

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

داده کړی با ریدماینر - مدلسازی داده

تألیف:

علی صدیقی منش برسام سوری محمد صدیقی منش

سارا فتحی پور دیمان شهاب طارقیان

سال ۱۳۹۸

سرشناسه: صدیقی منش، علی
عنوان و نام پدید آور: داده کاوی با رپیدماینر - مدلسازی داده - تألیف: علی صدیقی منش - برسام سوری - محمد صدیقی منش -
سارا فتحی پور دیمان - شهاب طارقیان

مشخصات نشر: قم، الهام نور، ۱۳۹۸.

مشخصات ظاهری: ۴۰۲ ص، مصور، نمودار

شابک: ۹۷۸-۶۰۰-۶۹۴۷-۶۰-۰ / ۴۵۰۰۰ ریال

وضعیت فهرست نویسی: فیا.

موضوع: رپید ماینر (منبع الکترونیکی)، داده کاوی، نرم افزار

موضوع: RapidMiner (Electronic resource)-Data Mining- Software

شماره کتبی افزوده: صدیقی منش، علی، ۱۳۶۶-

رده بندی کنگره: ۷۶/۹ OA

رده بندی دیویی: ۰۰۴۳۱۲

شماره کتبی شناسی ملی: ۶۰۴۶۶۰۳

Contact:

Email: Ali.sedighimanesh@gmail.com, 09193508510

Email: Mohammad.sedighimanesh@gmail.com, 09193523413

انتشارات الهام نور

نام کتاب: داده کاوی با رپیدماینر - مدلسازی داده

تألیف: علی صدیقی منش - برسام سوری - محمد صدیقی منش - سارا فتحی پور دیمان - شهاب
طارقیان

تیراژ / چاپخانه: ۳۰۰ نسخه / شریعت

قطع / تعداد صفحات / صحافی: وزیری / ۲۰۲ صفحه / براتی

نوبت چاپ / قیمت: اول / ۱۳۹۸ / ۴۵۰۰۰ تومان

شابک: ۹۷۸-۶۰۰-۶۹۴۷-۶۰-۰

کلیه حقوق محفوظ و مخصوص مؤلفین می باشد.

فهرست کتاب

۱۳	فصل چهارم - مدل سازی
۱۳	۴-۱- پیشگویانه
۱۳	Create Formula
۱۷	۴-۱-۱- تنبل (Lazy)
۲۹	۴-۱-۲- بیزی (Bayesian)
۳۵	۴-۱-۳- درخت ها
۷۲	۴-۱-۴- قوانین
۸۱	۴-۱-۵- شبیه های عمیق
۹۸	۴-۱-۶- توابع
۱۲۴	۴-۱-۷- رگرسیون ایستیک
۱۳۵	۴-۱-۸- ماشین های بردار پشتیبانی
۱۶۱	۴-۱-۹- تحلیل تشفیصی
۱۷۰	۴-۱-۱۰- روش های دسته جمعی
۱۹۳	۴-۲- قطعه بندی
۱۹۳	Agglomerative Clustering
۱۹۸	Cluster Model Visualizer
۲۰۰	DBSCAN
۲۰۵	Expectation Maximization Clustering

۲۰۹.....	Extract Cluster Prototypes
۲۱۸.....	Support Vector Clustering
۲۲۲.....	Top Down Clustering
۲۲۵.....	k-Means
۲۳۵.....	K-Means (Kernel)
۲۴۰.....	K-Medoids
۲۴۶.....	۴-۳- وابستگی ها
۲۴۶.....	Apply Association Rules
۲۴۹.....	Create Association Rules
۲۵۲.....	FP-Growth
۲۶۲.....	Generalized Sequential Patterns
۲۶۵.....	۴-۴ همبستگی ها
۲۷۹.....	Mutual Information Matrix
۲۸۱.....	۴-۵- شباهت ها
۲۸۱.....	Cross Distances
۲۸۶.....	Data to Similarity
۲۹۴.....	Similarity to Data
۲۹۶.....	۴-۶- وزن صفات
۲۹۶.....	Data to Weights

