داده کاوی

(مفاهیم و تکنیک‌ها)

نویسنده:

ژیاوی هان – میشلین کمبر- ژان پی

مترجم:

دکتر مهدی اسماعیلی

نیاز دانش

پیش گفتار

مترجم

اگر بر جای من غیری گزیند دوست حاکم اوست

حرامم باد اگر من جان بجای دوست بگزینم

داده‌ها یکی از مولفه‌های ارزشمند دنیای امروزی تلقی می‌شود و بدون شک جمع‌آوری و تحلیل داده‌ها برای سازمان‌ها و شرکت‌ها و حتی زندگی روزمره‌ی افراد سودمند است. پیشرفت‌هایی که در رسانه‌ها و ابزارهای ذخیره‌سازی داده‌ها صورت گرفته است، دغدغه‌های ما را در زمینه‌ی جمع‌آوری و نگهداری داده‌ها تا حدودی حل نموده است، اما مشکل چگونگی تحلیل این داده‌ها است. امروزه شما با حجم انبوهی از داده‌ها روبرو هستید و تحلیل حجم بالایی از داده‌ها نیز یک ضرورت است. یکی از علومی که امروزه در دنیا از اهمیت ویژه‌ای برخوردار است و بی‌شک می‌توان پیشرفت روزافزون آن را در بسیاری از کاربردهای علمی و صنعتی مشاهده نمود، علم داده‌کاوی است. داده‌کاوی همچون هر کاوش و تحلیل دیگری به دنبال گنجی از اطلاعات در میان اقیانوسی از داده‌ها است و رشد بی‌رویه‌ی داده‌ها اهمیت آن را دو چندان کرده است.

صاحبنظران، بویژه منبع‌شناسان در زمینه‌ی داده‌کاوی بهتر می‌دانند که کتاب پیش روی شما (منبع اصلی) یکی از ارزنده‌ترین کتاب‌هایی است که در این حوزه به رشته‌ی تحریر درآمده است. از اینرو از میان کتاب‌های فراوانی که در این حوزه وجود دارد، به انتخاب و ترجمه‌ی کتاب مزبور همت گماشتیم. هدف این کتاب، بیان مفاهیم و تکنیک‌های داده‌کاوی است. موضوعات به گونه‌ای در کتاب تنظیم شده‌اند که هم برای مبتدیان و هم برای حرفه‌ای‌ها مناسب هستند، چون ابتدا مفاهیم و پایه‌های نظری بررسی می‌شوند و در ادامه شما می‌توانید روش‌های پیشرفته‌ی داده‌کاوی را مطالعه کنید.

مطالب کتاب در سیزده فصل گردآوری و تهیه شده است. مقدمه‌ای بر داده‌کاوی و همچنین بیانی اجمالی از مطالب کتاب را می‌توانید در فصل اول مطالعه بفرمایید. گونه‌های متفاوت داده‌ها وخصوصیات هر یک از آنها در فصل دوم بررسی می‌شوند و در فصل سوم به سراغ روش‌های پیش‌پردازش داده‌ها می‌رویم. فصل‌های چهارم و پنجم به موضوع انبارش داده‌ها تخصیص داده شده است. کاوش الگوهای مکرر، مشارکت‌ها وهمبستگی‌ها موضوع فصل‌های ششم و هفتم را تشکیل می‌دهند. مفاهیم پایه‌ای دسته‌بندی در فصل هشتم بیان می‌شوند و به دنبال آن روش‌های پیشرفته‌ی آن در فصل نهم شرح داده شده‌اند. تحلیل خوشه‌بندی و روش‌های پیشرفته‌ی آن موضوع فصل‌های دهم و یازدهم است. فصلی مجزا برای انواع داده‌های پرت و روش‌های شناسایی آن در نظر گرفته شده است. در فصل دوازدهم بیشتر درباره‌ی داده‌های پرت صحبت می‌کنیم. موضوع آخرین فصل کتاب هم روندهاومرزهای تحقیق در حوزه‌ی داده‌کاوی است.

در ترجمه‌ی این کتاب تلاش شده است ضمن اصالت جمله‌بندی‌های کتاب، مطالب به گونه‌ای ترجمه شوند تا خوانندگان محترم بتوانند مفاهیم آن را به راحتی درک کنند. واژه‌های معادل در این متن پیشنهادی هستند و چه بسا برای برخی از اصطلاحات برابرهای دیگر (و احیاناً بهتر) بتوان یافت. آنچه مسلم است این است که اگر

چه گاهی تبدیل واژه‌ها آنچنانکه باید و شاید انجام نشده است اما در ادای جملات و بیان موضوع امانت مراعات و تلاش فراوان شده است تا خوانندگان گرامی با متنی مبهم و گیج کننده روبرو نشوند. بخش پایانی هر فصل در کتاب اصلی به معرفی منابع می‌پردازد که در ترجمه‌ی کتاب وجود ندارد. نه اینکه این بخش اهمیت نداشته باشد که به زعم مترجم یکی از بخش‌های ارزشمند کتاب به خصوص برای دوستان محقق و پژوهشگر محسوب می‌شود. اما از آنجا که کتاب دارای صفحات بالایی است و ما علاقه‌مند بودیم کتاب را در یک جلد و آنهم با مبلغی مناسب در اختیار خوانندگان محترم قرار دهیم، تصمیم گرفته شد تا این بخش در ترجمه قرار نداشته باشد. اما شما به سهولت می‌توانید به مطالب این بخش‌ها دسترسی داشته باشید.

