دانشکده مهندسی

پایان‌نامه کارشناسی ارشد در رشته مهندسی کامپیوتر(نرم‌افزار)

بررسی کاوش وابستگی‌های تابعی تقریبی

توسط:

حسین فرجی

استاد راهنما:

دکتر محمدهادی صدرالدینی

اسفند ۱۳۸۸


بسم الله الرحمن الرحيم

چکیده

بررسی کاوش وابستگی‌های تابعی تقریبی

بوسیله‌ی:

حسین فرجی

حجم روز افزون داده‌ها در فایلها، پایگاه داده‌ها و دیگر انباره‌های داده، توسعه روش‌های تجزیه و تحلیل و استخراج اطلاعات مفید و ضمنی موجود در داده‌ها را ایجاب می‌نماید. این اطلاعات می‌توانند در پروسه‌های تصمیم‌گیری سازمان‌ها بسیار موثر باشند. داده‌کاوی کشف الگوهای پنهان واطلاعات مفید از پایگاه داده‌ها است. داده‌کاوی یکی از قدم‌های مهم در پروسه کشف دانش است.

ارتباطات پنهان و نسبی بین خصیصه‌های موجود در پایگاه داده‌های رابطه‌ای را می‌توان توسط روش‌های کاوش وابستگی‌های تابعی تقریبی بدست آورد. همچنین ارتباطات پنهان بین مقادیر این خصیصه‌ها را می‌توان توسط قوانین انجمنی بیان نمود. در این پایان‌نامه یکی از بهترین روش‌های کاوش وابستگی‌های تابعی تقریبی بهبود داده شده است. در این بهبود علاوه بر تولید نتایج مفیدتر کارایی الگوریتم نیز در بسیاری از موارد بهبود می‌یابد. همچنین با استفاده از مفهوم وابستگی‌های تابعی تقریبی روش جدیدی بنام AR-Miner برای کاوش قوانین انجمنی از پایگاه داده‌های رابطه‌ای ارائه می‌شود. روش جدید علاوه بر کارا بودن می‌تواند قوانین انجمنی را مستقیما از داخل پایگاه داده‌های رابطه‌ای استخراج کند.

در انتها با استفاده از الگوریتم ارائه شده برای کاوش وابستگی‌های تابعی تقریبی، خصیصه‌های شرکت کننده در دسته‌بندی ماژول‌های معیوب نرم افزار را بدون کم شدن کارایی دسته‌بند کاهش می‌دهیم.

Abstract

An investigation into mining of approximate functional dependencies

By:

Hossein Faraji

With the enormous amount of data stored in files, databases and other repositories, it is increasingly important to develop powerful means for extraction, analysis and interpretation of such data. Extraction of interesting knowledge could help in decision-making tasks. Data Mining refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.

Mining of approximate functional dependencies can extract approximate dependencies between attributes of relational datasets. Moreover dependencies between values can be expressed by association rules. In this thesis, one of the best methods for mining approximate functional dependencies has been improved. The quality of generated dependencies is increased by pruning redundant dependencies. In addition, in many cases the run time of the algorithm has been improved. Moreover, a new method called AR-Miner has been proposed for association rule mining base on approximate functional dependencies. This new method uses the relational database directly without the need for converting the dataset into a transactional form.

Finally, we applied our proposed method to a software detection fault classification system. In this setup, we were able to decrease the number of features used for classification without any loss of accuracy.

