.632 مجموعه داده اولیه است)

ارزیابی در الگوریتم‌های خوشه بندی

به منظور ارزیابی الگوریتم‌های خوشه بندی می‌توان آنها به دو دسته تقسیم نمود:

شاخص‌های **ارزیابی بدون ناظر،** که گاهی در متون علمی با نام معیارهای داخلی شناخته می‌شوند، به آن دسته از معیارهایی گفته می‌شود که تعیین کیفیت عملیات خوشه بندی را با توجه به اطلاعات موجود در مجموعه داده بر عهده دارند. در مقابل، معیارهای ارزیابی با ناظر با نام معیارهای خارجی نیز شناخته می‌شوند، که با استفاده از اطلاعاتی خارج از حیطه مجموعه داده‌های مورد بررسی، عملکرد الگوریتم‌های خوشه بندی را مورد ارزیابی قرار می‌دهند.

از آنجا که مهمترین وظیفه یک الگوریتم خوشه بندی آن است که بتواند به بهترین شکل ممکن فاصله درون خوشه ای را کمینه و فاصله بین خوشه ای را بیشینه نماید، کلیه معیارهای ارزیابی بدون ناظر سعی در سنجش کیفیت عملیات خوشه بندی با توجه به دو فاکتور تراکم خوشه ای و جدائی خوشه ای دارند. برآورده شدن هدف کمینه سازی درون خوشه ای و بیشینه سازی میان خوشه ای به ترتیب در گرو بیشینه نمودن تراکم هر خوشه و نیز بیشینه سازی جدایی میان خوشه‌ها می‌باشد. طیف وسیعی از معیارهای ارزیابی بدون ناظر وجود دارد که همگی در ابتدا تعریفی برای فاکتورهای تراکم و جدائی ارائه می‌دهند سپس توسط تابع F(Cohesion, Separation) مرتبط با خود، به ترکیب این دو فاکتور می‌پردازند. ذکر این نکته ضروری است که نمی‌توان هیچ کدام از معیارهای ارزیابی خوشه بندی را برای تمامی کاربردها مناسب دانست.

ارزیابی با ناظر الگوریتم‌های خوشه بندی، با هدف آزمایش و مقایسه عملکرد روش‌های خوشه بندی با توجه به حقایق مربوط به رکوردها صورت می‌پذیرد. به بیان دیگر هنگامی که اطلاعاتی از برچسب رکوردهای مجموعه داده مورد بررسی در اختیار داشته باشیم، می‌توانیم از آنها در عملیات ارزیابی عملکرد الگوریتم‌های خوشه بندی بهره بریم. لازم است در نظر داشته باشید در این بخش از برچسب رکوردها تنها در مرحله ارزیابی استفاده می‌شود و هر گونه بهره برداری از این برچسب‌ها در مرحله یادگیری مدل، منجر به تبدیل شدن روش کاوش داده از خوشه بندی به دسته بندی خواهد شد. مشابه با روش‌های بدون ناظر طیف وسیعی از معیارهای ارزیابی با ناظر نیز وجود دارد که در این قسمت با استفاده از روابط زیر به محاسبه معیارهای **Rand Index** و **Jaccard** می پردازیم به ترتیب در **رابطه I و II** نحوه محاسبه آنها نمایش داده شده است:

$$I: \quad RI = \frac{TP + TN}{TP + FP + TN + FN}$$

$$II: \quad Jaccard = \frac{TP}{TP + FP + TN}$$

Rand Index را می‌توان به عنوان تعداد تصمیمات درست در خوشه بندی در نظر گرفت.
TP : به تعداد زوج داده هایی گفته می‌شود که باید در یک خوشه قرار می‌گرفتند، و قرار گرفته اند.
TN : به تعداد زوج داده هایی گفته می‌شود که باید در خوشه‌های جداگانه قرار داده می‌شدند و به درستی در خوشه‌های جداگانه جای داده شده اند.
FN : به تعداد زوج داده هایی گفته می‌شود که باید در یک خوشه قرار می‌گرفتند ولی در خوشه‌های جداگانه قرار داده شده اند.
FP : به تعداد زوج داده هایی اشاره دارد که باید در خوشه‌های متفاوت قرار می‌گرفتند ولی در یک خوشه قرار گرفته اند.

ارزیابی در الگوریتم‌های کشف قوانین انجمنی

به منظور ارزیابی الگوریتم‌های کشف قوانین انجمنی از آنجایی که این الگوریتم‌ها پتانسیل این را دارند که الگوها و قوانین زیادی تولید نمایند، جهت ارزیابی این قوانین به عواملی همچون شخص استفاده کننده از قوانین و نیز حوزه ای که مجموعه داده مورد بررسی به آن تعلق دارد، وابستگی زیادی پیدا می‌کنیم و بدین ترتیب کار پیدا کردن قوانین جذاب، به آسانی میسر نیست. فرض کنید قانونی با نام R داریم که به شکل **A=>B** می‌باشد، که در آن A و B زیر مجموعه ای از اشیاء می‌باشند.