در اینجا لازم می‌دانم از همه‌ی اساتید و دانشجویان به خاطر راهنمایی‌های ارزشمندشان در حین ترجمه‌ی این کتاب سپاسگزاری کنم. همچنین از مدیریت محترم انتشارات نیاز دانش نیز به خاطر آماده‌سازی، چاپ و پخش این کتاب تشکر می‌کنم. رهین محبت بی‌دریغ خانواده‌ام هستم که با فراهم‌سازی محیطی مناسب مرا یاری نمودند. با وجود همه‌ی سعی و تلاشی که در تمام مراحل آماده‌سازی این کتاب انجام گرفته است، یقین دارم که عاری از اشتباه نیست، چرا که تنها مکتوب بی‌نقص همان معجزه‌ی جاوید قرآن کریم است. در آخر ضمن سپاسگزاری از همه کسانی که مرا یاری داده‌اند و با پذیرش مسئولیت هرگونه کاستی احتمالی، امیدوارم که این اندک مفید افتد.

مهدی اسماعیلی
۱۳۹۳

فهرست مطالب

فصل ۱ **مقدمه** ۱۱

۱-۱ چرا داده‌کاوی؟ ۱۱
۱-۱-۱ حرکت به سوی عصراطلاعات ۱۱
۱-۱-۲ داده‌کاوی به عنوان سیر تکاملی فن‌آوری اطلاعات ۱۲
۱-۲ داده‌کاوی چیست؟ ۱۵
۱-۳ چه نوعی ازداده‌ها کاوش می‌شوند؟ ۱۷
۱-۳-۱ داده‌های پایگاه داده‌ها ۱۸
۱-۳-۲ انبارهای داده ۱۹
۱-۳-۳ داده‌های تراکنشی ۲۱
۱-۳-۴ گونه‌های دیگر داده‌ها ۲۲
۱-۴ چه نوعی از الگوها کاوش می‌شوند؟ ۲۳
۱-۴-۱ توصیف و تفکیک ۲۴
۱-۴-۲ کاوش الگوهای مکرر، مشارکت‌ها و وابستگی‌ها ۲۵
۱-۴-۳ دسته‌بندی و رگرسیون برای تحلیل پیشگویی ۲۶
۱-۴-۴ تحلیل خوشه ۲۸
۱-۴-۵ تحلیل داده‌های پرت ۲۹
۱-۴-۶ آیا کلیه‌ی الگوها جالب هستند؟ ۲۹
۱-۵ کدام فن‌آوری‌ها استفاده می‌شوند؟ ۳۱
۱-۵-۱ آمار ۳۱
۱-۵-۲ یادگیری ماشین ۳۲
۱-۵-۳ سیستم‌های پایگاه داده‌ها و انبارهای داده ۳۳
۱-۵-۴ بازیابی اطلاعات ۳۴
۱-۶ هدف داده‌کاوی کدام نوع از برنامه‌های کاربردی است؟ ۳۴
۱-۶-۱ هوش تجاری ۳۵
۱-۶-۲ موتورهای جستجوی وِب ۳۵
۱-۷ موضوعات عمده در داده‌کاوی ۳۶
۱-۷-۱ متدولوژی کاوش ۳۶
۱-۷-۲ تعامل کاربر(دخالت کاربر) ۳۸
۱-۷-۳ کارآمدبودن و قابلیت مقیاس‌پذیری ۳۸
۱-۷-۴ تنوع گونه‌های پایگاه داده‌ها ۳۹
۱-۷-۵ داده‌کاوی و جامعه ۳۹
۱-۸ خلاصه ۴۱
۱-۹ تمرین‌ها ۴۳

فصل ۲ **شناخت داده‌ها** ۴۵

۲-۱ نمونه‌ها و انواع صفات‌خاصه ۴۶
۲-۱-۱ یک صفت‌خاصه چیست؟ ۴۶
۲-۱-۲ صفات‌خاصه‌ی اسمی ۴۷
۲-۱-۳ صفات‌خاصه‌ی دودویی ۴۷
۲-۱-۴ صفات‌خاصه‌ی ترتیبی ۴۸
۲-۱-۵ صفات‌خاصه‌ی عددی ۴۸
۲-۱-۶ صفات‌خاصه‌ی گسسته در برابر پیوسته ۴۹
۲-۲ آمار توصیفی ۵۰
۲-۲-۱ محاسبه‌ی شاخص‌های مرکزی: میانگین، میانه و مُد ۵۰
۲-۲-۲ محاسبه‌ی شاخص‌های پراکندگی داده‌ها ۵۳
۲-۲-۳ نمایش گرافیکی آمار توصیفی ۵۷
۲-۳ مصورسازی داده‌ها ۶۱
۲-۳-۱ روش‌های مصورسازی پیکسل‌گرا ۶۲
۲-۳-۲ تکنیک‌های تصویرکردن هندسی ۶۳
۲-۳-۳ تکنیک‌های مصورسازی مبتنی بر شمایل ۶۶
۲-۳-۴ تکنیک‌های سلسله‌مراتبی مصورسازی ۶۷
۲-۳-۵ مصورسازی داده‌ها و روابط پیچیده ۶۸
۲-۴ محاسبه‌ی تشابه و عدم‌تشابه ۷۱
۲-۴-۱ ماتریس داده‌ها در مقابل ماتریس عدم‌تشابه ۷۱
۲-۴-۲ محاسبه‌ی همسایگی برای صفات‌خاصه‌ی اسمی ۷۳
۲-۴-۳ محاسبه‌ی همسایگی برای صفات‌خاصه‌ی دودویی ۷۴