فهرست مطالب

عنوان ..... صفحه

فهرست اشکال

عنوان ..... صفحه

فهرست جداول

عنوان ..... صفحه

۱. مقدمه

۱-۱ کلیاتی در مورد داده‌کاوی

از هنگامی که رایانه در تحلیل و ذخیره سازی داده ها بکار رفت (۱۹۵۰)، پس از حدود ۲۰ سال، حجم داده ها در پایگاه داده ها دو برابر شد. ولی پس از گذشت دو دهه و همزمان با پیشرفت فن آوری اطلاعات (IT) هر دو سال یکبار حجم داده ها، دو برابر شد. همچنین تعداد پایگاه داده ها با سرعت بیشتری رشد نمود. این در حالی است که تعداد متخصصین تحلیل داده ها و آمارشناسان با این سرعت رشد نکرد. حتی اگر چنین امری اتفاق می‌افتاد، بسیاری از پایگاه داده‌ها چنان گسترش یافته‌اند که شامل چند صد میلیون یا چند صد میلیارد رکورد ثبت شده هستند و امکان تحلیل و استخراج اطلاعات با روش های معمول آماری از دل انبوه داده‌ها مستلزم چند روز کار با رایانه های موجود است. حال با وجود سیستم های یکپارچه اطلاعاتی، سیستم های یکپارچه بانکی و تجارت الکترونیک، لحظه به لحظه به حجم داده ها در پایگاه داده های مربوط اضافه شده و باعث به وجود آمدن انبارهای (توده های) عظیمی از داده‌ها شده است به طوری که ضرورت کشف و استخراج سریع و دقیق دانش از این پایگاه داده‌ها را بیش از پیش نمایان کرده است، چنانکه در عصر حاضر گفته می‌شود: « اطلاعات طلاست».

هم اکنون در هر کشور، در سازمان ها و شرکت ها برای امور بازرگانی، پرسنلی، آموزشی، آماری پایگاه داده هایی ایجاد یا خریداری شده است، به طوری‌که این پایگاه داده ها برای مدیران، برنامه ریزان، پژوهشگران جهت تصمیم گیری‌های راهبردی، تهیه گزارش‌های مختلف، توصیف وضعیت جاری خود می‌تواند مفید باشد. داده‌کاوی[1] یا استخراج و کشف سریع و دقیق اطلاعات با ارزش و پنهان از این پایگاه داده‌ها از جمله اموری است که هر کشور، سازمان و شرکتی به منظور توسعه علمی، فنی و اقتصادی خود به آن نیاز دارد.

[1] Data mining

در کشور ما نیز سازمان ها، شرکت ها و مؤسسات دولتی و خصوصی به طور فزاینده ولی آهسته در حال ایجاد یا خرید نرم افزارهای پایگاه داده ها و مکانیزه کردن سیستم های اطلاعات خود هستند، همچنین با توجه به فصول دهم و یازدهم قانون برنامه سوم توسعه در خصوص داد و ستدهای الکترونیکی و همچنین تأکید بر برخورداری کشور از فن آوری های جدید اطلاعات برای دستیابی آسان به اطلاعات داخلی و خارجی، دولت مکلف شده است امکانات لازم برای دستیابی آسان به اطلاعات، زمینه سازی برای اتصال کشور به شبکه های جهانی و ایجاد زیر ساخت های ارتباطی و شاهراه های اطلاعاتی فراهم کند. واضح است این امر باعث ایجاد پایگاه های عظیم داده ها شده و ضرورت استفاده از داده کاوی را بیش از پیش نمایان می‌سازد.

داده کاوی و کشف دانش در پایگاه داده ها از جمله موضوع هایی هستند که همزمان با ایجاد و استفاده از پایگاه داده ها در اوایل دهه ۸۰ برای جستجوی دانش در داده ها شکل گرفت.

شاید بتوان لوول (۱۹۸۳) را اولین شخصی دانست که گزارشی در مورد داده کاوی تحت عنوان « شبیه سازی فعالیت داده کاوی » ارائه نمود. همزمان با او پژوهشگران و متخصصان علوم رایانه، آمار، هوش مصنوعی، یادگیری ماشین و. . . نیز به پژوهش در این زمینه و زمینه‌های مرتبط با آن پرداخته اند.

پژوهش جدی روی موضوع داده کاوی از اوایل دهه ۹۰ شروع شد. پژوهش ها و مطالعه‌های زیادی در این زمینه صورت گرفته، همچنین سمینارها، دوره های آموزشی و کنفرانس هایی نیز برگزار شده است. نتایج پایه های نظری داده کاوی در تعدادی از مقاله های پژوهشی آورده شده است. مثلاً پیاتتسکی و شاپیرو [۲] در سال ۱۹۹۱ « استقلال آماری قاعده ها در داده کاوی» را بررسی نموده اند. هافمن و نش در سال ۱۹۹۵ استفاده از داده‌کاوی و انباره داده[۳] توسط بانک های آمریکا را بررسی نموده و بیان کردند که چگونه این سیستم ها برای بانک های آمریکا قدرت رقابت بیشتری ایجاد می‌کنند. چت فیلد مشکلات ایجاد شده توسط داده کاوی را بررسی نمود و همچنین مقاله ای تحت عنوان « مدل های خطی غیر دقیق داده کاوی و استنباط آماری » ارایه نمود. هندری نیز دیدگاه اقتصاد سنجی روی داده کاوی را تهیه کرد. در این سال انجمن داده کاوی همزمان با اولین کنفرانس بین المللی «کشف دانش و داده کاوی» شروع به کار کرد. این کنفرانس توسعه یافته چهار

