

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

داده کاوی بار پدماینر - دسترسی به داده ها ، ترکیب و پاکسازی

تألیف:

شهاب طارقیان

برسام سوری

علی صدیقی منش

محمد صدیقی منش

مجتبی احمدی

سال ۱۳۹۸

سرشناسه: صدیقی منش، علی
عنوان و نام پدید آور: داده کاوی با رییدماینر- دسترسی به داده ها ، ترکیب و پاکسازی - تألیف: علی صدیقی منش- برسام سوری-
شهاب طارقیان- مجتبی احمدی- محمد صدیقی منش
مشخصات نشر: قم، الهام نور، ۱۳۹۸.
مشخصات ظاهری: ۴۷۳ ص، مصور، نمودار
شابک: ۹۷۸-۶۰۰-۶۹۴۷-۵۹-۴ / ۵۰۰۰۰ ریال
وضعیت فهرست نویسی: فیبا.
موضوع: ریید ماینر (منبع الکترونیکی)، داده کاوی، نرم افزار
موضوع: RapidMiner (Electronic resource)-Data Mining- Software
شابنامۀ افزوده: صدیقی منش، علی، ۱۳۶۶-
رده بندی کنگره: QA ۷۶ / ۹
رده بندی دیویی: ۰۰۶/۳۱۲
شماره کتابشناسی ملی: ۶۰۴۶۶۰۴

Contact:

Email: Ali.sedighimanesh@gmail.com, 09193508519

Email: Mohammad.sedighimanesh@gmail.com, 09193523413

انتشارات الهام نور

نام کتاب: داده کاوی با رییدماینر- دسترسی به داده ها ، ترکیب و پاکسازی
تألیف: علی صدیقی منش، برسام سوری، شهاب طارقیان، مجتبی احمدی، محمد صدیقی منش
تیراژ / چاپخانه: ۳۰۰ نسخه / شریعت
قطع / تعداد صفحات / صحافی: وزیری / ۴۷۳ صفحه / براتی
نوبت چاپ / قیمت: اول ۱۳۹۸ / ۵۰۰۰ تومان

شابک: ۹۷۸-۶۰۰-۶۹۴۷-۵۹-۴

کلیه حقوق محفوظ و مخصوص مؤلفین می باشد.

امروزه با پیشرفت فناوری و حضور گسترده‌ی آن در زندگی مردم، شاهد کاربرد پررنگ داده و اطلاعات هستیم. یکی از روش‌های استخراج اطلاعات از داده‌های خام داده‌کاوی یا Data Mining می‌باشد. به فرایند تجزیه و تحلیل داده‌ها و کشف الگوهای مفید در یک مجموعه‌ی داده، داده‌کاوی می‌گویند. در اطراف ما داده‌های زیادی وجود دارد که رشد آن‌ها بصورت چشم‌گیری در حال افزایش است. بنابراین ما نیاز به تصمیم‌گیری‌های کاربردی و هوشمندانه در زمینه‌های مختلف داریم تا الگوهای مورد نیاز خود را از این داده‌ها استخراج نماییم. در صورتی که با ابزارهای موجود در این زمینه آشنا نباشیم، این فرایند می‌تواند بسیار مشکل باشد.

تعداد بسیار زیادی ابزار مفید برای داده‌کاوی وجود دارد که با استفاده از آن‌ها می‌توانیم به اعمالی چون کشف رابطه‌ی میان داده‌ها، خوشه‌بندی داده‌ها، طبقه‌بندی آن‌ها و ... بپردازیم.

یکی از برترین نرم‌افزارها در حوزه داده‌کاوی رپیدماینر^۱ می‌باشد که این نرم‌افزار با زبان جاوا نوشته شده است. شاید بتوان یکی از مزایای این نرم‌افزار را عدم نیاز کاربران به کدنویسی دانست که کار را برای آن‌ها آسان نموده است. قابلیت‌ها و ابزارهای موجود در رپیدماینر به کاربران این امکان را می‌دهد تا به داده‌کاوی بر روی مجموعه داده‌ی خود بپردازند. همچنین رپیدماینر با فراهم آوردن الگوریتم‌های گسترده‌ی داده‌کاوی و یادگیری ماشین، شما را در راستای پیش‌بینی خروجی‌های آینده‌ی یاری می‌کند.