پیشتر به معرفی دو معیار Support و Confidence پرداختیم. می‌دانیم از نسبت تعداد تراکنش هایی که در آن اشیاء A و B هر دو حضور دارند، به کل تعداد رکوردها **Support** بدست می‌آید که دارای مقداری عددی بین صفر و یک می‌باشد و هر چه این میزان بیشتر باشد، نشان می‌دهد که این دو شیء بیشتر با هم در ارتباط هستند. کاربر می‌تواند با مشخص کردن یک آستانه برای این معیار، تنها قوانینی را بدست آورد که Support آنها بیشتر از مقدار آستانه باشد، بدین ترتیب می‌توان با کاهش فضای جستجو، زمان لازم جهت پیدا کردن قوانین انجمنی را کمینه کرد. البته باید به ضعف این روش نیز توجه داشت که ممکن است قوانین با ارزشی را بدین ترتیب از دست دهیم. در واقع استفاده از این معیار به تنهایی کافی نیست. معیار **Confidence** نیز مقداری عددی بین صفر و یک می‌باشد، که هر چه این عدد بزرگتر باشد بر کیفیت قانون افزوده خواهد شد. استفاده از این معیار به همراه Support مکمل مناسبی برای ارزیابی قوانین انجمنی خواهد بود. ولی مشکلی که همچنان وجود دارد این است که امکان دارد قانونی با Confidence بالا وجود داشته باشد ولی از نظر ما ارزشمند نباشد.

از معیارهای دیگر قوانین انجمنی می‌توان به معیار **Lift** که با نام‌های Intersect Factor یا Interestingness نیز شناخته می‌شود اشاره کرد، که این معیار میزان استقلال میان اشیاء A و B را نشان می‌دهد که می‌تواند مقدار عددی بین صفر تا بی نهایت باشد. در واقع Lift میزان هم اتفاقی بین ویژگی‌ها را در نظر می‌گیرد و میزان رخداد تکی بخش تالی قانون (یعنی شیء B) را در محاسبات خود وارد می‌کند. (بر خلاف معیار Confidence)

مقادیر نزدیک به عدد یک معرف این هستند که A و B مستقل از یکدیگر می‌باشند، بدین ترتیب نشان دهنده قانون جذابی نمی‌باشند. چنانچه این معیار از عدد یک کمتر باشد، نشان دهنده این است که A و B با یکدیگر رابطه منفی دارند. هر چه مقدار این معیار بیشتر از عدد یک باشد، نشان دهنده این است که A اطلاعات بیشتری درباره B فراهم می‌کند که در این حالت جذابیت قانون A=>B بالاتر ارزیابی می‌شود. در ضمن این معیار نسبت به سمت چپ و راست قانون متقارن است در واقع اگر سمت چپ و راست قانون را با یکدیگر جابجا کنیم، مقدار این معیار تغییری نمی‌کند. از آنجائی که این معیار نمی‌تواند به تنهایی برای ارزیابی مورد استفاده قرار گیرد، و حتماً باید در کنار معیارهای دیگر باشد، باید مقادیر آن بین بازه صفر و یک نرمال شود. ترکیب این معیار به همراه Support و Confidence جزو بهترین روش‌های کاوش قوانین انجمنی است. مشکل این معیار حساس بودن به تعداد نمونه‌های مجموعه داده، به ویژه برای مجموعه تراکنش‌های کوچک می‌باشد. از این رو معیارهای دیگری برای جبران این نقص معرفی شده اند.

معیار **Conviction** برخی ضعف‌های معیارهای Confidence و Lift را جبران می‌نماید. محدوده قابل تعریف برای این معیار در حوزه 0.5 تا بی نهایت قرار می‌گیرد که هر چه این مقدار بیشتر باشد، نشان دهنده این است که آن قانون جذاب‌تر می‌باشد. بر خلاف Lift این معیار متقارن نمی‌باشد و مقدار این معیار برای دلالت‌های منطقی یعنی در جایی که Confidence قانون یک می‌باشد برابر با بی نهایت است و چنانچه A و B مستقل از هم باشند، مقدار این معیار برابر با عدد یک خواهد بود.

$$Conf\ (A \rightarrow B) = \frac{SUP(A \cup B)}{SUP(A)}$$

$$Lift\ (A \rightarrow B) = \frac{Conf(A \rightarrow B)}{SUP(B)}$$

$$Conv\ (A \rightarrow B) = \frac{1 - SUP(B)}{1 - Conf(A \rightarrow B)}$$

معیار **Leverage** که در برخی متون با نام Novelty (جدید بودن) نیز شناخته می‌شود، دارای مقداری بین 0.25- و 0.25+ می‌باشد. ایده مستتر در این معیار آن است که اختلاف بین میزان هم اتفاقی سمت چپ و راست قانون با آن مقداری که مورد انتظار است به چه اندازه می‌باشد.