[2] Piatetsky-Shapiro
[3] Data warehouse

دوره آموزشی بین المللی در پایگاه هـای داده درسـال ۱۹۸۹ تـا ۱۹۹۴ بـود. انجمـن مـذکور، یـک سازمان علمی بـه نـام ACM- SIGKDD را ایجـاد نمـود. ایمیلنـسکی[۴] و منـیلا[۵] در سـال ۱۹۹۶ دیدگاهی از داده کاوی به عنوان «پرس و جو کننده از پایگاه های استنتاجی[۶]» را پیـشنهاد کردنـد. فایاد، پیاتتسکی – شاپیرو و اودورأسامی پیشرفت های کشف دانش و داده کـاوی را عنـوان کردنـد. منیلا در سال ۱۹۹۷ خلاصه ای از مطالعه روی اساس داده کاوی ارایه نمود. باربـارا و همکـاران نیـز دیدگاه کاهش داده هـا روی داده کـاوی را در گـزارش کـاهش دادههـای نیوجرسـی ارایـه نمودنـد. همچنین می‌توان برای کاربرد داده‌کاوی در مدیریت مالی می‌توان، تحلیـل دادههـای مـالی و مـدل سازی مالی بنینگاه و چاچ کز و هیگینز [۷] را ملاحظه کرد فریدمن نیز مقاله‌ای در ارتبـاط بـا مفهـوم آمار و داده‌کاوی ارایه نمود. هند [۸] در سـال ۱۹۹۸ مقالـه ای تحـت عنـوان « داده کـاوی : آمـار یـا بیشتر؟ » ارائه نمود. کلینبرگ[۹] پائودیمیتریو و راغـان[۱۰] دیـدگاه اقتـصاد سـنجی روی داده کـاوی و عملکرد داده کاوی به عنوان یک مسئله بهینه را ارائه نمودند. در این سال نیز کنفرانس هـای ناحیـه ای و بین‌المللی در مورد داده کاوی برگزار شد که از جمله می‌توان بـه کنفـرانس آسـیا و اقیانوسـیه درباره کشف دانش و داده کاوی اشاره کرد. هنـد و همکـاران و اسـمیت در سـال ۲۰۰۰ بحـث‌هـای مقایسه‌ای بین آمار و داده‌کاوی را ارائه کردند. سری و استاوا، کولی، رش پانـد و تـن اسـتفاده از وب در کاوش داده ها و کاربردهای آن را ارایه کردنـد. کلادیـو کانورسـانو و همکـاران در سـال ۲۰۰۲ « مدل آمیختـه چندگانـه جمـع پـذیر تعمـیم یافتـه » بـرای داده کـاوی را بررسـی نمودنـد. پـائلو و گیانلوکاپاسرون، « داده کاوی ساختارهای پیوند برای مدل رفتار مصرف کننده » را ارایه نمودند.

عبارت داده کاوی مترادف با یکی از عبارت های استخراج دانش، برداشت اطلاعات، وارسی داده هـا و حتی لایروبی کردن داده هاست کـه در حقیقـت کـشف دانـش در پایگـاه داده هـا[۱۱] (KDD) را توصیف می‌کند. بنابراین ایده ای که مبنـای داده کـاوی اسـت یـک فرآینـد بـا اهمیـت از شـناخت الگوهای بالقوه مفید، تازه و درنهایت قابل درک در داده هاست. واژه کشف دانش در پایگـاه داده هـا

[4] Imielnski
[5] Mannila
[6] Inductive databases
[7] Benninga, Czaczkes, Higgins
[8] Hand
[9] Kleinberg
[10] Paodimitriou , Raghavan
[11] Knowledge Discovery of Database

