

خوشه‌بندی در یادگیری ماشین

چگونه خوشه‌بندی به داده‌های کاوی معنا می‌بخشد؟

مؤلف: دکتر شهین پوربهرامی

(عضو هیأت علمی دانشگاه فنی و حرفه‌ای تبریز)

نیاز دانش

فهرست مطالب

عنوان	شماره صفحه
یادداشت نویسنده	۷
مقدمه	۷
بخش اول: مفاهیم پایه	۹
فصل ۱ معرفی	۱۱
۱-۱ تئوری یادگیری و داده کاوی	۱۳
۲-۱ ساخت همسایگی در خوشه بندی	۱۵
۳-۱ مروری اجمالی بر روش های خوشه بندی مؤثر برای کلان داده ها با ابعاد بالا	۱۶
۴-۱ الگوریتم های خوشه بندی جریان داده مبتنی بر ابعاد بالا و چگالی	۱۷
۵-۱ الگوریتم های خوشه بندی جریان داده پیشرفته دائمی	۱۸
۶-۱ الگوریتم های کارآمد انرژی برای جمع آوری و خوشه بندی داده های جریان حسگر	۱۹
فصل ۲ خوشه بندی سلسله مراتبی و افرازی	۲۱
۱-۲ الگوریتم های خوشه بندی سلسله مراتبی	۲۲
۲-۲ خوشه بندی با روش Single-Link	۲۳
۳-۲ خوشه بندی با روش Complete-Link	۲۴
۴-۲ خوشه بندی با روش Average-Link	۲۵
۵-۲ خوشه بندی با روش Group Average-Link	۲۵
۶-۲ خوشه بندی با روش Median Distance	۲۶
۷-۲ خوشه بندی با روش Ward	۲۷
۸-۲ روش های افرازی (مسطح)	۲۸
فصل ۳ معیار شباهت بین خوشه ها	۳۳
۱-۳ معیارهای تشابه	۳۴
۲-۳ شاخص جاکارد	۳۴
۳-۳ فاصله اقلیدوسی	۳۵
۴-۳ مربع فاصله اقلیدوسی	۳۵

۳۶ میانگین فاصله اقلیدسی
۳۶ فاصله منهتن (فاصله قدر مطلق)
۳۶ فاصله مینکوفسکی
۳۷ معیار شباهت کسینوسی
۳۷ نتیجه‌گیری در انتخاب معیار شباهت‌ها

فصل ۴ خوشه‌بندی مبتنی بر گراف هندسی..... ۳۹

۴۰ گراف گابریل
۴۱ الگوریتم خوشه‌بندی مثلث دلونی
۴۱ الگوریتم ساخت همسایگی (NC)
۴۲ گراف نزدیک‌ترین همسایگی و بتا اسکلت
۴۲ گراف همسایگی مبتنی بر زاویه
۴۴ خوشه‌بندی براساس ساخت همسایگی با آپولونیوس
۵۶ الگوریتم k نزدیک‌ترین همسایگی

فصل ۵ خوشه‌بندی مبتنی بر چگالی..... ۵۷

۵۸ الگوریتم DBSCAN و مشتقات آن
۶۰ الگوریتم DMM-Stream
۶۰ الگوریتم DBCURE
۶۱ الگوریتم خوشه‌بندی مبتنی بر چگالی پیک
۶۳ الگوریتم DPC-PCA و DPC-KNN-PCA

فصل ۶ خوشه‌بندی مبتنی بر گراف..... ۶۷

۶۸ الگوریتم K - گراف بهینه‌ی وابسته
۶۸ پارتیشن‌بندی گراف نويز مبتنی بر چگالی
۷۱ الگوریتم خوشه‌بندی K - نزدیک‌ترین همسایه ترکیبی
۷۱ الگوریتم RM-CRAG
۷۲ الگوریتم CP-Cluster

فصل ۷ خوشه‌بندی مبتنی بر مشبک.....	۷۳
۱-۷ الگوریتم مبتنی بر اطلاعات آماری مشبک و خوشه‌بندی موجی.....	۷۴
۲-۷ الگوریتم رویکردهای ادغام فواصل تطبیقی با داده‌کاوی فضایی.....	۷۶
۳-۷ افزایش متعامد.....	۷۷
۴-۷ خوشه‌بندی مشبک با تغییر جهت محور.....	۷۸
۵-۷ الگوریتم پالایش توری تطبیقی و BANG.....	۷۹
۶-۷ خوشه‌بندی در جستجو.....	۷۹

فصل ۸ خوشه‌بندی فازی.....	۸۱
۱-۸ الگوریتم خوشه‌بندی C میانگین.....	۸۳
۲-۸ الگوریتم خوشه‌بندی C میانگین برای داده‌های نویزی.....	۸۴
۳-۸ الگوریتم خوشه‌بندی C میانگین فازی شبه نظارتی: رویکردی مبتنی بر هسته.....	۸۵
۴-۸ الگوریتم خوشه‌بندی C میانگین مبتنی بر آنتروپی.....	۸۵
۵-۸ الگوریتم خوشه‌بندی C میانگین مبتنی بر آنتروپی برای داده‌های نویزی.....	۸۶
۶-۸ الگوریتم خوشه‌بندی C میانگین با استفاده از یادگیری وزن ویژگی‌ها.....	۸۷