رپیدماینر نرم‌افزاری توانمند در مباحث داده‌کاوی، یادگیری ماشینی، تحلیل و پیش‌بینی و تحلیل کسب و کار می‌باشد. این برنامه کاربردهای زیادی هم در امر تجاری و کسب و کار و هم در امور تحقیقاتی، آموزشی، یادگیری و ... دارد. محققین و افراد علاقه‌مند به مباحث داده‌کاوی و یادگیری ماشینی می‌توانند با استفاده از این برنامه تمامی مراحل مورد نیاز را از آماده‌سازی اطلاعات اولیه گرفته تا بررسی کردن نتایج، ارزیابی و اعتبار سنجی و بهینه‌سازی خروجی را در یک محیط یکپارچه و واحد انجام دهند. هسته اصلی این نرم‌افزار به صورت متن باز و رایگان عرضه شده است که بر اساس آن محصولات رایگان و تجاری زیادی نوشته شده است.

این محصول در سال‌های ابتدایی عرضه یعنی سال ۲۰۰۱ با نام YALE (Yet Another Learning Environment) شناخته می‌شد. به مرور زمان و همزمان با افزایش قابلیت‌ها در سال‌های متمادی نهایتاً در سال ۲۰۰۷ نام نرم‌افزار به نام کنونی آن یعنی رپیدماینر (RapidMiner) تغییر پیدا کرد. این نرم‌افزار به جهت سادگی کارکرد آن در سال‌های اخیر توجه بسیاری از محققین و دانشجویان را به خود جلب کرده است و از طرفی مدل توسعه متن باز این محصول نیز باعث شده است تا برنامه‌نویسان تمایل بیشتری به توسعه آن داشته باشند. این نرم‌افزار از الگوریتم‌های مختلفی برای آماده‌سازی و مدل‌سازی اطلاعات استفاده می‌کند که همین طیف گسترده گزینه‌های قابل انتخاب باعث شده است تا پروژه‌های آکادمیک زیادی لاقلاً برای قسمت‌هایی از کارشان از این محصول قدرتمند استفاده کنند. محیط گرافیکی این برنامه به

^۱ RapidMiner

صورت مطلوب و زیبایی طراحی شده و کاربران بدون دانش برنامه نویسی می‌توانند به راحتی با قسمت‌های مختلف برنامه ارتباط برقرار کنند.

نقاط قوت نرم‌افزار رپیدماینر:

- (۱) ارائه گزارش و رونوشت از مراحل اجرای الگوریتم.
- (۲) ظاهر پرداخته و آراسته.
- (۳) نمای گرافیکی خوب
- (۴) قابلیت تطابق با فایل‌های خروجی بسیاری از نرم‌افزارها مانند Excel.
- (۵) امکان تصحیح و خطایابی بسیار سریع.
- (۶) وجود آموزش‌های ویدیویی مناسب که برای این نرم‌افزار تهیه شده و در وب قرار داده شده‌است.
- (۷) مستندات شامل راهنمای بسیاری از عملگرها در نرم‌افزار.
- (۸) امکان اجرای هم‌زمان الگوریتم‌های یادگیری متفاوت در نرم‌افزار و مقایسه آن‌ها با یک‌دیگر در ابزار در نظر گرفته شده‌است.

امکانات ویژه در این ابزار:

به دلیل پیاده سازی و توسعه این نرم‌افزار با استفاده از زبان جاوا، امکان کار در سیستم‌های عامل مختلف از جمله ویندوز، لینوکس و سیستم‌های مکintosh برای این نرم‌افزار وجود دارد.

امکانات متن‌کاوی نیز در این نرم‌افزار پیش‌بینی شده‌است.

کلیه الگوریتم‌های یادگیری مدل در نرم‌افزار داده‌کاوی وکا (WEKA) پس از به هنگام‌سازی رپیدماینر به نرم‌افزار اضافه خواهند شد.