در اوایل دهه ۸۰ در مراجعه به مفهوم کلی، گسترده، سطح بالا و به دنبال جستجوی دانش در اطلاعات شکل گرفته است. داده کاوی کاربرد سطح بالای فنون و ابزار بکار برده شده برای معرفی و تحلیل داده های تصمیم گیرندگان است. اصطلاح داده کاوی را آمارشناسان، تحلیل‌گران داده ها و انجمن سیستم های اطلاعات مدیریت به کار برده اند، در حالی که پژوهشگران یادگیری ماشین و هوش مصنوعی از KDD بیشتر استفاده می‌کنند. در ادامه چند تعریف از داده کاوی ارائه می‌شود.

۱- «داده کاوی یا به تعبیر دیگر کشف دانش در پایگاه داده ها، استخراج غیر بدیهی اطلاعات بالقوه مفید از روی داده هایی است که قبلاً، ناشناخته مانده اند. این مطلب برخی از روش های فنی مانند خوشه بندی، خلاصه سازی داده ها، فراگیری قاعده های رده بندی، یافتن ارتباط شبکه ها، تحلیل تغییرات و کشف بی قاعدگی را شامل می‌شود » (پیاتتسکی شاپیرو، ماتئوس کریستوفر)

۲- « داده کاوی در حقیقت کشف ساختارهای جالب توجه، غیر منتظره و با ارزش از داخل مجموعه وسیعی از داده ها می‌باشد و فعالیتی است که اساساً با آمار و تحلیل دقیق داده ها منطبق است» هند (۱۹۹۸).

۳- « داده کاوی فرآیند کشف رابطه ها، الگوها و روندهای جدید معنی داری است که به بررسی حجم وسیعی از اطلاعات ذخیره شده در انبارهای داده با فناوری های تشخیص الگو (مانند ریاضی و آمار) می‌پردازد». (سایت http://www.spss.com)

کشف دانش در پایگاه داده ها در جهت کشف اطلاعات مفید از مجموعه بزرگ داده هاست. دانش کشف شده می‌تواند قاعده ای باشد تا ویژگی های داده ها، الگوهایی که به طور متناسب رخ می‌دهند، خوشه بندی موضوع های درون پایگاه داده ها و غیره را توصیف کند.

یک کاربر سیستم KDD بایستی درک بالایی از قلمرو داده ها به منظور انتخاب زیر مجموعه صحیحی از داده ها، رده مناسبی از الگوها و معیار خوبی برای الگوهای جالب داشته باشد. بنابراین سیستم KDD باید ابزارهایی با اثر تعاملی داشته باشد نه سیستم های تجزیه و تحلیل خودکار. لذا کشف دانش از پایگاه داده‌ها یک فرآیند شامل گام‌های زیر است:

۱- درک قلمرو

۲- آماده کردن مجموعه داده ها

۳- کشف الگوها (داده کاوی)

۴- پردازش بعد از کشف الگو

۵- استفاده از نتایج.

در فرآیند بالا، داده‌های خام از منابع مختلفی جمع‌آوری می‌شوند و از طریق استخراج، ترجمه و فرآیندهای بازخوانی به انبار داده‌ها وارد می‌شوند. در بخش مهیاسازی داده‌ها، داده‌ها از انبار خارج شده و به صورت یک فرمت مناسب برای داده‌کاوی درمی‌آیند. در بخش کشف الگو با روش‌های داده‌کاوی برای پاسخ به سؤال‌های خاصی که به ذهن می‌رسند، الگوریتم‌هایی را استخراج می‌کنند و از این الگوریتم‌ها برای ساخت الگو استفاده می‌شود. در بخش تجزیه و تحلیل الگو، الگوها به یک دانش مفید و قابل استفاده تبدیل می‌شوند و پس از بهبود آن‌ها، الگوهایی که کارا محسوب می‌شوند در یک سیستم اجرایی به کار گرفته خواهند شد.