۲۹۸	Weight by Chi Squared Statistic
۳۰۰	Weight by Component Model
۳۰۴	Weight by Deviation
۳۰۸	Weight by Gini Index
۳۱۰	Weight by Information Gain
۳۱۴	Weight by PCA
۳۱۷	Weight by Relief
۳۱۹	Weight by Rule
۳۲۱	Weight by SVM
۳۲۵	Weight by Uncertainty
۳۲۷	Weight by User Specification
۳۳۰	Weight by Value Average
۳۳۲	Weights to Data
۳۳۴	 ۴-۷- بهینه‌سازی
۳۳۴	 ۴-۷-۱ پارامترها
۳۵۳	 ۴-۷-۲ انتخاب صفات
۳۷۷	 ۴-۷-۳ تولید صفات
۳۹۴	 ۴-۷-۴ وزن‌دهی صفات

فهرست اشکال

شکل (۴-۱): فرآیند آموزشی "فرمول مدل Logistic Regression".....	۱۴
شکل (۴-۲): فرآیند آموزشی "گروه‌بندی مدل‌ها و بکارگیری مدل گروه‌بندی شده حاصله".....	۱۶
شکل (۴-۳): فرآیند آموزشی "استفاده از اپراتور Default Model با روش mode".....	۱۸
شکل (۴-۴): فرآیند آموزشی "طبقه‌بندی مجموعه داده Golf-Testset با استفاده از اپراتور k-NN".....	۲۷
شکل (۴-۵): فرآیند آموزشی "تفاوت کارایی بین داده‌های نرمال شده و داده‌های نرمال نشده".....	۲۷
شکل (۴-۶): فرآیند آموزشی "بهینه‌سازی k در یک مدل k-NN و ثبت کردن تمامی نتایج".....	۲۸
شکل (۴-۷): فرآیند آموزشی "بکار بردن بیز ساده روی مجموعه داده Iris".....	۳۱
شکل (۴-۸): فرآیند آموزشی "مقدمه‌ای بر اپراتور Naive Bayes (Kernel)".....	۳۴
شکل (۴-۹): فرآیند آموزشی "مقدمه‌ای بر اپراتور CHAID".....	۳۸
شکل (۴-۱۰): فرآیند آموزشی "مقدمه‌ای بر اپراتور Decision Stump".....	۴۰
شکل (۴-۱۱): فرآیند آموزشی "آموزش دادن یک مدل درخت تصمیم".....	۴۴
شکل (۴-۱۲): فرآیند آموزشی "آموزش دادن یک مدل درخت تصمیم و استفاده از آن برای پیش‌بینی نتیجه".....	۴۵
شکل (۴-۱۳): فرآیند آموزشی "رگرسیون".....	۴۵
شکل (۴-۱۴): فرآیند آموزشی "مقدمه‌ای بر اپراتور Decision Tree (Multiway)".....	۴۷
شکل (۴-۱۵): فرآیند آموزشی "مقدمه‌ای بر اپراتور Decision Tree (Weight-Based)".....	۵۰
شکل (۴-۱۶): فرآیند آموزشی "طبقه‌بندی با استفاده از GBT".....	۵۶
شکل (۴-۱۷): فرآیند آموزشی "طبقه‌بندی با اپراتور Split Validation و با استفاده از GBT".....	۵۷
شکل (۴-۱۸): فرآیند آموزشی "رگرسیون با استفاده از GBT".....	۵۷
شکل (۴-۱۹): فرآیند آموزشی "شروع کار با ID3".....	۶۱
شکل (۴-۲۰): فرآیند آموزشی "تولید یک مجموعه از درخت‌های تصادفی با استفاده از اپراتور Random".....	۶۶
شکل (۴-۲۱): فرآیند آموزشی "جنگل تصادفی برای رگرسیون".....	۶۶