فهرست مطالب

۳	 مقدمه کتاب
۲۲	 فصل اول - دسترسی به داده‌ها
۲۲	 Copy Repository Entry
۲۳	 Delete Repository Entry
۲۳	 Move Repository Entry
۲۴	 Rename Repository Entry
۲۵	 Retrieve
۲۷	 Store
۳۰	 ۱-۱ - فایل‌ها
۳۰	 ۱-۱ - ۱-۱ - فوآندن
۴۰	 ۱-۱ - ۲ نوشتن
۷۴	 ۱-۲ - پایگاه داده
۷۴	 Read Database
۷۸	 Update Database
۸۱	 Write Database
۸۴	 ۱-۳ NoSQL
۸۴	 ۱-۳ - ۱-۳ Cassandra
۹۲	 ۱-۳ - ۲ MongoDB
۹۷	 ۱-۳ - ۳ Solr
۱۰۳	 ۱-۴ - ۱-۴ کاربرد
۱۰۳	 Trigger Zapier

۱۰۵	Salesforce	۱- ۴- ۱
۱۱۲	Mozenda	۱- ۴- ۲
۱۱۳	Qlik	۱- ۴- ۳
۱۱۹	Splunk	۱- ۴- ۵
۱۲۰	Cloud Storage	۱- ۵- ۵
۱۲۰	Amazon S3	۱- ۵- ۱
۱۲۴	Azure Blob Storage	۱- ۵- ۲
۱۲۸	Azure Data Lake Storage	۱- ۵- ۳
۱۳۲	Dropbox	۱- ۵- ۴
۱۳۴	Google Storage	۱- ۵- ۵
۱۳۹	فصل دوم - ترکیب کردن	
۱۳۹	۱- ۲- صفات	
۱۳۹	Reorder Attributes	
۱۴۲	۲- ۱- ۱- نام ها و نقش ها	
۱۶۱	۲- ۱- ۲- انواع	
۲۳۰	۲- ۱- ۳- انتقال	
۲۵۵	۲- ۱- ۴- تولید	
۲۸۲	۲- ۲- مثال ها	
۲۸۲	۲- ۲- ۱- فیلتر کردن	
۲۸۸	۲- ۲- ۲- نمونه برداری	
۳۰۱	۲- ۲- ۳- مرتب سازی	
۳۰۵	۲- ۳- جداول	
۳۰۵	۲- ۳- ۱- گروه بندی	
۳۱۱	۲- ۳- ۲- چرخاندن	

۳۲۰	۲-۳-۳ اتصال
۳۳۳	۲-۴-۴ مقادیر
۳۳۳	Adjust Date
۳۸۱	فصل سوم - پاک‌سازی
۳۸۱	۳-۱-۱ نرمال‌سازی
۳۸۱	Normalize
۳۸۹	۳-۲-۲ طرف‌بندی
۳۸۹	Discretize by Binning
۴۱۷	۳-۳-۳ مقادیر مفقوده
۴۱۷	Declare Missing Value
۴۳۸	۳-۴-۴ تکراری‌ها
۴۳۸	Remove Duplicates
۴۴۳	۳-۵-۵ Outlierها
۴۴۳	Detect Outlier (COF)
۴۵۴	۳-۶-۶ کاهش ابعاد
۴۵۴	Fourier Transformation
۴۵۶	Generalized Hebbian Algorithm