فصل ۹ راهکارهای اعتبارسنجی و مقایسه‌ی خوشه‌ها.....	۸۹
۱-۹ ابزارهای اعتبارسنجی.....	۹۰
۲-۹ توابع ارزیابی خوشه‌ها.....	۹۱
۳-۹ معیارهای نسبی.....	۹۵
۴-۹ معیارهای خروجی.....	۹۶
۵-۹ معیارهای درونی.....	۹۷
۶-۹ شاخص‌های اعتبارسنجی.....	۹۷

فصل ۱۰ معایب و مزایای راهکارهای خوشه‌بندی.....	۱۰۱
۱-۱۰ ارزیابی کلی الگوریتم‌های خوشه‌بندی از دیدگاه مختلف.....	۱۰۲
کتابشناسی.....	۱۰۹

یادداشت نویسنده

مقدمه

خوشه‌بندی در یادگیری ماشین یک حوزه‌ی مطالعاتی گسترده و کاربردی است. هدف استخراج دانش رویدادهای کاربردی از جمله داده‌های دریافتی از حسگرها، تشخیص موضوعات نو ظهور استخراج اطلاعات درون داده‌ها برای بازاریابی و مسائلی است که کاربرد وسیع‌شان در آینده بیش از حد دور نیستند. تجزیه و تحلیل داده‌ها در حوزه‌های مختلف، ما را قادر به انجام عمل استخراج دانش از داده‌های خام می‌کند یا تحلیل در یک پایگاه داده را به طور مرتب سریعتر از حد غیرپردازشی آن می‌کند. تحلیل فوق‌العاده‌ی از داده‌ها در دنیای تجارت الکترونیک با استفاده از خوشه‌بندی صورت می‌پذیرد. خوشه‌بندی؛ در ساختار بدون ناظر داده‌کاوی تقریباً غیرقابل جایگزین است. سؤال اساسی این کتاب ساده است: چه چیزی از محاسبات داده‌ها می‌تواند در یادگیری ماشین مشارکت کند؟ ما طبیعتاً انتظار پردازش با دقت بالا از روش‌های خوشه‌بندی داریم، اما چه نوع از دقت و روی چه نوع از داده‌ها؟ یا آیا پردازش هر نوع از داده با اشکال و ابعاد مختلف امکان‌پذیر است؟ هر فرم از پیچیدگی محاسباتی طبیعی است؟

روش‌های یادگیری بدون ناظر به طور کلی و روش‌های خوشه‌بندی به طور خاص، پتانسیل کشف الگوهای غیرمنتظره را دارند که منجر به بینش‌های ارزشمندی از داده‌ها می‌شوند. خوشه‌بندی شاید یکی از مواردی باشد که به طور گسترده در زمینه داده‌کاوی و یادگیری ماشین مورد مطالعه قرار گرفته است. خوشه‌بندی شامل طیف گسترده‌ای از حوزه‌های مانند متن، بازاریابی، شبکه‌های اجتماعی و داده‌های بیولوژیکی است. علاوه بر این، الگوریتم‌های خوشه‌بندی بستگی زیادی به دامنه داده‌ها و سناریوی مسئله دارند.

بنابراین، این کتاب به سه جنبه اصلی خوشه‌بندی داده‌ها می‌پردازد: اولین مجموعه از فصل‌ها روی روش‌های پایه و کلاسیک برای طبقه‌بندی داده‌ها متمرکز خواهند شد. این‌ها شامل روش‌هایی مانند خوشه‌بندی سلسله‌مراتبی، خوشه‌بندی افرازی است. فصل‌های بعدی به

حوزه‌ها و سناریوهای جدید خوشه‌بندی مانند روش‌های مبتنی بر هندسه، روش‌های مبتنی بر گراف، روش‌های مبتنی بر مشبک و روش‌های مبتنی بر چگالی متمرکز می‌شوند. مجموعه فصل‌ها به دلیل ایجاد ذهنیت از فرآیند خوشه‌بندی و روش‌های مختلفی که می‌توانند دانشی یکسان از مجموعه داده‌های یکسان را ارائه دهند، به بینش‌های تفصیلی مختلفی از فرآیند خوشه‌بندی متمرکز خواهند کرد. پاسخ به اینکه چگونه می‌دانیم یک خوشه‌بندی خاص خوب است یا نیازهای آن کاربرد خاص را برطرف می‌کند؟ روش‌های بیشمار وجود دارد که می‌توان این موضوعات را بررسی کرد.

این کتاب به گونه‌ای طراحی شده است که بتواند در پوششی کامل مباحث خوشه‌بندی نوین تلفیق شده با روش‌های پایه در حوزه‌های ذکر شده را بیان کند و امیدوارم که این امر به عنوان مرجعی کامل برای دانشجویان و محققان باشد.

شهین پوربهرامی

تبریز، آبان ۱۳۹۸