شکل (۴-۲۲):	فرآیند آموزشی "مقایسه بین درخت تصمیم و جنگل تصادفی"	۶۷
شکل (۴-۲۳):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Random Tree"	۷۱
شکل (۴-۲۴):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Rule Induction"	۷۴
شکل (۴-۲۵):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Subgroup Discovery"	۷۸
شکل (۴-۲۶):	فرآیند آموزشی "تولید قوانین از یک درخت تصمیم"	۸۰
شکل (۴-۲۷):	فرآیند آموزشی "طبقه‌بندی با اپراتور Split Validation و با استفاده از یادگیری عمیق"	۹۱
شکل (۴-۲۸):	فرآیند آموزشی "رگرسیون با استفاده از یادگیری عمیق"	۹۱
شکل (۴-۲۹):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Neural Net"	۹۵
شکل (۴-۳۰):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Perceptron"	۹۷
شکل (۴-۳۱):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Gaussian Process"	۱۰۰
شکل (۴-۳۲):	فرآیند آموزشی "طبقه‌بندی با استفاده از GLM"	۱۰۸
شکل (۴-۳۳):	فرآیند آموزشی "رگرسیون با استفاده از GLM"	۱۰۸
شکل (۴-۳۴):	فرآیند آموزشی "بکار بردن اپراتور Linear Regression در مجموعه داده Polynomial"	۱۱۲
شکل (۴-۳۵):	فرآیند آموزشی "بکار بردن اپراتور Local Polynomial Regression در مجموعه داده Polynomial"	۱۱۵
شکل (۴-۳۶):	فرآیند آموزشی "بکار بردن اپراتور Polynomial Regression در مجموعه داده Polynomial"	۱۱۹
شکل (۴-۳۷):	فرآیند آموزشی "مقدمه‌ای بر اپراتور RVM"	۱۲۲
شکل (۴-۳۸):	فرآیند آموزشی "بکار بردن اپراتور Vector Linear Regression در مجموعه داده Polynomial"	۱۲۴
شکل (۴-۳۹):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Logistic Regression"	۱۲۸
شکل (۴-۴۰):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Logistic Regression"	۱۳۴
شکل (۴-۴۱):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Fast Large Margin"	۱۳۷
شکل (۴-۴۲):	فرآیند آموزشی "شروع کار با SVM"	۱۴۵