طی سال‌های گذشته جریان سریعی از تمایل به داده‌کاوی در بازارهای نرم‌افزاری به وجود آمده است. بیشتر کاربران نرم‌افزارهای داده‌کاو با تفکر استفاده تجاری از این نرم‌افزارها، خواهان استفاده از آن شده‌اند. نرم‌افزارهای داده‌کاو معمولاً سه روش مختلف را برای استفاده از داده‌کاوی به کار می‌برند. ۱) اکتشاف ۲) استفاده از مدل‌های پیشگویی ۳) استفاده از آنالیز بحث و جدل.

اکتشاف، فرآیند جستجو در داده‌هاست تا الگوهای مخفی موجود در داده‌ها را بدون هیچ ایدهٔ از پیش تعیین شده‌ای مشخص نماید. در نرم‌افزارهای داده‌کاوی مبتنی بر مدل‌های پیشگویی، الگوهایی که از یک بانک داده کشف می‌شوند، برای پیش‌بینی آینده به کار می‌روند. مدل‌های پیش‌بینی به کاربر اجازه می‌دهند تا داده‌های نامشخص را به کار ببرد و این مقادیر نامشخص توسط نرم‌افزار کشف شود.

در مدل‌های جدلی نیز الگوهای یافت شده از داده‌ها برای تعیین مقادیر غیرعادی به کار می‌رود. برای تعیین مقادیر غیر عادی، ابتدا می‌بایست مقادیر عادی شناخته شود تا بر این اساس مقادیر غیرعادی و منحرف شناخته شوند.

نرم‌افزارهای داده‌کاوی در حال حاضر از فعالیت کمتری نسبت به سایر نرم‌افزارهای هوشمند برخوردار هستند. با این وجود فعالیت تجاری این نرم‌افزار را می‌توان در شش بخش کلی، دسته‌بندی داده‌ها، برآورد مقادیر نامشخص، پیش‌بینی مقادیر نامشخص، گروه‌بندی تقریبی داده‌ها، خوشه‌بندی داده‌ها و تشریح روابط بین داده‌ها تقسیم کرد.

اگر چه دانش به طور انحصاری محصول فناوری اطلاعات نیست، ولی فناوری اطلاعات به طور لاینفکی در ایجاد دانش و فرآیند مدیریت دانش از سال های اول مشارکت داشته است. امروزه مدیریت دانش از مسئولیت های فناوری اطلاعات به شمار می‌رود. زیرا در جمع‌آوری، تبدیل دانش و انتقال داده‌ها، اطلاعات و دانش نقش کلیدی دارد.

از منظر مدیریت دانش، هدف داده‌کاوی، کشف دانش سازمانی پنهان در اطلاعات خام است. اینگونه نیست که هر بینش حاصل از داده‌کاوی دانش می‌سازد، بلکه در عوض بسیاری از نتایج به دست آمده، اطلاعات مدیریت، یا هوش سازمانی است. مثلاً در سازمان‌های تجاری، دانش با ارزش در مورد مشتری، محصول و بازار را می‌توان از طریق داده‌کاوی به دست آورد. داده‌کاوی ابزار مفیدی برای مدیران دانش است که کشف را با تحلیل تلفیق می‌کنند. تلفیقی که اغلب منجر به ایجاد دانش می‌شود [6-1].

امروزه حجم زیادی از اطلاعات در پایگاه های داده مربوط به شرکت ها، مراکز تجاری و دولتی ذخیره می‌شود. استفاده معمول از این داده‌ها انجام عملیات گزارش‌گیری برای کاربران و مدیران است. استفاده دیگری که امروزه از حجم انبوه داده های ذخیره شده در پایگاه های داده و انبار های داده می شود انجام عملیات داده کاوی است. در عملیات داده کاوی ما به دنبال الگوهای پنهان و احتمالاً سودمند هستیم. برخی از این الگوها در انجام تصمیم گیری ها می توانند به مدیران کمک کرده و یا برای کاربران و مشتریان مفید باشند. الگوهایی که در عملیات مختلف داده کاوی پیدا می شوند انواع گوناگونی دارند. یک نوع پرکاربرد و معروف از این الگوها قوانین یا قواعد وابستگی هستند. معروف ترین کاربرد قوانین وابستگی در تحلیل سبد خرید برای فروشگاه ها و مراکز تجاری است. به عنوان مثال پس از کاوش پایگاه داده مربوط به یک فروشگاه زنجیره ای ممکن است مشخص شود که مشتریانی که از این فروشگاه شیر می خرند به احتمال ۶۰٪ کره نیز خواهند خرید. یافتن چنین قواعدی می تواند در چیدن قفسه ها، راهنمایی مشتریان و مسائل مدیریتی سودمند باشد. عملیات یافتن قواعد وابستگی را کشف یا کاوش قواعد وابستگی گویند.