معیار **Jaccard** که دارای مقداری عددی بین صفر و یک است، علاوه بر اینکه نشان دهنده وجود نداشتن استقلال آماری میان A و B می‌باشد، درجه همپوشانی میان نمونه‌های پوشش داده شده توسط هر کدام از آنها را نیز اندازه گیری می‌کند. به بیان دیگر این معیار فاصله بین سمت چپ و راست قانون را بوسیله تقسیم تعداد نمونه هایی که توسط هر دو قسمت پوشش داده شده اند بر نمونه هایی که توسط یکی از آنها پوشش داده شده است، محاسبه می‌کند. مقادیر بالای این معیار نشان دهنده این است که A و B تمایل دارند، نمونه‌های مشابهی را پوشش دهند. لازم است به این نکته اشاره شود از این معیار برای فهمیدن میزان همبستگی میان متغیرها استفاده می‌شود که از آن می‌توان برای یافتن قوانینی که دارای همبستگی بالا ولی Support کم هستند، استفاده نمود. برای نمونه در مجموعه داده سبد خرید، قوانین نادری که Support کمی دارند ولی همبستگی بالایی دارند، توسط این معیار می‌توانند کشف شوند.

$$Leve\ (A \rightarrow B) = SUP(A \cup B) - SUP(A) \times SUP(B)$$

$$Jaac\ (A \rightarrow B) = \frac{SUP(A \cup B)}{SUP(A) + SUP(B) - SUP(A \cup B)}$$

معیار **Coefficient (ϕ)** نیز به منظور اندازه گیری رابطه میان A و B مورد استفاده قرار می‌گیرد که محدوده این معیار بین 1- و 1+ می‌باشد.
از دیگر معیارهای ارزیابی کیفیت قوانین انجمنی، طول قوانین بدست آمده می‌باشد. به بیان دیگر با ثابت در نظر گرفتن معیارهای دیگر نظیر Support، Confidence و Lift قانونی برتر است که طول آن کوتاه‌تر باشد، بدلیل فهم آسانتر آن.

$$\Phi\ (A \rightarrow B) = \frac{Leve\ (A \rightarrow B)}{\sqrt{SUP(A) \times SUP(B) \times (1 - SUP(A)) \times (1 - SUP(B))}}$$

در نهایت با استفاده از **ماتریس وابستگی** (Dependency Matrix)، می‌توان اقدام به تعریف معیارهای متنوع ارزیابی روش‌های تولید قوانین انجمنی پرداخت. در عمل معیارهای متعددی برای ارزیابی مجموعه قوانین بدست آمده وجود دارد و لازم است با توجه به تجارب گذشته در مورد میزان مطلوب بودن آنها تصمیم گیری شود. بدین ترتیب که ابتدا معیارهای برتر در مسئله مورد کاوش پس از مشورت با خبرگان حوزه شناسائی شوند، پس از آن قوانین انجمنی بدست آمده از حوزه کاوش، مورد ارزیابی قرار گیرند.

نظرات خوانندگان

نویسنده: محمد باقر سیف اللهی
تاریخ: ۱۳۹۳/۰۹/۱۲ ۲۲:۳۳

خسته نباشید میگم. مطالب مفیدی است ولی نکته ای رو لازم می‌دونم بیان کنم و آن، اینکه مفاهیم بیان شده در حد مطالعه خوب است ولی بدون نمونه و مثال، مفاهیم بسیار سختی دارند. برای مثال روش تهیه ROC از روی جدول اطلاعاتی مرتبط با TP - TN - FN وFP و TP Rate و FP rate مشکل است خصوصا با داده‌های زیاد. و یا روش‌های Rule Pruning و Rule Growing در رده بندی و امثالهم. (که البته مفاهیم بسیار سنگین‌تری نیز وجود دارند)
به همین دلیل پیشنهاد می‌کنم در صورت امکان، در انتهای آموزش خود، تا جاییکه امکان دارد با مثال و شکل این موارد بیان شوند.