شکل (۴-۱۴):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Support Vector Machine (Evolutionary)"	۱۵۱...
شکل (۴-۱۴):	فرآیند آموزشی "SVM با کرنل rbf"	۱۵۶.....
شکل (۴-۱۴):	فرآیند آموزشی "مقدمه‌ای بر اپراتور SVM (PSO)"	۱۶۱.....
شکل (۴-۱۴):	فرآیند آموزشی "مقدمه‌ای بر اپراتور LDA"	۱۶۴.....
شکل (۴-۱۴):	فرآیند آموزشی "مقدمه‌ای بر اپراتور QDA"	۱۶۶.....
شکل (۴-۱۴):	فرآیند آموزشی "مقدمه‌ای بر اپراتور RDA"	۱۶۹.....
شکل (۴-۱۴):	فرآیند آموزشی "استفاده از اپراتور AdaBoost برای تولید یک درخت تصمیم بهتر"	۱۷۲.....
شکل (۴-۵۰):	فرآیند آموزشی "استفاده از اپراتور Bagging برای تولید یک درخت تصمیم بهتر"	۱۷۶.....
شکل (۴-۵۱):	فرآیند آموزشی "استفاده از اپراتور Bayesian Boosting برای تولید یک درخت تصمیم بهتر"	۱۸۰.....
شکل (۴-۵۲):	فرآیند آموزشی "استفاده از اپراتور Linear Regression برای طبقه‌بندی"	۱۸۲.....
شکل (۴-۵۳):	فرآیند آموزشی "استفاده از اپراتور MetaCost برای تولید یک درخت تصمیم بهتر"	۱۸۵.....
شکل (۴-۵۴):	فرآیند آموزشی "استفاده از اپراتور SVM برای طبقه‌بندی چندجمله‌ای"	۱۸۷.....
شکل (۴-۵۵):	فرآیند آموزشی "مقدمه‌ای بر Stacking (انباشتن)"	۱۹۱.....
شکل (۴-۵۶):	فرآیند آموزشی "استفاده از اپراتور Vote برای طبقه‌بندی"	۱۹۲.....
شکل (۴-۵۷):	فرآیند آموزشی "فوشه‌بندی تجمعی مجموعه داده Ripley-Set"	۱۹۸.....
شکل (۴-۵۸):	فرآیند آموزشی "تجسم‌سازی فوشه برای Iris"	۱۹۹.....
شکل (۴-۵۹):	فرآیند آموزشی "فوشه‌بندی مجموعه داده Ripley-Set با استفاده از اپراتور DBSCAN"	۲۰۵.....
شکل (۴-۶۰):	فرآیند آموزشی "فوشه‌بندی مجموعه داده Ripley-Set با استفاده از اپراتور Expectation Maximization Clustering"	۲۰۹.....
شکل (۴-۶۱):	فرآیند آموزشی "استخراج Centroid Table پس از بکارگیری اپراتور K-Means روی Ripley-Set"	۲۱۱.....
شکل (۴-۶۲):	فرآیند آموزشی "مسطح‌سازی مدل فوشه‌بندی تجمعی"	۲۱۴.....
شکل (۴-۶۳):	فرآیند آموزشی "فوشه‌بندی تصادفی روی مجموعه داده Ripley-Set"	۲۱۷.....

شکل (۴-۶۴): فرآیند آموزشی "فوشه‌بندی مجموعه داده Ripley-Set با کمک اپراتور Support Vector	۲۲۱
....."Clustering	
شکل (۴-۶۵): فرآیند آموزشی "فوشه‌بندی به روش بالا به پایین روی مجموعه داده Ripley-Set".	۲۲۴
شکل (۴-۶۶): فرآیند آموزشی "فوشه‌بندی مجموعه داده Iris".	۲۳۳
شکل (۴-۶۷): فرآیند آموزشی "فوشه‌بندی مجموعه داده Ripley-Set با استفاده از اپراتور Kernel K	
....."Means	۲۳۹
شکل (۴-۶۸): فرآیند آموزشی "فوشه‌بندی مجموعه داده Ripley-Set با استفاده از اپراتور K-Medoids".	۲۴۵
شکل (۴-۶۹): فرآیند آموزشی "بکار بردن قوانین وابستگی".	۲۴۸
شکل (۴-۷۰): فرآیند آموزشی "مقدمه‌ای بر اپراتور Create Association Rules".	۲۵۱
شکل (۴-۷۱): فرآیند آموزشی "مقدمه‌ای بر اپراتور FP-Growth".	۲۵۸
شکل (۴-۷۲): فرآیند آموزشی "فالت‌های ورودی برای اپراتور FP-Growth".	۲۶۰
شکل (۴-۷۳): فرآیند آموزشی "مقدمه‌ای بر اپراتور GSP".	۲۶۵
شکل (۴-۷۴): فرآیند آموزشی "ماتریس ANOVA برای مجموعه داده Golf".	۲۶۷
شکل (۴-۷۵): فرآیند آموزشی "ماتریس همبستگی برای مجموعه داده Golf".	۲۷۲
شکل (۴-۷۶): فرآیند آموزشی "ماتریس کوواریانس برای مجموعه داده Polynomial".	۲۷۵
شکل (۴-۷۷): فرآیند آموزشی "Grouped ANOVA برای مجموعه داده Golf".	۲۷۸
شکل (۴-۷۸): فرآیند آموزشی "ماتریس اطلاعات متقابل برای مجموعه داده Polynomial".	۲۸۰
شکل (۴-۷۹): فرآیند آموزشی "مقدمه‌ای بر اپراتور Cross Distances".	۲۸۵
شکل (۴-۸۰): فرآیند آموزشی "مقدمه‌ای بر اپراتور Data to Similarity".	۲۹۰
شکل (۴-۸۱): فرآیند آموزشی "مقدمه‌ای بر اپراتور Data to Similarity Data".	۲۹۴
شکل (۴-۸۲): فرآیند آموزشی "مقدمه‌ای بر اپراتور Similarity to Data".	۲۹۵
شکل (۴-۸۳): فرآیند آموزشی "تولید بردار وزنی که وزن همه صفات در آن برابر با ۱.۰ است".	۲۹۷
شکل (۴-۸۴): فرآیند آموزشی "محاسبه کردن وزن صفات مجموعه داده Golf".	۲۹۹