۲-۴-۴ عدم‌تشابه داده‌های عددی: فاصله‌ی مینکوفسکی ۷۶
۲-۴-۵ محاسبه‌ی همسایگی برای صفات‌خاصه‌ی ترتیبی ۷۸
۲-۴-۶ محاسبه‌ی عدم‌تشابه برای مخلوطی از انواع صفات‌خاصه ۷۹
۲-۴-۷ تشابه کسینوسی ۸۱
۲-۵ خلاصه ۸۳
۲-۶ تمرین‌ها ۸۴

فصل ۳ پیش‌پردازش داده‌ها ۸۷
۳-۱ مروری بر پیش‌پردازش داده‌ها ۸۸
۳-۱-۱ کیفیت داده‌ها: چرا پیش‌پردازش داده‌ها؟ ۸۸
۳-۱-۲ وظایف اصلی در پیش‌پردازش داده‌ها ۸۹
۳-۲ پالایش داده‌ها ۹۱
۳-۲-۱ مقادیر ناموجود ۹۱
۳-۲-۲ داده‌های نویزدار ۹۳
۳-۲-۳ پالایش داده‌ها به عنوان یک فرایند ۹۴
۳-۳ یکپارچه‌سازی داده‌ها ۹۷
۳-۳-۱ مشکل شناسایی موجودیت ۹۷
۳-۳-۲ افزونگی و تحلیل همبستگی ۹۸
۳-۳-۳ تکرار تاپل ۱۰۱
۳-۳-۴ تشخیص و حل تصادم میان مقادیر داده‌ها ۱۰۲
۳-۴ کاهش داده‌ها ۱۰۲
۳-۴-۱ مروری بر راهبردهای کاهش داده‌ها ۱۰۲
۳-۴-۲ تبدیل‌های موجک ۱۰۳
۳-۴-۳ تحلیل مولفه‌های اصلی ۱۰۵
۳-۴-۴ انتخاب زیرمجموعه‌ای از صفات‌خاصه ۱۰۶
۳-۴-۵ رگرسیون و مدل‌های لگاریتمی-خطی: کاهش پارامتری داده‌ها ۱۰۸
۳-۴-۶ هیستوگرام‌ها ۱۰۸
۳-۴-۷ خوشه‌بندی ۱۱۰
۳-۴-۸ نمونه‌گیری ۱۱۰
۳-۴-۹ تجمیع در مکعب داده‌ها ۱۱۲
۳-۵ تبدیل داده‌ها و گسسته‌سازی داده‌ها ۱۱۳
۳-۵-۱ مروری اجمالی بر روی راهبردهای تبدیل داده‌ها ۱۱۳
۳-۵-۲ تبدیل داده‌ها با کمک نرمال‌سازی ۱۱۵
۳-۵-۳ گسسته‌سازی با کمک بسته‌بندی ۱۱۷
۳-۵-۴ گسسته‌سازی با تحلیل هیستوگرام ۱۱۷
۳-۵-۵ گسسته‌سازی با کمک خوشه‌بندی، درخت تصمیم و تحلیل همبستگی ۱۱۸
۳-۵-۶ ایجاد سلسله‌مراتب مفهومی برای داده‌های اسمی ۱۱۹
۳-۶ خلاصه ۱۲۲
۳-۷ تمرین‌ها ۱۲۳

