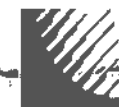
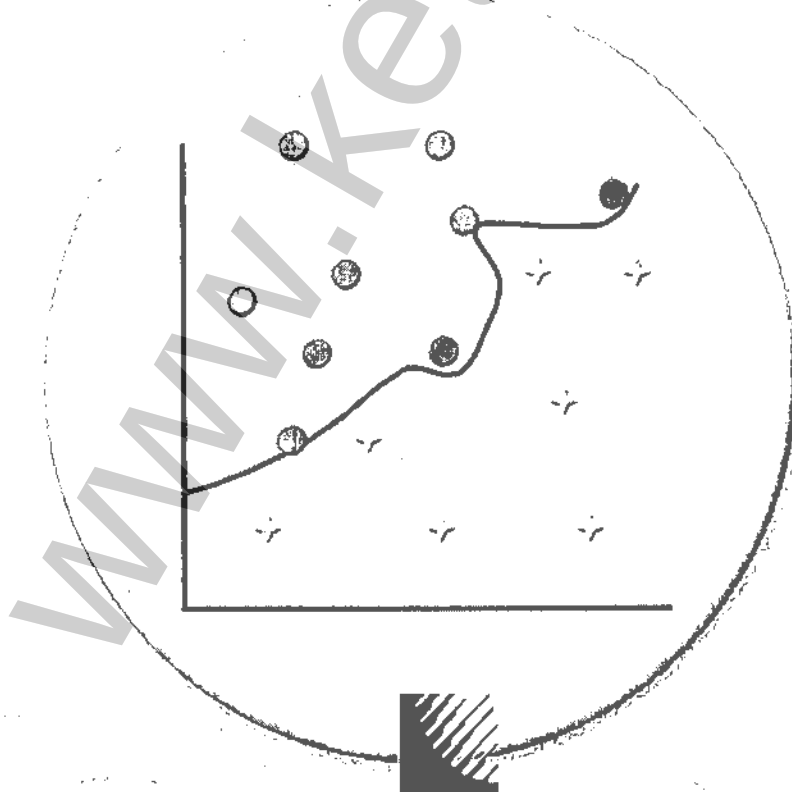


داده کاوی

(Data Mining)

مهندس امیر علیخانزاده
(عضو هیات علمی دانشگاه فردوسی مشهد)

مهمد کانتاردزیک



علوم رایانه

کانتاردزیک ، مهمد -

داده کاوی (Data Mining) / تألیف : مهمد کانتاردزیک ؛ ترجمه: امیر علیخانزاده.

بابل : نشر علوم رایانه ، ۱۳۸۵ .

۳۴۴ ص . : مصور ، جدول.

ISBN : 964-8996-13-X

فهرست نویسی براساس اطلاعات فیبا (فهرست نویسی پیش از انتشار).

۱. داده کاوی - شبکه های عصبی . الف. علیخانزاده، امیر، ۱۳۳۵ - مترجم. ب. عنوان

۰۰۵/۷۳

د ک ۸ ۷۶/۹/۸ QA



www.olomrayaneh.net

تلفن ۰۱۱۱-۳۲۶۰۷۷۲

بابل، صندوق پستی ۸۹۱-۴۷۱۳۵

علوم رایانه

داده کاوی (Data Mining)

تألیف : مهمد کانتاردزیک

ترجمه : امیر علیخانزاده

ناشر : علوم رایانه

چاپ اول : تابستان ۱۳۸۵

جلد ۲۰۰۰

شابک: ۱۳-۸۹۹۶-۹۶۴-X

قیمت: ۴۰۰۰ تومان

حروفچینی و صفحه‌آرایی : علوم رایانه

بسم الله الرحمن الرحيم

یکی از پایه‌های توسعه فرهنگی و رشد و شکوفایی
استعدادها، مقوله کتاب و کتابخوانی است. نیاز به کتاب به سرعت
در حال رشد است و انتشار نتیجه مطالعات پژوهشگران و
اندیشمندان پاسخگوی این نیاز خواهد بود.

جهت تحقق این امر و گام برداشتن به سمت خودکفایی،
چاپ کتب علمی، به‌خصوص در زمینه تکنولوژی جدید، یعنی
کامپیوتر، که توسعه سایر فناوری‌ها نیز به آن وابسته است،
ضروری است.