۱-۲ داده‌کاوی

سرعت تولید و جمع آوری داده ها در پایگاههای داده به صورت روزافزونی زیاد شده است. استفاده گسترده از بارکد درفروش تولیدات، کامپیوتری شدن تعداد زیادی از کارهای تجاری، اداری و دولتی و پیشرفت در زمینه ابزار جمع آوری داده ها ما را با حجم زیادی از داده ها مواجه کرده است. امروزه پایگاه داده در زمینه های تجاری، اداری، علمی، مهندسی و زمینه های دیگر استفاده می شوند. تعداد چنین پایگاههای داده ای به دلیل نیاز مبرم به جمع آوری و گزارش گیری از داده ها و همچنین وجود سیستم های پایگاه داده قدرتمند در حال افزایش است. این چنین رشد فزاینده ای در داده ها و پایگاههای داده یک نیاز ضروری برای ابزار جدیدی که بتوانند به طور هوشمند و خودکار این داده ها را پردازش کرده و به اطلاعات و به دانش های سودمند تبدیل کنند، بوجود آورده است. در نتیجه داده کاوی به یک زمینه تحقیقاتی با اهمیت فراوان تبدیل شده است.

از داده کاوی همچنین به عنوان کشف دانش در پایگاههای داده یاد می شود، که به معنی فرآیند استخراج اطلاعات غیر صریح و احتمالاً سودمندی از پایگاههای داده است که در گذشته نا شناخته و پنهان بوده اند. با انجام عملیات داده کاوی دانش های جالب و گاه غیر منتظره، نظم ها و الگوهای پنهان، یا اطلاعات سطح بالا می توانند از مجموعه ای از داده های مرتبط در پایگاه داده استخراج شوند و از زوایای مختلف مورد بررسی قرار گیرند. بنابراین پایگاههای داده حجیم را می توان به عنوان منبعی غنی و قابل اطمینان برای تولید و وارسی برخی دانش ها و اطلاعات در نظر گرفت.

کاوش اطلاعات و دانش از پایگاههای داده حجیم به عنوان یک موضوع کلیدی برای محققینی که در زمینه پایگاههای داده و یادگیری ماشین کارمی کنند و به فرصتی برای کسب درآمد های بیشتر توسط شرکت های صنعتی و تجاری تبدیل شده است. دانش های کشف شده توسط داده کاوی می توانند در مدیریت اطلاعات، پردازش گزارش ها، انجام تصمیم گیری ها و بسیاری زمینه های دیگر استفاده ۱۴ شوند. به علت وجود گسترده داده ها در حجم زیاد و نیاز مبرم به تبدیل این داده ها به اطلاعات و دانش مفید برای کاربرد های مختلف، داده کاوی در سال های اخیر توجه زیادی را به خود جلب کرده است [7,8].

داده کاوی موضوعی وابسته به کاربرد است و کاربردهای مختلف نیازمند روش ها و تکنیک های داده کاوی مختلفی هستند. کاوش قواعد وابستگی[۱۲] ، دسته بندی[۱۳] ، خوشه بندی[۱۴] ، پیش بینی[۱۵] و تحلیل سری های زمانی[۱۶] از جمله مهمترین روش ها و تکنیک های داده کاوی به شمار می آیند. در ادامه هر کدام از این روش ها به صورت خلاصه توضیح داده می شوند. کشف یا کاوش قواعد وابستگی در پایگاه داده های رابطه ای[۱۷] یا تراکنشی[۱۸] که موضوع اصلی این مطالعه نیز است، اخیراً جذابیت زیادی را در انجمن های مربوط به پایگاههای داده بوجود آورده است. در این تکنیک داده کاوی وابستگی ها و ارتباطات بین داده های موجود در یک پایگاه داده بدست می آیند. نتیجه این عملیات داده کاوی دسته ای از قواعد است که به آنها قواعد وابستگی گفته می‌شود.