فهرست اشکال

- شکل (۱): تکامل فن آوری سیستم پایگاه داده ۱۴
- شکل (۲): ما دارای داده های غنی اما اطلاعات ضعیف هستیم. ۱۶
- شکل (۳): داده کاوی جستجوی دانش (الگوهای پالپ) در داده های شما. ۱۷
- شکل (۴): داده کاوی به عنوان یک گام از فرایند کشف دانش ۱۹
- شکل (۵): معماری یک سیستم داده کاوی معمولی ۲۰
- شکل (۱-۱): فرآیند آموزشی "فراخوانی یک مجموعه مثال با استفاده از اپراتور Retrieve". ۲۶
- شکل (۲-۱): فرآیند آموزشی "ذخیره کردن یک مجموعه مثال با استفاده از اپراتور Store". ۲۸
- شکل (۳-۱): فرآیند آموزشی "ذخیره کردن یک مدل با استفاده از اپراتور Store". ۲۸
- شکل (۴-۱): فرآیند آموزشی "استفاده از اشیائی که اپراتور Store ذخیره کرده است". ۲۹
- شکل (۵-۱): فرآیند آموزشی "اصول ARFF". ۳۲
- شکل (۶-۱): فرآیند آموزشی "خواندن یک فایل ARFF با کمک اپراتور Read ARFF". ۳۳
- شکل (۷-۱): فرآیند آموزشی "نوشتن و سپس خواندن داده ها از یک پایگاه داده Access". ۳۵
- شکل (۸-۱): فرآیند آموزشی "خواندن یک فایل CSV". ۴۲
- شکل (۹-۱): فرآیند آموزشی "خواندن یک ExampleSet از یک فایل Excel". ۴۸
- شکل (۱۰-۱): فرآیند آموزشی "کاربرد اپراتور Read SAS". ۴۹
- شکل (۱۱-۱): فرآیند آموزشی "خواندن یک فایل SPSS". ۵۱
- شکل (۱۲-۱): فرآیند آموزشی "نوشتن و خواندن یک فایل sparse". ۵۳
- شکل (۱۳-۱): فرآیند آموزشی "نوشتن و خواندن یک فایل XFF". ۵۹
- شکل (۱۴-۱): فرآیند آموزشی "اصول ARFF". ۶۲
- شکل (۱۵-۱): فرآیند آموزشی "خواندن یک فایل ARFF با کمک اپراتور Read ARFF". ۶۲
- شکل (۱۶-۱): فرآیند آموزشی "نوشتن و سپس خواندن داده ها از یک پایگاه داده Access". ۶۴
- شکل (۱۷-۱): فرآیند آموزشی "نوشتن مجموعه داده Labor-Negotiations در یک فایل CSV". ۶۶
- شکل (۱۸-۱): فرآیند آموزشی نوشتن مجموعه داده Labor-Negotiation در یک فایل Excel. ۶۸
- شکل (۱۹-۱): فرآیند آموزشی نوشتن مجموعه داده label دار یک فرمت تعریف شده توسط کاربر. ۷۲
- شکل (۲۰-۱): فرآیند آموزشی خواندن ExampleSet از یک پایگاه داده MySQL. ۷۶
- شکل (۲۱-۱): فرآیند آموزشی به روز رسانی یک ExampleSet در یک پایگاه داده MySQL. ۸۰
- شکل (۲۲-۱): فرآیند آموزشی نوشتن یک ExampleSet در یک پایگاه داده MySQL. ۸۳
- شکل (۲-۱): فرآیند آموزشی "انتخاب صفات با تعیین عبارات منظمی که بر نام آنها منطبق می شود". ۱۴۱