عین‌اله جعفر نژاد قمی
انتشارات علوم رایانه

فهرست مطالب

۱- مفاهیم داده‌کاوی

۱-۱	مقدمه	۹
۱-۲	ریشه‌های داده‌کاوی	۱۲
۱-۳	فرآیند داده‌کاوی	۱۴
۱-۴	مجموعه داده‌های بزرگ	۱۸
۱-۵	انبارهای داده	۲۲
۱-۶	ساختار این کتاب	۲۶
۱-۷	مرور سئوالات و مسائل	۲۷
۱-۸	منابع برای مطالعه	۲۸

۲- آماده‌سازی داده‌ها

۲-۱	نمایش داده‌های خام	۳۱
۲-۲	ویژگی‌های داده‌های اولیه (خام)	۳۵
۲-۳	تبدیل (تغییر وضعیت) داده‌های خام	۳۷
۲-۴	داده‌های از دست رفته	۴۰
۲-۵	داده‌های وابسته به زمان	۴۲
۲-۶	تحلیل داده‌های نامنتطبق	۴۷
۲-۷	مرور سئوالات و مسائل	۵۰
۲-۸	منابع برای مطالعه	۵۲

۳- کاهش داده‌ها

۳-۱	ابعاد مجموعه داده‌های بزرگ	۵۴
۳-۲	کاهش ویژگی‌ها	۵۶
۳-۳	سنجش آنتروپی برای رتبه‌بندی ویژگی‌ها	۶۲
۳-۴	تجزیه و تحلیل عنصر اصلی	۶۵
۳-۵	کاهش مقادیر	۶۷
۳-۶	گسسته‌سازی ویژگی: تکنیک ChiMerge	۷۱
۳-۷	کاهش موارد (حالات)	۷۴
۳-۸	مرور سئوالات و مسائل	۷۸
۳-۹	منابع برای مطالعه	۷۹

۴- یادگیری از داده‌ها

۴-۱	ماشین یادگیری	۸۳
۴-۲	تئوری یادگیری آماری	۸۸
۴-۳	انواع روش‌های یادگیری	۹۴
۴-۴	اعمال مشترک یادگیری	۹۶
۴-۵	برآورد مدل	۱۰۲
۴-۶	مرور سئوالات و مسائل	۱۰۷
۴-۷	منابع برای مطالعه	۱۰۸

۵- روش‌های آماری

۵-۱	استنتاج آماری	۱۰۹
۵-۲	تشخیص تفاوت‌ها در مجموعه داده	۱۱۱
۵-۳	استنتاج Bayesian	۱۱۳
۵-۴	رگرسیون پیشگرو	۱۱۶
۵-۵	تحلیل واریانس	۱۲۲
۵-۶	رگرسیون Logistic	۱۲۵
۵-۷	مدل Log خطی	۱۲۶
۵-۸	تحلیل ممیز خطی	۱۳۱
۵-۹	مرور سئوالات و مسائل	۱۳۲
۵-۱۰	منابع برای مطالعه	۱۳۳

۶- تجزیه و تحلیل دسته‌ها

۶-۱	مفاهیم دسته‌بندی	۱۳۵
۶-۲	سنجش‌ها و معیارهای تشابه	۱۳۹
۶-۳	دسته‌بندی سلسله مراتبی متراکم	۱۴۳
۶-۴	دسته‌بندی جزئی	۱۴۷
۶-۵	دسته‌بندی افزایشی	۱۵۱
۶-۶	مرور سئوالات و مسائل	۱۵۵
۶-۷	منابع برای مطالعه	۱۵۷