شکل (۴- ۸۵):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Sonar با کمک PCA".	۳۰۲
شکل (۴- ۸۶):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Polynomial".	۳۰۵
شکل (۴- ۸۷):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Polynomial".	۳۰۷
شکل (۴- ۸۸):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Golf".	۳۰۹
شکل (۴- ۸۹):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Golf".	۳۱۱
شکل (۴- ۹۰):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Golf".	۳۱۳
شکل (۴- ۹۱):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Sonar با کمک PCA".	۳۱۶
شکل (۴- ۹۲):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Polynomial".	۳۱۸
شکل (۴- ۹۳):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Golf".	۳۲۰
شکل (۴- ۹۴):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Polynomial".	۳۲۲
شکل (۴- ۹۵):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Golf با استفاده از مدل جنگل تصادفی".	۳۲۴
شکل (۴- ۹۶):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Golf".	۳۲۶
شکل (۴- ۹۷):	فرآیند آموزشی "تنظیم دستی وزن صفات مجموعه داده Golf".	۳۲۹
شکل (۴- ۹۸):	فرآیند آموزشی "مماسبه کردن وزن صفات مجموعه داده Golf و ذخیره کردن آنها در یک ExampleSet".	۳۳۳
شکل (۴- ۹۹):	فرآیند آموزشی "مقدمه‌ای بر اپراتور Clone Parameters".	۳۳۶
شکل (۴- ۱۰۰):	فرآیند آموزشی "یافتن مقادیر بهینه برای پارامترهای اپراتور SVM از طریق اپراتور Optimize Parameters (Evolutionary)".	۳۴۱
شکل (۴- ۱۰۱):	فرآیند آموزشی "یافتن مقادیر بهینه برای پارامترهای اپراتور SVM".	۳۴۶
شکل (۴- ۱۰۲):	فرآیند آموزشی "ساختن یک مدل با استفاده از یک مجموعه پارامتر بهینه".	۳۵۲
شکل (۴- ۱۰۳):	فرآیند آموزشی "کاهش صفات مجموعه داده Polynomial".	۳۵۶
شکل (۴- ۱۰۴):	فرآیند آموزشی "کاهش صفات مجموعه داده Polynomial از طریق Forward Selection".	۳۶۰

- شکل (۴-۱۰۵): فرآیند آموزشی "کاهش صفات مجموعه داده Polynomial" ۳۶۶
- شکل (۴-۱۰۶): فرآیند آموزشی "کاهش صفات مجموعه داده Polynomial" ۳۷۰
- شکل (۴-۱۰۷): فرآیند آموزشی "کاهش صفات مجموعه داده Polynomial" ۳۷۶
- شکل (۴-۱۰۸): فرآیند آموزشی "اعمال کردن GGA روی مجموعه داده Polynomial" ۳۸۱
- شکل (۴-۱۰۹): فرآیند آموزشی "اعمال کردن YAGGA روی مجموعه داده Polynomial" ۳۸۶
- شکل (۴-۱۱۰): فرآیند آموزشی "اعمال کردن YAGGA2 روی مجموعه داده Polynomial" ۳۹۳
- شکل (۴-۱۱۱): فرآیند آموزشی "مماسیه کردن وزن صفات مجموعه داده Polynomial" ۳۹۸
- شکل (۴-۱۱۲): فرآیند آموزشی "مماسیه کردن وزن صفات مجموعه داده Polynomial" ۴۰۲