فصل ۴ انبارش داده‌ها و پردازش تحلیلی برخط ۱۲۷
۴-۱ انبار داده‌ها: مفاهیم پایه‌ای ۱۲۷
۴-۱-۱ انبار داده‌ها چیست؟ ۱۲۸
۴-۱-۲ تفاوت‌های میان سیستم‌های پایگاه داده عملیاتی و انبارهای داده‌ها ۱۳۰
۴-۱-۳ چرا یک انبار داده‌ی مجزا؟ ۱۳۱
۴-۱-۴ انبارش داده‌ها: یک معماری چند ردیفی ۱۳۲
۴-۱-۵ مدل‌های انبار داده‌ها ۱۳۲
۴-۱-۶ استخراج، تبدیل و بارگذاری ۱۳۵
۴-۱-۷ مخزن مِتاداده‌ها ۱۳۵
۴-۲ مدل‌سازی انبار داده‌ها: مکعب داده‌ها و OLAP ۱۳۶
۴-۲-۱ مکعب داده‌ها: یک مدل داده‌ای چندبُعدی ۱۳۶
۴-۲-۲ ستاره‌ای، برفگونه و صورفلکی: شِماهایی برای مدل‌های داده‌ای چندبُعدی ۱۳۹
۴-۲-۳ ابعاد: نقش سلسله‌مراتب‌های مفهومی ۱۴۲
۴-۲-۴ سنجه‌ها: طبقه‌بندی و محاسبه‌ی آنها ۱۴۴
۴-۲-۵ عملیات معمول OLAP ۱۴۵
۴-۲-۶ مدل شبکه ستاره‌ای برای پرس‌وجو از پایگاه داده‌های چندبُعدی ۱۴۸
۴-۳ طراحی و کاربرد انبار داده‌ها ۱۴۹
۴-۳-۱ یک چارچوب تحلیل کسب‌وکار برای طراحی انبار داده‌ها ۱۴۹
۴-۳-۲ فرایند طراحی انبار داده‌ها ۱۵۰
۴-۳-۳ کاربرد انبار داده‌ها در پردازش اطلاعات ۱۵۱
۴-۳-۴ از پردازش تحلیلی برخط تا داده‌کاوی چندبُعدی ۱۵۳
۴-۴ پیاده‌سازی انبار داده‌ها ۱۵۴
۴-۴-۱ مروری بر محاسبه‌ی موثر مکعب داده‌ها ۱۵۵
۴-۴-۲ شاخص‌بندی داده‌های OLAP: شاخص نگاشت‌بیتی و شاخص پیوند ۱۵۸
۴-۴-۳ پردازش موثر پرسش‌های OLAP ۱۶۱
۴-۴-۴ معماری‌های سرویس‌دهنده‌ی OLAP ۱۶۲
۴-۵ تعمیم داده‌ها با کمک استنتاج صفت‌گرا ۱۶۳
۴-۵-۱ استنتاج صفت‌گرا برای توصیف داده‌ها ۱۶۵
۴-۵-۲ پیاده‌سازی مؤثر فرایند استنتاج صفت‌گرا ۱۷۰
۴-۵-۳ استنتاج صفت‌گرا برای مقایسه‌های کلاسی ۱۷۲
۴-۶ خلاصه ۱۷۶
۴-۷ تمرین‌ها ۱۷۸

فصل ۵ تکنولوژی مکعب داده‌ها ۱۸۳

۵-۱ محاسبه‌ی مکعب داده‌ها: مفاهیم مقدماتی ۱۸۴
۵-۱-۱ ساخت مکعب: مکعب کامل، مکعب کوه یخی، مکعب بسته و پوسته‌ی مکعب ۱۸۴
۵-۱-۲ راهبردهای کلی برای محاسبه‌ی مکعب داده‌ها ۱۸۸
۵-۲ روش‌های محاسبه‌ی مکعب داده‌ها ۱۹۰
۵-۲-۱ تجمیع چندراهه آرایه برای محاسبه‌ی مکعب کامل ۱۹۰
۵-۲-۲ BUC: محاسبه‌ی مکعب‌های کوه‌یخی از شبه‌مکعب نوک به سمت پایین ۱۹۵
۵-۲-۳ Star-Cubing: محاسبه‌ی مکعب‌های کوه‌یخی با کمک ساختار ستاره-درختی ۱۹۹
۵-۲-۴ پیش‌محاسبه‌ی قطعات پوسته برای تسریع OLAP در کار با ابعاد بالا ۲۰۴
۵-۳ پردازش گونه‌های پیشرفته‌ای از پرسش‌ها ۲۱۳
۵-۳-۱ مکعب‌های نمونه‌گیری: کاوش مبتنی بر OLAP بر روی داده‌های نمونه ۲۱۳
۵-۳-۲ مکعب‌های رتبه‌بندی: محاسبه‌ی موثر پرسش‌های رتبه‌ای ۲۱۹
۵-۴ تحلیل داده‌های چندبُعدی در فضای مکعب ۲۲۱
۵-۴-۱ مکعب‌های پیش‌گویانه: کاوش پیش‌گویانه در فضای مکعب ۲۲۱
۵-۴-۲ مکعب‌های چندجانبه: تجمیع پیچیده در چندین سطح از دانه‌بندی ۲۲۳
۵-۴-۳ کاوش اکتشافی و مبتنی بر استثناء فضای مکعب ۲۲۵
۵-۵ خلاصه ۲۲۹
۵-۶ تمرین‌ها ۲۳۰