۲۵۲	۹-۸. منابع برای مطالعه	۱۵۹	۷-درختان تصمیم و قوانین تصمیم
		۱۶۱	۷-۱. درخت تصمیم
۲۵۵	۱۰-۱. الگوریتم‌های ژنتیک	۱۶۲	۷-۲. الگوریتم C4.5: تولید درخت تصمیم
۲۵۶	۱۰-۱. اصول الگوریتم‌های ژنتیک	۱۷۰	۷-۳. مقادیر ویژگی ناشناخته
	۱۰-۲. بهینه‌سازی با استفاده از الگوریتم‌های ژنتیک	۱۷۴	۷-۴. هرس کردن درخت تصمیم
۲۵۹	۱۰-۳. نمایش ساده‌ای برای یک الگوریتم ژنتیک	۱۷۶	۷-۵. الگوریتم C4.5: تولید قوانین تصمیم
۲۶۵	۱۰-۴. الگوریتم‌های ژنتیک	۱۷۶	۷-۶. محدودیت‌های درختان تصمیم و قوانین تصمیم
۲۷۱	۱۰-۵. مسئله فروشنده دوره‌گرد	۱۷۹	۷-۷. روش طبقه‌بندی شرکت‌پذیری (انجمنی)
۲۷۲	۱۰-۶. یادگیری ماشین با استفاده از الگوریتم‌های ژنتیک	۱۸۱	۷-۸. مرور سئوالات و مسائل
۲۷۶	۱۰-۷. مرور سئوالات و مسائل	۱۸۶	۷-۹. منابع برای مطالعه
۲۸۱	۱۰-۸. منابع برای مطالعه		۸-قوانین انجمنی
۲۸۲		۱۸۹	۸-۱. تحلیل سید خرید
	۱۱-۱. مجموعه‌های فازی و منطق فازی	۱۹۰	۸-۲. الگوریتم APRIORI
۲۸۶	۱۱-۲. اعمال مجموعه فازی	۱۹۱	۸-۳. مجموعه اقلام‌های تکراری و روابط انجمنی
۲۹۱	۱۱-۳. اصل توسعه و روابط فازی	۱۹۴	۸-۴. افزایش راندمان و کارایی الگوریتم Apriori
۲۹۷	۱۱-۴. منطق فازی و سیستم‌های استنباط فازی	۱۹۵	۸-۵. روش الگوی افزایش تکراری
۳۰۱	۱۱-۵. ارزیابی چندفاکتوری	۱۹۷	۸-۶. کاوش قوانین انجمنی چندبعدی
۳۰۵	۱۱-۶. توسعه مدل‌های فازی از داده‌ها	۲۰۰	۸-۷. کاوش وب (وب کاری)
۳۰۸	۱۱-۷. مرور سئوالات و مسائل	۲۰۲	۸-۸. الگوریتم‌های هیتس و لاگسام
۳۱۲	۱۱-۸. منابع برای مطالعه	۲۰۵	۸-۹. کاوش الگوهای پیمایش مسیر
۳۱۵		۲۱۲	۸-۱۰. کاوش متن (متن کاری)
	۱۲-۱. ادراک و تجسم‌سازی فکری	۲۱۶	۸-۱۱. مرور سئوالات و مسائل
۳۱۷	۱۲-۲. تجسم‌سازی عصبی و تجسم‌سازی اطلاعات	۲۲۱	۸-۱۲. منابع برای مطالعه
۳۱۸	۱۲-۳. مختصات موازی	۲۲۳	
۳۱۹	۱۲-۴. تجسم‌سازی شعاعی		۹-شبکه‌های عصبی و مصنوعی
۳۲۶	۱۲-۵. نگاشت‌های خردسازماندهی کوهن	۲۲۷	۹-۱. مدل یک نورون مصنوعی
۳۲۹	۱۲-۶. سیستم‌های تجسم‌سازی برای داده‌کاو	۲۳۰	۹-۲. معماری‌های شبکه‌های عصبی مصنوعی
۳۳۱	۱۲-۷. مرور سئوالات و مسائل	۲۳۲	۹-۳. فرآیند یادگیری
۳۳۲	۱۲-۸. منابع برای مطالعه	۲۳۶	۹-۴. وظایف یادگیری
۳۳۸		۲۳۹	۹-۵. مفاهیم چندلایه‌ای
۳۳۹		۲۴۶	۹-۶. شبکه‌های رقابتی و یادگیری رقابتی
		۲۵۰	۹-۷. مرور سئوالات و مسائل

پیشگفتار

به صورت سنتی، تحلیل‌گران فرآیند استخراج اطلاعات مفید از داده‌های ذخیره شده را اجرا می‌کنند اما افزایش حجم داده‌ها در علوم و تجارت مدرن، به نام شیوه‌های مبتنی بر کامپیوتر نامیده می‌شود. مادامی که اندازه و پیچیدگی مجموعه داده‌ها رشد می‌کند، یک تغییر و جابه‌جایی غیرقابل تصویری از تحلیل داده‌های دستی به سمت تحلیل داده‌های خودکار و تا حدودی غیر مستقیم با استفاده از ابزار پیچیده و پیشرفته حاصل می‌شود. در حقیقت، فناوری‌های جدید کامپیوتر، شبکه و سنسورها، جمع‌آوری داده‌ها و سازماندهی آنها را تقریباً به یک امر بدون تلاش و زحمت تبدیل نموده‌اند. هرچند داده‌های به دست آمده برای آنکه مفید باشند، باید با استفاده از داده‌های ذخیره شده به اطلاعات و دانش مفید تبدیل شود. داده‌کاوی در حقیقت فرآیند کاملی از اعمال متدولوژی مبتنی بر کامپیوتر می‌باشد که شامل تکنیک‌های جدید برای اکتشاف دانش از داده‌ها است.