- شکل (۲-۲): فرآیند آموزشی "معاوضه نقشهای صفاتی از مجموعه داده Golf". ۱۴۳.....
- شکل (۲-۳): فرآیند آموزشی "تغییر نام چندین صفت". ۱۴۴.....
- شکل (۲-۴): فرآیند آموزشی "تغییر نام صفات با استفاده از توضیحات سافتمانی آنها". ۱۴۶.....
- شکل (۲-۵): فرآیند آموزشی "تغییر نام تمامی صفات با استفاده از مقادیر مثال". ۱۴۸.....
- شکل (۲-۶): فرآیند آموزشی "تغییر نام صفات مجموعه داده Sonar". ۱۵۳.....
- شکل (۲-۷): فرآیند آموزشی "تغییر نام صفات مجموعه داده Sonar". ۱۵۷.....
- شکل (۲-۸): فرآیند آموزشی "اپراتور Set Role روی مجموعه داده Titanic". ۱۶۰.....
- شکل (۲-۹): فرآیند آموزشی "مقدمه‌ای بر اپراتور Date to Nominal". ۱۶۶.....
- شکل (۲-۱۰): فرآیند آموزشی "مقدمه‌ای بر اپراتور Date to Numerical". ۱۶۹.....
- شکل (۲-۱۱): فرآیند آموزشی "تغییر مقادیر عددی به مبالغ پولی". ۱۷۵.....
- شکل (۲-۱۲): فرآیند آموزشی "مدس زدن نوع یک صفت پس از پیش‌پردازش". ۱۷۹.....
- شکل (۲-۱۳): فرآیند آموزشی "تبدیل صفات مجموعه داده Golf از اسمی به دو اسمی". ۱۸۴.....
- شکل (۲-۱۴): فرآیند آموزشی "مقدمه‌ای بر اپراتور Nominal to Date". ۱۹۰.....
- شکل (۲-۱۵): فرآیند آموزشی "تبدیل اسمی به عددی از طریق انواع مختلف کدگذاری". ۱۹۷.....
- شکل (۲-۱۶): فرآیند آموزشی "بک‌گیری اپراتور Nominal to Text روی مجموعه داده Golf". ۲۰۱.....
- شکل (۲-۱۷): فرآیند آموزشی "تبدیل صفات عددی در مجموعه داده Sonar به صفات دو اسمی". ۲۰۶.....
- شکل (۲-۱۸): فرآیند آموزشی "تبدیل صفات عددی در مجموعه داده Sonar به صفات چند اسمی". ۲۱۱.....
- شکل (۲-۱۹): فرآیند آموزشی "تبدیل صفات مجموعه داده Golf از عدد صمیم به نوع حقیقی". ۲۱۵.....
- شکل (۲-۲۰): فرآیند آموزشی "تبدیل اسمی به عددی با کمک اپراتور Parse Numbers". ۲۲۰.....
- شکل (۲-۲۱): فرآیند آموزشی "تبدیل صفات مجموعه داده Iris از نوع حقیقی به عدد صمیم". ۲۲۵.....
- شکل (۲-۲۲): فرآیند آموزشی "مقدمه‌ای بر اپراتور Text to Nominal". ۲۲۹.....
- شکل (۲-۲۳): فرآیند آموزشی "مذف دو صفت اول از مجموعه داده Golf". ۲۳۱.....
- شکل (۲-۲۴): فرآیند آموزشی "مذف صفات همبسته از مجموعه داده Sonar". ۲۳۴.....
- شکل (۲-۲۵): فرآیند آموزشی "مذف صفات اسمی بی‌استفاده از یک ExampleSet". ۲۳۷.....
- شکل (۲-۲۶): فرآیند آموزشی "مذف صفات عددی بی‌استفاده از یک ExampleSet". ۲۳۸.....
- شکل (۲-۲۷): فرآیند آموزشی "انتخاب صفات از نمونه داده‌های Titanic". ۲۴۳.....
- شکل (۲-۲۸): فرآیند آموزشی "کامبردهای مختلف اپراتور Select Attributes". ۲۴۴.....
- شکل (۲-۲۹): فرآیند آموزشی "انتخاب صفات با کمک عبارات منظم". ۲۴۵.....
- شکل (۲-۳۰): فرآیند آموزشی "انتخاب صفات تصادفی از مجموعه داده Sonar". ۲۴۶.....
- شکل (۲-۳۱): فرآیند آموزشی "انتخاب صفات از مجموعه داده Sonar". ۲۴۶.....
- شکل (۲-۳۲): فرآیند آموزشی "کار کردن روی زیرمجموعه‌ای از مجموعه داده Golf". ۲۵۴.....