فصل ۶ کاوش الگوهای مکرر، مشارکت‌ها و همبستگی‌ها: مفاهیم پایه و روش‌ها ۲۳۵

۶-۱ مفاهیم پایه ۲۳۶
۶-۱-۱ تحلیل سبد خرید: یک مثال انگیزشی ۲۳۶
۶-۱-۲ مجموعه‌اقلام مکرر، مجموعه اقلام بسته و قوانین انجمنی ۲۳۷
۶-۲ روش‌های کاوش مجموعه‌اقلام مکرر ۲۴۰
۶-۲-۱ الگوریتم Apriori: یافتن مجموعه‌اقلام مکرر با کمک تولید کاندیداهای محدود ۲۴۰
۶-۲-۲ ایجاد قوانین انجمنی از مجموعه‌اقلام مکرر ۲۴۵
۶-۲-۳ بهبود Apriori ۲۴۵
۶-۲-۴ رویکرد رشد الگو برای کاوش مجموعه‌اقلام مکرر ۲۴۷
۶-۲-۵ کاوش مجموعه‌اقلام مکرر با کمک قالب عمودی داده‌ها ۲۵۱
۶-۲-۶ کاوش الگوهای بسته و ماکسیمال ۲۵۳
۶-۳ کدامیک ازالگوها جالب هستند؟ روش‌های ارزشیابی الگو ۲۵۵
۶-۳-۱ قوانین قوی الزاماً جالب نیستند ۲۵۵
۶-۳-۲ از تحلیل مشارکت تا تحلیل همبستگی ۲۵۶
۶-۳-۳ مقایسه‌ای میان سنجه‌های ارزشیابی الگو ۲۵۸
۶-۴ خلاصه ۲۶۲
۶-۵ تمرین‌ها ۲۶۳

فصل ۷ مباحث پیشرفته در کاوش الگو ۲۶۷

۷-۱ کاوش الگو: یک نقشه‌ی راه ۲۶۷
۷-۲ کاوش الگو در فضای چندسطحی و چندبُعدی ۲۷۰
۷-۲-۱ کاوش مشارکت‌های چندسطحی ۲۷۱
۷-۲-۲ کاوش مشارکت‌های چندبُعدی ۲۷۴
۷-۲-۳ کاوش قوانین انجمنی کمّی ۲۷۶
۷-۲-۴ کاوش الگوهای نادر و الگوهای منفی ۲۷۸
۷-۳ کاوش الگوی مکرر مبتنی بر محدودیت ۲۸۱
۷-۳-۱ کاوش قوانین انجمنی با کمک مِتاقاعده ۲۸۲
۷-۳-۲ تولید الگو بر اساس محدودیت: هرس نمودن فضای الگو و فضای داده‌ها ۲۸۳
۷-۴ کاوش داده‌هایی با ابعاد بالا و الگوهای بسیار طولانی ۲۸۸
۷-۴-۱ کاوش الگوهای بسیار طولانی با کمک ائتلاف الگوها ۲۸۹
۷-۵ کاوش الگوهای فشرده یا تقریبی ۲۹۳
۷-۵-۱ کاوش الگوهای فشرده با کمک خوشه‌بندی الگو ۲۹۴
۷-۵-۲ استخراج k الگوی برتر با آگاهی از افزونگی ۲۹۷
۷-۶ کشف و کاربرد الگو ۲۹۹
۷-۶-۱ حاشیه‌نویسی معنایی الگوهای مکرر ۲۹۹
۷-۶-۲ کاربردهایی از کاوش الگو ۳۰۴
۷-۷ خلاصه ۳۰۷
۷-۸ تمرین‌ها ۳۰۸

فصل ۸ دسته‌بندی: مفاهیم پایه ۳۱۱

۸-۱ مفاهیم پایه ۳۱۱
۸-۱-۱ دسته‌بندی چیست؟ ۳۱۲
۸-۱-۲ رویکرد عمومی برای دسته‌بندی ۳۱۲
۸-۲ استقراء درخت تصمیم ۳۱۴
۸-۲-۱ استقراء درخت تصمیم ۳۱۵
۸-۲-۲ سنجه‌های انتخاب صفت‌خاصه ۳۱۹
۸-۲-۳ هرس نمودن درخت ۳۲۷
۸-۲-۴ مقیاسپذیری و استقراء درخت تصمیم ۳۲۹
۸-۲-۵ کاوش بصری جهت استقراء درخت تصمیم ۳۳۱
۸-۳ روش‌های دسته‌بندی بیزی ۳۳۳
۸-۳-۱ نظریه‌ی بیز ۳۳۳
۸-۳-۲ دسته‌بندی بیزی ساده ۳۳۴
۸-۴ دسته‌بندی مبتنی بر قاعده ۳۳۸
۸-۴-۱ دسته‌بندی با کمک قواعدی به شکل IF-THEN ۳۳۸
۸-۴-۲ استخراج قواعد از یک درخت تصمیم ۳۴۰
۸-۴-۳ استقراء قاعده با کمک الگوریتم پوشای متوالی ۳۴۲
۸-۵ ارزشیابی و انتخاب مدل ۳۴۷
۸-۵-۱ متریک‌هایی جهت ارزشیابی کارایی دسته‌بند ۳۴۷
۸-۵-۲ روش holdout و زیرنمونه‌گیری تصادفی ۳۵۳
۸-۵-۳ اعتبارسنجی متقابل ۳۵۳
۸-۵-۴ روش bootstrap ۳۵۴
۸-۵-۵ انتخاب مدل با کمک آزمون‌های آماری ۳۵۴
۸-۵-۶ مقایسه‌ی مدل‌ها بر اساس هزینه–سود و منحنی‌های ROC ۳۵۶
۸-۶ تکنیک‌هایی جهت بهبود صحت دسته‌بندی ۳۶۰
۸-۶-۱ معرفی روش‌های تلفیقی ۳۶۰
۸-۶-۲ روش bagging ۳۶۱
۸-۶-۳ روش boosting و الگوریتم AdaBoost ۳۶۲
۸-۶-۴ جنگل‌های تصادفی ۳۶۵
۸-۶-۵ بهبود صحت دسته‌بندی در داده‌هایی با عدم‌تعادل کلاس ۳۶۶
۸-۷ خلاصه ۳۶۹
۸-۸ تمرین‌ها ۳۷۰