باید توجه داشت، دنیای مدرن در حقیقت دنیای داده‌گرا می‌باشد و ما در محاصره داده‌ها، چه عددی و چه انواع دیگری قرار گرفته‌ایم. این داده‌ها باید تحلیل و پردازش شوند تا تبدیل به اطلاعاتی گردند که به عنوان ابزاری در جهت اطلاع‌رسانی، آموزش و سایر روش‌ها به درک و تصمیم‌گیری ما کمک کنند. درواقع این دوران، عصر اینترنت، اینترنت انبار داده‌ها و مراکز داده‌ای و بالاخره پارادایم مبانی تحلیل داده‌های کلاسیک به تغییرات مناسب در این مبحث می‌باشد. مجموعه‌های بسیار بزرگی از داده‌ها (گاهی اوقات صدها میلیون رکورد منفرد) در انبارهای داده‌ای متمرکز ذخیره می‌شوند به‌نحوی که این کتاب برای دامنه وسیعی از خوانندگان تدوین شده است. این طیف شامل دانشجویانی است که علاقه‌مند به یادگیری این زمینه از فناوری و فرآیندهای داده‌کاوی می‌باشند و تا تحلیل‌گران و برنامه‌نویسانی که به‌طور مستقیم درگیر انتخاب کاربردهای داده‌کاوی می‌باشند، متفاوت می‌باشد. این کتاب تکنیک‌های کاملاً علمی برای تحلیل تعداد زیادی از فرآیندهای تصمیم‌گیری را مطرح می‌کند، خیلی از تعاریف، طبقه‌بندی‌ها و توضیحات این تکنیک‌های ارائه شده در این کتاب کاملاً جدید نبوده و اطلاعات بیشتری در مراجع انتهایی کتاب ارائه شده است. یکی از اهداف اصلی این کتاب، تمرکز بر شیوه معنایی و مفهومی تمام مراحل یک فرآیند داده‌کاوی داشته و از نمایش نمونه‌های کافی استفاده می‌کنند. این اطمینان وجود دارد که نمونه‌ها و مثال‌های مفید به خواننده این امکان را بدهد که از این قابلیت‌ها برای انتخاب ساختار تکنیک‌ها و ابزاری برای کاربردهای داده‌کاوی استفاده کند. درک بهتر از جزئیات پیاده‌سازی برای اکثر تکنیک‌های معرفی شده موجب می‌شود تا چالش لازم برای ایجاد ابزار موردنیاز خواننده ایجاد شود که این خود باعث بهبود روش‌ها و تکنیک‌های اعمال شده می‌باشد.

برای آموزش داده‌کاوی، باید بر مفاهیم و خواص روش‌های اعمال شده به جای بیان جزئیات عملی ابزار مختلف داده‌کاوی تکیه نمود. برخلاف همه جاذبه‌های ابزار مبتنی بر کامپیوتر، شیوه‌های موجود باید به خیلی از سئوالات پاسخ دهد از جمله اینکه فرآیندهای مربوطه چگونه طراحی می‌شوند و چگونه و به چه شکلی مورد استفاده قرار می‌گیرند. در اینجا یک درک عمیق‌تر از روش‌ها و مدل‌ها به‌دست می‌آید به‌نحوی که این روش‌ها مشخص می‌کنند که چگونه باید عمل کنند و چه پیش‌نیازهایی برای کاربردهای موفقیت‌آمیز این فناوری داده‌کاوی به تحلیل‌گرها اجازه