مقدمه ای بر فصل چهارم

مدلسازی داده¹ به صورت کلی در مهندسی نرم افزار، فرآیند مستندسازی طراحی یک سیستم نرم افزاری پیچیده و تبدیل آن به دیگرامی است که به سادگی درک شود. این دیگرام از متن و نشانه های خاصی برای نشان دادن جریان داده ها استفاده می کند. از این دیگرام می توان به عنوان برنامه کاری برای ساخت و یا مهندسی مجدد یک سیستم استفاده نمود.

معمولاً مدل داده ها در فازهای تحلیل و طراحی پروژه در جهت اطمینان از این که کلیه جوانب یک سیستم درستی درک شده است ایجاد می شود. یک مدل داده ای را می توان فلوچارتی در نظر گرفت که روابط بین داده ها را نمایش می دهد. از آنجا که نشان دادن همه روابط ممکن در داده ها کار زمان بر و سنگینی است، نمی توان این کار در یک مرحله انجام داد.

بنابراین معمولاً به ترتیب مدل های مفهومی، منطقی و فیزیکی داده ها ساخته می شوند. این مدل های داده ای به دینفعان پروژه سیستم اطلاعاتی امکان می دهد خطاها را به سرعت شناسایی کرده و قبل از شروع کار کد نویسی تغییرات لازم را اعمال کنند. متخصصان مدلسازی داده ها اغلب از چند مدل برای دیدن یک داده استفاده می کنند تا مطمئن شوند که همه فرایندها، موجودیتها و روابط بین داده ها به درستی شناسایی شده اند.

اگر مدلسازی داده ها در طراحی سیستم های اطلاعاتی سازمانی مورد نظر باشد، به عنوان فرآیندی برای تعریف و تحلیل نیازمندی های مورد نیاز داده ها برای پشتیبانی از فرآیندهای کسب و کار تعریف می شود. بنابراین در فرآیند مدلسازی داده ها، متخصصین مدلسازی داده ها باید با دینفعان کسب و کار و همچنین کاربران سیستم های اطلاعاتی به صورت تنگاتنگی کار کنند تا بتوانند نیازمندی های سیستم را به درستی شناسایی کنند.

در مبحث هوش تجاری و خصوصاً تحلیل داده ها در کسب و کار، مدلسازی داده شیوه مهیا کردن داده ها برای تحلیل و اکتشاف داده هاست. نرم افزارهای هوش تجاری مدرن دلی ابزارهای قدرتمندی برای خودکار نمودن کار طاقت فرسا و زمانبر آماده سازی مدلسازی و profiling داده ها هستند. یعنی هدف از Data Modeling در ابزارهای هوش تجاری، تحلیل داده هاست و نه استفاده از داده ها در سیستم های تراکنشی.

مدلسازی داده ها کلید اصلی موفقیت در هوش کسب و کار (BI) است. مدلسازی داده ها در هوش تجاری باید کسب و کار محور باشد یعنی باید بر حل مسائل کاری متمرکز باشد. این کار با تعریف واضحی از کسب و کار خاصی که سیستم برای آن طراحی می شود، شروع می شود. در تعریف این کسب و کار خاص، شناخت هدف کسب و کار و این که چگونه داده ها به آن کسب و کار کمک می کند ضروری است. در هوش تجاری، یک مدل داده برای یک بخش از یک کسب و کار به سختی برای بخش های دیگر آن مناسب است.

¹ Data Modeling