فصل ۹ دسته‌بندی: روش‌های پیشرفته ۳۷۳

۹-۱ شبکه‌های بیز ۳۷۳
۹-۱-۱ مفاهیم و مکانیزم‌ها ۳۷۴
۹-۱-۲ آموزش شبکه‌های بیز ۳۷۵
۹-۲ دسته‌بندی با کمک الگوریتم پس‌انتشار ۳۷۷
۹-۲-۱ شبکه‌ی عصبی پیش‌خور چندلایه ۳۷۸
۹-۲-۲ معرفی یک توپولوژی شبکه ۳۷۹
۹-۲-۳ پس‌انتشار ۳۸۰
۹-۲-۴ درون جعبه سیاه: پس‌انتشار و قابلیت تفسیرپذیری ۳۸۵
۹-۳ ماشین‌های بردار پشتیبان ۳۸۷
۹-۳-۱ تفکیک‌پذیری خطی داده‌ها ۳۸۷
۹-۳-۲ تفکیک‌پذیری غیرخطی ۳۹۱
۹-۴ دسته‌بندی با کمک الگوهای مکرر ۳۹۴
۹-۴-۱ دسته‌بندی مشارکتی ۳۹۵
۹-۴-۲ دسته‌بندی مبتنی برالگوهای مکرر جداساز ۳۹۸
۹-۵ یادگیرنده‌های کُند (یادگیری از طریق همسایه‌ها) ۴۰۱
۹-۵-۱ دسته‌بندهای k همسایه‌ی نزدیک ۴۰۲
۹-۵-۲ استدلال مبتنی بر مورد ۴۰۴
۹-۶ روش‌های دسته‌بندی دیگر ۴۰۴
۹-۶-۱ الگوریتم‌های ژنتیک ۴۰۵
۹-۶-۲ رویکرد مجموعه‌های راف ۴۰۵
۹-۶-۳ مجموعه‌های فازی ۴۰۶
۹-۷ چند موضوع دیگر درباره‌ی دسته‌بندی ۴۰۸
۹-۷-۱ دسته‌بندی چندکلاسه ۴۰۹
۹-۷-۲ دسته‌بندی نیمه‌ناظر ۴۱۰
۹-۷-۳ یادگیری فعال ۴۱۲
۹-۷-۴ یادگیری انتقالی ۴۱۳
۹-۸ خلاصه ۴۱۶
۹-۹ تمرین‌ها ۴۱۸

فصل ۱۰ تحلیل خوشه: مفاهیم پایه و روش‌ها ۴۲۱

۱۰-۱ تحلیل خوشه ۴۲۲
۱۰-۱-۱ تحلیل خوشه چیست؟ ۴۲۲
۱۰-۱-۲ نیازهای تحلیل خوشه ۴۲۳
۱۰-۱-۳ مروری بر روش‌های پایه‌ی خوشه‌بندی ۴۲۶
۱۰-۲ روش‌های افراز ۴۲۸
۱۰-۲-۱ روش k-means: تکنیکی مبتنی بر گرانیگاه ۴۲۹
۱۰-۲-۲ روش k-medoids: تکنیکی مبتنی بر شی نماینده ۴۳۲
۱۰-۳ روش‌های سلسله‌مراتبی ۴۳۵
۱۰-۳-۱ خوشه‌بندی سلسله‌مراتبی تجمیعی در مقابل تقسیمی ۴۳۷

۱۰-۳-۲ سنجه‌های فاصله در روش‌های الگوریتمیک ۴۳۹
۱۰-۳-۳ الگوریتم BIRCH: خوشه‌بندی سلسله‌مراتبی چندمرحله‌ای با کمک درختان ویژگی خوشه‌بندی ۴۴۱
۱۰-۳-۴ الگوریتم Chameleon: خوشه‌بندی سلسله‌مراتبی چندمرحله‌ای با کمک مدل‌سازی پویا ۴۴۳
۱۰-۳-۵ خوشه‌بندی سلسله‌مراتبی احتمالاتی ۴۴۵
۱۰-۴ روش‌های مبتنی بر چگالی ۴۴۸
۱۰-۴-۱ DBSCAN: خوشه‌بندی مبتنی بر چگالی بر اساس اتصال مناطقی با چگالی بالا ۴۴۸
۱۰-۴-۲ OPTICS: نظم‌دهی نقاط برای شناسایی ساختار خوشه‌بندی ۴۵۱
۱۰-۴-۳ DENCLUE:خوشه‌بندی بر اساس توابع توزیع چگالی ۴۵۳
۱۰-۵ روش‌های مبتنی بر گرید ۴۵۶
۱۰-۵-۱ تکنیک STING ۴۵۶
۱۰-۵-۲ CLIQUE: یک تکنیک خوشه‌بندی زیرفضا مشابه Apriori ۴۵۸
۱۰-۶ ارزشیابی خوشه‌بندی ۴۶۱
۱۰-۶-۱ ارزیابی روند خوشه‌بندی ۴۶۱
۱۰-۶-۲ تعیین تعداد خوشه‌ها ۴۶۳
۱۰-۶-۳ اندازه‌گیری کیفیت خوشه‌بندی ۴۶۴
۱۰-۷ خلاصه ۴۶۸
۱۰-۸ تمرین‌ها ۴۶۹