می‌دهند تا از روش‌های داده‌کاوی قدرتمند و جامعی استفاده کنند. مادامی که حجم و میزان داده‌ها زیاد و در حال رشد باشد، تعداد منابع محدود و بازه زمینه‌های مورد پوشش در بخش‌های صنعتی، تجاری، مالی و اعتباری و علمی کاملاً وسیع می‌باشد. در سال‌های اخیر رشد فزاینده‌ای در روش‌های استخراج دانش جدید از داده‌های اولیه و خام به وجود آمده است. در ارتباط با این موضوع، خط مشی و نظام‌های جدیدی از داده‌کاوی توسعه یافته است مخصوصاً استخراج اطلاعات ارزشمند از مجموعه داده‌های حجیم بزرگ. با توجه به هزینه‌های پایین داده و رو به کاهش استفاده از کامپیوتر، سنسورهای کم هزینه، ارتباطات پر سرعت و فناوری بانک اطلاعاتی (برای جمع‌آوری و ذخیره‌سازی داده‌ها)، استقبال متخصصین و دانشمندان از این ابزار مفید و مورد علاقه، آن‌چنان تعجب‌آور نیست.

در سال‌های اخیر، فناوری داده‌کاوی به یکی از داغ‌ترین موضوعات برای تصمیم‌سازان تبدیل شده است، زیرا فناوری داده‌کاوی یک 'هوش' علمی و تجاری بالارزش و تا حدودی پنهان را برای حجم زیادی از داده‌ها و سوابق یک کاربرد فراهم می‌سازد - هرچند اساساً داده‌کاوی یک فناوری جدید محسوب نمی‌شود. استخراج اطلاعات و دانش از داده‌های ذخیره شده یک مفهوم کاملاً تعریف شده در مباحث علمی و پزشکی است. آن‌چه که جدید است، همگرایی چندین نظام و تکنولوژی است که فرصت کاملاً متمایزی را برای داده‌کاوی در دنیای معاصر ایجاد کرده است.

اصولاً، این کتاب در جهت راهنمایی دانشجویان و علاقه‌مندان در مباحث جدید داده‌کاوی می‌باشد. هرچند، این مسئله قابل تأمل است که افراد با زمینه‌های عملی مختلف و نیز موقعیت‌های شغلی متفاوت به روش‌های گوناگونی از این کتاب بهره‌مند خواهند شد و به‌طور کلی هدف بیان روش‌ها و ابزار مختلف و متداول داده‌کاوی است به‌نحوی که برای اکثر افراد قابل استفاده باشد. بنابراین هر محقق و صاحب‌نظر در این زمینه نیاز دارد تا از این جنبه‌ها به منظور تحقق اهداف خود استفاده کند و در نتیجه روش شناختی و محدودیت‌های موجود را بهتر مورد توجه قرار دهد. در هر صورت این یک تلاش و کوشش قابل توجهی در جهت ارائه و بررسی چنین اصول و جنبه‌ها می‌باشد که در نتیجه منجر به ارائه مباحثی مانند آمار، یادگیری ماشین، گرافیک کامپیوتر، پایگاه داده‌ها، بازیابی اطلاعات، شبکه‌های عصبی، منطق فازی و محاسبات دقیق می‌گردد. در حقیقت در این کتاب، شیوه‌هایی مورد بررسی قرار می‌گیرد که برای الگوهای مهم و مدل‌های مربوطه به مجموعه داده‌های بزرگ حیاتی و مهم هستند.

هرچند، تأکید بر مفاهیم تکنولوژیکی کار ساده‌ای می‌باشد در طول مطالعه این کتاب به این نکته توجه داشته باشید که مفاهیم تکنولوژیکی به تنهایی تمام راه حل نمی‌باشد یکی از اهداف اصلی کتاب حداقل سازی مشکلات مربوط به داده‌کاوی به‌جای قول‌های غیرممکن از آنجه که به صورت معقول مورد انتظار می‌باشد. در حقیقت در اینجا از شیوه‌های هدف‌دار بیشتری استفاده شده است. الگوریتم‌ها و فرآیندهایی که در اینجا مطرح می‌شود، منجر به ارائه نتایج قابل اطمینان و مفیدی در زمینه داده‌کاوی می‌گردد.

نکته قابل توجه در اینجا عدم برتری هر تکنیک خاصی بر تکنیک دیگری می‌باشد. طراحان فرآیندهای داده‌کاوی باید زمینه علمی کافی برای انتخاب ابزارهای نرم‌افزاری و شیوه شناختی مناسبی داشته باشند. انتظار می‌رود که خوانندگان پس از مطالعه این کتاب قادر به اجرای فعالیت‌های اصلی در تمام مراحل فرآیندهای داده‌کاوی به صورت موفقیت‌آمیزی باشند.

مهد کاتاردزیک

لونیس ویلک و

۱ اگوست ۲۰۰۲