همچنین ابزارهای دیگری به غیر از SQL Server نیز مرور شوند (و یا صرفا معرفی شوند) تا با رویکرد عملیاتی ساختن و استفاده کاربردی از داده کاوی، بتوان از آنها بهره برد
(موردی بود که بنده مجبور بودم در یک پروسه، از 3 ابزار برای کارهای مختلف استفاده کنم) اسلاید هایی هم در این زمینه‌ها وجود دارند که برای شروع مناسب هستند + و +
موفق باشید

نویسنده: محمد رجبی
تاریخ: ۱۳۹۳/۰۹/۱۴ ۱۰:۴۲

با سلام و احترام، ضمن تشکر از Feedback ای که ارسال نمودید، حقیقتاً در ابتدای امر چنین قصدی داشتم ولی با مشورت دوستانم بر آن شدم، که مشخصاً به بیان مباحث تئوری موضوع بپردازم. از آنجا که به نظر می‌رسد بر خلاف رویه ماکروسافت که معمولاً مفاهیم را در مجموعه‌های آموزشی از مباحث پایه و مقدماتی شروع می‌کند و تا سطح پیشرفته؛ با جزئیات کامل به بیان موضوع می‌پردازد. متاسفانه در بحث داده کاوی چنین رویه ای را در پیش نگرفته و فرض را بر آن گذاشته است که خواننده با مفاهیم کلی علم داده کاوی آشناست و با این پیش فرض به بررسی الگوریتم‌ها و نحوه استفاده از آنها می‌پردازد. از این رو تصمیم گرفتم بیشتر خلاصه مباحث تئوری را بیان کنم و از آنجایی که به منظور انجام عملیات داده کاوی در گام نخستین شخص داده کاو می‌بایست از داده‌های مورد کاوش، شناخت و آگاهی کافی داشته باشد و همانطور که می‌دانیم، جهت اهداف آموزشی بانک اطلاعاتی Adventure Works (که حاوی اطلاعات حوزه‌های متفاوت در یک کمپانی می‌باشد) و بانک Adventure Works DW (که در واقع انبار داده حوزه فروش بانک Adventure Works است)، موجود می‌باشد. کلیه مثال‌های موجود در Books Online که برای هر الگوریتم و زمینه کاری متناظر با آن ارائه شده است روی این بانک‌های اطلاعاتی انجام می‌گیرد.
لینک زیر دانلود مجموعه آموزش SQL Server 2012 Tutorials - Analysis Services Data Mining می‌باشد. که شامل موارد زیر است:

Basic Data Mining Tutorial

```
Lesson 1: Preparing the Analysis Services Database

Lesson 2: Building a Targeted Mailing Structure

Lesson 3: Adding and Processing Models

Lesson 4: Exploring the Targeted Mailing Models

Lesson 5: Testing Models

Lesson 6: Creating and Working with Predictions
```

Intermediate Data Mining Tutorial

```
Lesson 1: Creating the Intermediate Data Mining Solution
```

```
Lesson 2: Building a Forecasting Scenario
```

```
Lesson 3: Building a Market Basket Scenario
```

```
Lesson 4: Building a Sequence Clustering Scenario
```

```
Lesson 5: Building Neural Network and Logistic Regression Models
```

Creating and Querying Data Mining Models with DMX: Tutorials

```
Lesson 1: Bike Buyer
```

```
Lesson 2: Market Basket
```

```
Lesson 3: Time Series Prediction
```

بدین ترتیب برای مخاطبان این دوره که ممکن است آشنائی با مفاهیم تئوری علم داده کاوی نداشته باشند، سعی شده است مطالب به گونه ای بیان شود که با مطالعه این مجموعه، سر نخ هایی از موضوع بدست آورند و طبیعتاً در صورت علاقه مندی به موضوع به مطالعه عمیق هر الگوریتم بپردازند.

از آنجا که با سونامی «تحصیلات تکمیلی» در کشور مواجه هستیم و بسیاری از پایان نامه‌ها پیرامون موضوع Data Mining می‌باشد و همچنین مشابه بسیاری از موضوعات دیگر؛ بدون در نظر گرفتن زیر ساخت‌ها و فلسفه پیدایش موضوع و دستاوردهای آن و ... پروژه‌های داده کاوی نیز به صورت وارداتی به کشور و به طبع سازمان‌ها تحمیل می‌شود و ... امیدوارم توانسته باشم، هم زبانان نا آشنا را تا حدی که در توان داشتم با موضوع آشنا کرده باشم. برای مطالعه منابع غیر از SQL Server کتاب‌های « داده کاوی کاربردی - RapidMiner » انتشارات نیاز دانش و همچنین کتاب « داده کاوی با کلمنتاین » انتشارات جهاد دانشگاهی واحد صنعتی امیر کبیر نیز به بیان موضوع می‌پردازد. موفق و سلامت باشید.