فصل ۱۱ روش‌های پیشرفته‌ی تحلیل خوشه ۴۷۳
۱۱-۱ خوشه‌بندی بر پایه‌ی مدل احتمالاتی ۴۷۳
۱۱-۱-۱ خوشه‌های فازی ۴۷۵
۱۱-۱-۲ خوشه‌های مبتنی بر مدل احتمالاتی ۴۷۷
۱۱-۱-۳ الگوریتم EM ۴۸۰
۱۱-۲ خوشه‌بندی داده‌هایی با ابعاد بالا ۴۸۳
۱۱-۲-۱ خوشه‌بندی داده‌هایی با ابعاد بالا: مشکلات، چالش‌ها و متدولوژی‌های اصلی ۴۸۴
۱۱-۲-۲ روش‌های خوشه‌بندی زیرفضا ۴۸۵
۱۱-۲-۳ خوشه‌بندی دوگانه ۴۸۷
۱۱-۲-۴ روش‌های کاهش ابعاد و خوشه‌بندی طیفی ۴۹۵
۱۱-۳ خوشه‌بندی گراف و شبکه ۴۹۷
۱۱-۳-۱ کاربردها و چالش‌ها ۴۹۷
۱۱-۳-۲ سنجه‌های تشابه ۵۰۰
۱۱-۳-۳ روش‌های خوشه‌بندی گراف ۵۰۳
۱۱-۴ خوشه‌بندی همراه با محدودیت‌ها ۵۰۶
۱۱-۴-۱ گروه‌بندی محدودیت‌ها ۵۰۶
۱۱-۴-۲ روش‌هایی برای خوشه‌بندی همراه با محدودیت‌ها ۵۱۰
۱۱-۵ خلاصه ۵۱۳
۱۱-۶ تمرین‌ها ۵۱۵

فصل ۱۲ شناسایی داده‌های پرت ۵۱۷
۱۲-۱ داده‌های پرت و تحلیل آنها ۵۱۸
۱۲-۱-۱ داده‌ی پرت چیست؟ ۵۱۸
۱۲-۱-۲ انواع داده‌های پرت ۵۱۹
۱۲-۱-۳ چالش‌های تشخیص داده‌های پرت ۵۲۲
۱۲-۲ روش‌های تشخیص داده‌های پرت ۵۲۳
۱۲-۲-۱ روش‌های باناظر، نیمه‌ناظر و بی‌ناظر ۵۲۳
۱۲-۲-۲ روش‌های آماری و روش‌های مبتنی بر مجاورت و روش‌های مبتنی بر خوشه‌بندی ۵۲۵
۱۲-۳ رویکردهای آماری ۵۲۷
۱۲-۳-۱ روش‌های پارامتری ۵۲۷
۱۲-۳-۲ روش‌های بدون پارامتر ۵۳۲
۱۲-۴ رویکردهای مبتنی بر مجاورت ۵۳۴
۱۲-۴-۱ تشخیص داده‌های پرت بر مبنای فاصله و یک روش با حلقه‌ی تودرتو ۵۳۴
۱۲-۴-۲ یک روش مبتنی بر گرید ۵۳۶
۱۲-۴-۳ تشخیص داده‌های پرت بر مبنای چگالی ۵۳۸
۱۲-۵ رویکردهای مبتنی بر خوشه بندی ۵۴۱
۱۲-۶ رویکردهای مبتنی بر دسته‌بندی ۵۴۴
۱۲-۷ کاوش داده‌های پرت بافتاری و گروهی ۵۴۶
۱۲-۷-۱ تبدیل روش‌های تشخیص داده‌های پرت بافتاری به روش‌های متداول برای این کار ۵۴۷
۱۲-۷-۲ مدلسازی رفتار نرمال با توجه به بافتها ۵۴۸
۱۲-۷-۳ کاوش داده‌های پرت گروهی ۵۴۸
۱۲-۸ تشخیص داده‌های پرت در داده‌هایی با ابعاد بالا ۵۵۰
۱۲-۸-۱ تعمیم روش‌های متداول تشخیص داده‌های پرت ۵۵۱
۱۲-۸-۲ یافتن داده‌های پرت در زیرفضاها ۵۵۱
۱۲-۸-۳ مدل‌سازی داده‌های پرت با ابعاد بالا ۵۵۳
۱۲-۹ خلاصه ۵۵۵
۱۲-۱۰ تمرین‌ها ۵۵۷