یکی دیگر از روش های مهم داده کاوی توانایی انجام عملیات دسته بندی در حجم زیاد داده هاست. این عملیات کاوش قوانین دسته بندی نیز نامیده می شود. در این روش اشیاء موجود در یک پایگاه داده بر اساس مقادیر چند خصوصیت از آنها، به دسته های مجزا تقسیم می شوند. در دسته بندی داده ها مجموعه ای از داده های آزمایشی تحلیل می شوند. برای مجموعه داده های آزمایشی برچسب کلاس ها مشخص است. برای هر کلاس داده های آزمایشی مدلی بر اساس خصوصیات داده ها ساخته می شود. حاصل عملیات دسته بندی می تواند یک درخت تصمیم یا مجموعه ای از قوانین دسته بندی باشد که برای فهم بهتر داده های موجود در پایگاه داده و همچنین دسته بندی داده هایی که در آینده به پایگاه داده اضافه می شوند به کار می رود. دسته بندی داده ها ارتباط تنگاتنگی با کاوش قواعد وابستگی دارد، به طوری که گاهی دسته بندی را به کمک قواعد وابستگی انجام می دهند.

به عنوان مثال برای فروشنده ی ماشین دسته بندی مشتریانش بر اساس تمایل و علاقه آنها به انواع مختلف ماشین مطلوبست به طوری که بتواند به مشتریانش خدمات بهتری ارائه دهد و

[12] Association Rules Mining
[13] Classification
[14] Clustering
[15] Prediction
[16] Time Series Analysis
[17] Relational
[18] Transactional

کاتالوگ های محصولات جدید مورد نظر آنها را برایشان بفرستد و درآمد حاصل از فـروش خـود را افزایش دهد.

خوشه بندی داده ها از مهمترین روش های داده کاوی به شمار می‌آید. درخوشه بندی مجموعه ای از داده‌ها گروه‌بندی می شوند. فرق خوشه بندی با دسته بندی داده‌ها در این است کـه در خوشـه بندی برخلاف دسته بندی تعداد کلاس ها در ابتدا مشخص نیستند. خوش بندی داده ها بر سـاس اصل مفهومی زیر صورت می گیرد:

"حداکثر کردن شباهت های بین اعضای هر کلاس و حداقل کردن شباهت ها بین اعضای مربـوط به کلاس های مختلف".

به عنوان مثالی از خوشه بندی، مجموعه ای از کالاها را می توان در ابتدا به صورت مجموعـه ای از کلاس های مختلف خوشه بندی کرده و سپس مجموعه ای از قوانین را بر اساس این چنین دسته بندی نتیجه گیری کرد.

پیش بینی یکی دیگر از تکنیک های داده کاوی محسوب می شود که در آن مقـادیر ممکـن بـرای متغیرهای نامعلوم پیش بینی می شوند. در پیشبینی ابتدا داده هایی که به متغیر نـامعلوم مربـوط هستند بوسیله ی برخی تحلیل های آماری پیدا می شوند. سـپس از برخـی روش هـای هوشـمند مانند شبکه های عصبی و الگوریتم ژنتیک برای انجام پیشبینی استفاده می شود. برای مثال مقدار حقوقی که یک کارمند می گیرد را می توان با استفاده از چگونگی توزیع حقوق کارمندان مشابه در آن اداره، پیشبینی کرد. روش های دیگری از جمله تحلیل رگرسیون[19] ، تحلیل وابستگی[20] ، درخت تصمیم[21] در انجام یک پیشبینی با کیفیت مؤثرند.

تحلیل سری های زمانی از دیگر روش های کاربردی داده کاوی است. در این روش حجم زیـادی از داده های سری زمانی برای یافتن خصوصیات جالب توجه و نظم های مشخص، تحلیل مـی شـوند. رخداد وقایع متوالی، مجموعه وقایعی که بعد از یک واقعه مشخص به وقوع می پیوندند، روند هـا و انحراف ها از جمله این نظم ها و پدیده های جالب توجه هستند. برای مثال می توان رونـد تغییـر

[19] Regression Analysis
[20] Correlation Analysis
[21] Decision Tree