- شکل (۲-۳۳): فرآیند آموزشی "مقادیر مطلق مجموعه داده". Ripley-Set..... ۲۵۸
- شکل (۲-۳۴): فرآیند آموزشی "تولید یک صفت که دارای مقدار میانگین صفات مقیاس از مجموعه داده Sonar است". ۲۶۳
- شکل (۲-۳۵): فرآیند آموزشی "تولید صفات از طریق function descriptions مختلف". ۲۶۷
- شکل (۲-۳۶): فرآیند آموزشی "تولید یک صفت ترکیبی در مجموعه داده Labor-Negotiations". ۲۶۹
- شکل (۲-۳۷): فرآیند آموزشی "تولید یک کپی از صفت Temperature در مجموعه داده Golf". ۲۷۱
- شکل (۲-۳۸): فرآیند آموزشی "اضافه کردن یک صفت فالی به مجموعه داده Golf". ۲۷۲
- شکل (۲-۳۹): فرآیند آموزشی "استفاده کردن از تابع power برای تولید صفات". ۲۷۵
- شکل (۲-۴۰): فرآیند آموزشی "باطل کردن صفت id در مجموعه داده Iris". ۲۷۶
- شکل (۲-۴۱): فرآیند آموزشی "تولید ماصلفرب صفات". ۲۷۸
- شکل (۲-۴۲): فرآیند آموزشی "مقدمه‌ای بر اپراتور Generate TFIDF". ۲۷۹
- شکل (۲-۴۳): فرآیند آموزشی "تخصیص وزنها طوری که مجموع اوزان در همه برپسب‌ها یکی باشد". ۲۸۱
- شکل (۲-۴۴): فرآیند آموزشی "فیلتر کردن مثال‌ها با استفاده از پارامتر invert filter". ۲۸۳
- شکل (۲-۴۵): فرآیند آموزشی "فیلتر کردن مثال‌ها با استفاده از custom_filters". ۲۸۶
- شکل (۲-۴۶): فرآیند آموزشی "فیلتر کردن مثال‌ها با استفاده از attribute_value_filter". ۲۸۷
- شکل (۲-۴۷): فرآیند آموزشی "فیلتر کردن مثال‌ها با استفاده از expression". ۲۸۷
- شکل (۲-۴۸): فرآیند آموزشی "نمونه‌برداری از مجموعه داده Ripley-Set". ۲۹۰
- شکل (۲-۴۹): فرآیند آموزشی "نمونه‌برداری فوراها اندازه از مجموعه داده Golf". ۲۹۳
- شکل (۲-۵۰): فرآیند آموزشی "نمونه‌برداری Kernard-Stone از مجموعه داده Iris". ۲۹۵
- شکل (۲-۵۱): فرآیند آموزشی "نمونه‌برداری لایه‌ای از مجموعه داده Golf". ۲۹۷
- شکل (۲-۵۲): فرآیند آموزشی "ایجاد پارتیشن‌هایی از مجموعه داده Golf با استفاده از اپراتور Split Data". ۳۰۰
- شکل (۲-۵۳): فرآیند آموزشی "بهم ریختن مجموعه داده Iris". ۳۰۲
- شکل (۲-۵۴): فرآیند آموزشی "مرتب کردن مجموعه داده Golf بر اساس Temperature". ۳۰۳
- شکل (۲-۵۵): فرآیند آموزشی "مرتب کردن بر اساس چندین صفت". ۳۰۴
- شکل (۲-۵۶): فرآیند آموزشی "تقلید کردن از یک مستجوبی تمیمی SQL با استفاده از اپراتور Aggregate". ۳۱۰
- شکل (۲-۵۷): فرآیند آموزشی "ادغام کردن چندین صفت که مشخصات یکسانی را اندازه می‌گیرند در یک صفت واحد". ۳۱۳
- شکل (۲-۵۸): فرآیند آموزشی "مقدمه‌ای بر اپراتور Pivot". ۳۱۶
- شکل (۲-۵۹): فرآیند آموزشی "سناریوهای مختلف Transpose". ۳۱۹
- شکل (۲-۶۰): فرآیند آموزشی "ترکیب کردن مجموعه‌های داده Golf و Golf-Testset". ۳۲۱
- شکل (۲-۶۱): فرآیند آموزشی "اشتراک دو ExampleSet". ۳۲۳