فصل ۱۳ روندها و مرزهای تحقیق در حوزه‌ی داده‌کاوی ۵۵۹
۱۳-۱ کاوش انواع داده‌های پیچیده ۵۵۹
۱۳-۱-۱ کاوش توالی‌ها: سری‌های زمانی، توالی‌های نمادین و توالی‌های زیستی ۵۶۰
۱۳-۱-۲ کاوش گراف‌ها و شبکه‌ها ۵۶۵
۱۳-۱-۳ کاوش گونه‌های دیگر داده‌ها ۵۶۸
۱۳-۲ متدولوژی‌های دیگر داده‌کاوی ۵۷۲
۱۳-۲-۱ داده‌کاوی آماری ۵۷۲
۱۳-۲-۲ دیدگاه‌هایی در مورد پایه‌های نظری داده‌کاوی ۵۷۴
۱۳-۲-۳ داده‌کاوی تصویری و صوتی ۵۷۵
۱۳-۳ کاربردهای داده‌کاوی ۵۸۱
۱۳-۳-۱ داده‌کاوی برای تحلیل داده‌های مالی ۵۸۱
۱۳-۳-۲ داده‌کاوی برای صنایع خرده‌فروشی و مخابرات ۵۸۲
۱۳-۳-۳ داده‌کاوی در علوم و مهندسی ۵۸۴
۱۳-۳-۴ داده‌کاوی برای تشخیص نفوذ و پیشگیری از آن ۵۸۶
۱۳-۳-۵ داده‌کاوی و سیستم‌های پیشنهاد دهنده ۵۸۸
۱۳-۴ داده‌کاوی و جامعه ۵۹۱
۱۳-۴-۱ داده‌کاوی غیرمحسوسی که در همه جا حضور دارد ۵۹۱
۱۳-۴-۲ محرمانگی، امنیت و اثرات اجتماعی داده‌کاوی ۵۹۳
۱۳-۵ روندهای داده‌کاوی ۵۹۶
۱۳-۶ خلاصه ۵۹۸
۱۳-۷ تمرین‌ها ۵۹۹

فصل ۱

مقدمه

این کتاب مقدمه‌ای بر موضوع داده‌کاوی[1] است که گاه با عنوان کشف دانش[2] از داده‌ها نیز شناخته می‌شود. داده‌کاوی حوزه‌ی جدیدی تلقی می‌شود و به سرعت در حال رشد است. کتاب بر روی مفاهیم و تکنیک‌های اصلی داده‌کاوی متمرکز می‌شود که با کمک آنها می‌توان الگوهای جالبی را از داده‌ها و در برنامه‌های کاربردی متفاوت کشف نمود. تاکید ما بر روی تکنیک‌هایی است که جهت توسعه‌ی ابزارهای داده‌کاوی موثر، کارا و مقیاس‌پذیر از آنها استفاده می‌شود.

فصل حاضر بدین ترتیب سازمان‌دهی شده است. در بخش۱-۱ خواهیم آموخت که چرا تقاضا برای داده‌کاوی افزایش یافته است و چرا بخشی از سیر تکامل تکنولوژی اطلاعات تلقی می‌شود. بخش ۱-۲ به معرفی داده‌کاوی در فرایند کشف دانش می‌پردازد. در ادامه جنبه‌های زیادی در مورد داده کاوی را بحث خواهیم کرد، نوع داده‌هایی که می‌توانند کاوش شوند (بخش ۱-۳)، گونه‌هایی از دانش که نتیجه‌ی کاوش است (بخش ۱-۴)، انواع تکنولوژی‌هایی که استفاده می‌شوند (بخش ۱-۵) و برنامه‌های کاربردی که در آنها از داده‌کاوی استفاده شده است (بخش ۱-۶). بدین ترتیب از منظرهای متفاوتی به داده‌کاوی خواهیم پرداخت. در بخش ۱-۷ که بخش پایانی فصل را تشکیل می‌دهد چالش‌های تحقیقاتی این حوزه بررسی می‌شوند.

۱-۱ چرا داده‌کاوی؟

ما در دنیایی زندگی می‌کنیم که روزانه حجم وسیعی از داده‌ها تولید می‌شود. یکی از نیازهای مهم تحلیل این داده‌های حجیم است. بخش ۱-۱-۱ نگاهی به این موضوع دارد که چگونه داده‌کاوی با مهیا ساختن ابزارهایی جهت کشف دانش از داده‌ها این نیاز را مرتفع می‌سازد. در بخش ۱-۱-۲ نیز مشاهده خواهیم کرد که چگونه داده‌کاوی را می‌توان نتیجه سیر تکاملی تکنولوژی اطلاعات دانست.

۱-۱-۱ حرکت به سوی عصراطلاعات

جمله‌ی "ما درحال زندگی در عصر اطلاعات هستیم" شاید یک جمله‌ی رایج تلقی شود، اما بهرحال و در واقع ما در حال زندگی درعصر داده‌ها هستیم. قرار گرفتن حجم وسیع داده‌ها در مقیاس ترابایت و یا حتی پتابایت بر روی شبکه‌های کامپیوتری، شبکه جهانی وِب و رسانه‌های ذخیره‌سازی متفاوت که داده‌ها را از علوم متفاوتی جمع‌آوری و نگهداری می‌کنند، این موضوع را به خوبی اثبات می‌کند. رشد انفجاری حجم داده‌ها نتیجه‌ی مکانیزه شدن جامعه‌ی ماست و همچنین توسعه‌ی ابزارهای قدرتمند برای جمع‌آوری و نگهداری داده‌ها است. کسب‌وکارهای

---

[1] Data Mining [2] Knowledge Discovery