- شکل (۲-۶): فرآیند آموزشی "بررسی کردن انواع مختلف اتصال و پیوند". ۳۷۰
- شکل (۲-۶): فرآیند آموزشی "مقدمه‌ای بر اپراتور Set Minus". ۳۷۲
- شکل (۲-۶): فرآیند آموزشی "تولید ابرمجموعه‌هایی از مجموعه‌های داده Iris و Golf". ۳۷۳
- شکل (۲-۶): فرآیند آموزشی "اتحاد مجموعه‌های داده Golf و Golf-Testset". ۳۷۳
- شکل (۲-۶): فرآیند آموزشی "اتحاد مجموعه‌های داده Iris و Golf". ۳۷۴
- شکل (۲-۶): فرآیند آموزشی "اتحاد مجموعه‌های داده Golf (با صفت id) و Iris". ۳۷۴
- شکل (۲-۶): فرآیند آموزشی "ایجاد چندین نوع تنظیم در یک صفت تاریخ". ۳۷۵
- شکل (۲-۶): فرآیند آموزشی "بکار بردن اپراتور Cut روی label مجموعه داده Iris". ۳۷۵
- شکل (۲-۶): فرآیند آموزشی "نکاشت کردن چندین مقدار". ۳۷۵
- شکل (۲-۶): فرآیند آموزشی "مقدمه‌ای بر اپراتور Merge". ۳۷۶
- شکل (۲-۶): فرآیند آموزشی "تغییر نکاشت صفت Wind از مجموعه داده Golf". ۳۷۶
- شکل (۲-۶): فرآیند آموزشی "استفاده از پارامترهای replace what و replace by". ۳۷۶
- شکل (۲-۶): فرآیند آموزشی "جایگزین کردن زیررشته‌ها با استفاده از یک لغت‌نامه". ۳۷۷
- شکل (۲-۶): فرآیند آموزشی "مقدمه‌ای بر اپراتور Set Data". ۳۷۷
- شکل (۲-۶): فرآیند آموزشی "تقسیم‌های مرتب و نامرتب". ۳۷۷
- شکل (۲-۶): فرآیند آموزشی "هدف کردن خواص فالی از ابتدا و انتهای مقادیر صفات". ۳۷۸
- شکل (۳-۱): فرآیند آموزشی "نرمال‌سازی صفات Age و Passenger Fare برای مجموعه داده Titanic". ۳۸۶
- شکل (۳-۲): فرآیند آموزشی "بکار بردن اپراتور Scale by Weights روی مجموعه داده Golf". ۳۸۸
- شکل (۳-۳): فرآیند آموزشی "میداسازی صفات عددی از مجموعه داده Golf با کمک ظرف‌بندی". ۳۹۴
- شکل (۳-۴): فرآیند آموزشی "میداسازی و تقسیم مجموعه داده Sonar با کمک آنالوژی". ۴۰۰
- شکل (۳-۵): فرآیند آموزشی "میداسازی و تقسیم صفت Temperature از مجموعه داده Golf با کمک فراوانی". ۴۰۵
- شکل (۳-۶): فرآیند آموزشی "میداسازی و تقسیم صفت Temperature از مجموعه داده Golf". ۴۱۱
- شکل (۳-۷): فرآیند آموزشی "میداسازی و تقسیم صفات عددی از مجموعه داده Golf". ۴۱۶
- شکل (۳-۸): فرآیند آموزشی "اعلان یک مقدار اسمی بعنوان مقدار مفقوده". ۴۲۰
- شکل (۳-۹): فرآیند آموزشی "مقدمه‌ای بر اپراتور Fill Data Gaps". ۴۲۲
- شکل (۳-۱۰): فرآیند آموزشی "استفاده از طرح K-NN برای تمییز زدن مقادیر مفقوده". ۴۲۷
- شکل (۳-۱۱): فرآیند آموزشی "جایگزین کردن مقادیر نامتناهی با مقادیر مفقوده". ۴۳۲
- شکل (۳-۱۲): فرآیند آموزشی "جایگزین کردن مقادیر مفقوده در مجموعه داده Labor Negotiations". ۴۳۷
- شکل (۳-۱۳): فرآیند آموزشی "هدف کردن مقادیر تکراری از مجموعه داده Golf بر اساس صفات Wind و Outlook". ۴۴۱
- شکل (۳-۱۴): فرآیند آموزشی "کشف outliera از یک ExampleSet". ۴۴۶

- شکل (۳-۱۵): فرآیند آموزشی "کشف outlierها از یک ExampleSet" ۴۴۸
- شکل (۳-۱۶): فرآیند آموزشی "کشف outlierها از یک ExampleSet" ۴۵۰
- شکل (۳-۱۷): فرآیند آموزشی "کشف outlierها از یک ExampleSet" ۴۵۳
- شکل (۳-۱۸): فرآیند آموزشی "مقدمه‌ای بر اپراتور Fourier Transformation" ۴۵۵
- شکل (۳-۱۹): فرآیند آموزشی "کاهش ابعاد مجموعه داده Polynomial با استفاده از اپراتور GHA" ۴۵۸
- شکل (۳-۲۰): فرآیند آموزشی "کاهش ابعاد مجموعه داده Sonar با استفاده از اپراتور Independent Component Analysis" ۴۶۱
- شکل (۳-۲۱): فرآیند آموزشی "کاهش ابعاد مجموعه داده Polynomial با استفاده از اپراتور Principal Component Analysis" ۴۶۴
- شکل (۳-۲۲): فرآیند آموزشی "مقدمه‌ای بر اپراتور Principal Component Analysis (Kernel)" ۴۶۸
- شکل (۳-۲۳): فرآیند آموزشی "کاهش ابعاد مجموعه داده Sonar با استفاده از اپراتور Self-Organizing Map" ۴۷۰
- شکل (۳-۲۴): فرآیند آموزشی "کاهش ابعاد مجموعه داده Sonar با استفاده از اپراتور Singular Value Decomposition" ۴۷۳

مقدمه ای بر فصل اول

با توجه به دسترسی گسترده به مقادیر بسیار عظیمی از داده‌ها و نیاز قریب‌الوقوع برای تبدیل داده به اطلاعات^۱ مفید و دانش، در سال‌های اخیر، داده کاوی توجه زیادی را در صنعت اطلاعات و در جامعه به خود جلب نموده است. اطلاعات و دانش به دست آمده می‌تواند برای کاربردهای مختلف از تحلیل بازار، تشخیص تقلب، و حفظ مشتری، کنترل تولید و اکتشاف علم استفاده شود. داده کاوی می‌تواند به عنوان نتیجه‌ی چرخه طبیعی فناوری اطلاعات دیده شود. صنعت سیستم پایگاه داده، شاهد یک مسیر تکاملی در توسعه (پیشرفت) عوامل زیر بود (شکل ۱):

مجموعه داده و ایجاد پایگاه داده مدیریت داده (از جمله ذخیره سازی و بازیابی اطلاعات^۲ و پردازش تراکنشی^۳ پایگاه داده)

داده کاوی را می‌توان به عنوان نتیجه تکامل طبیعی فناوری اطلاعات دید. چنانکه در شکل زیر مشهود است، صنعت پایگاه داده و مدیریت داده طی توسعه برخی کارکردهای حیاتی فناوری اطلاعات تکامل یافته‌اند. گردآوری داده‌ها و ساخت پایگاه داده، مدیریت داده (شامل ذخیره‌سازی و بازیابی داده و پردازش تراکنش‌های پایگاه داده) و تحلیل پیشرفته داده (شامل انبار داده و داده کاوی)، از این جمله هستند. تکامل اولیه مکانیزم‌های گردآوری داده و ساخت پایگاه داده به‌عنوان پیش‌شرطی برای توسعه‌های آتی مکانیزم‌های موثر جهت ذخیره‌سازی و بازیابی داده‌ها، پردازش کوئری و تراکنش محسوب می‌شود. امروزه، سیستم‌های پایگاه داده متعددی امکان پردازش کوئری و تراکنش را به‌عنوان یک قابلیت معمول فراهم می‌کنند. تحلیل‌های پیشرفته داده طی زمان کام بعدی محسوب می‌شود.

^۱ information

^۲ Information Retrieval

^۳ Transaction processing