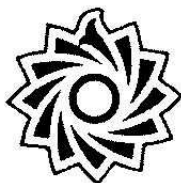


بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه تربیت دبیر شهید رجایی

پاک سازی داده ها

رفع ناسازگاری، تکرار، مقدار از دست رفته و اغتشاش

تألیف:

دکتر نگین دانشپور

عضو هیأت علمی دانشگاه تربیت دبیر شهید رجایی

مهدیه عطایان

سرشناسه	دانشپور، نگین، ۱۳۵۵ -
عنوان و نام پدیدآور	پاک‌سازی داده‌ها: رفع ناسازگاری، تکرار، مقدار از دست‌رفته و اغتشاش / تألیف نگین دانشپور، مهدیه عطاییان، ویراستار ادبی ساغر سلمانی‌نژاد
مشخصات نشر	تهران: دانشگاه تربیت دبیر شهید رجایی، ۱۴۰۰
مشخصات ظاهری	۲۲۰ ص: جدول
شابک	۹۷۸-۶۲۲-۶۵۸۹-۲۰-۸
وضعیت فهرست نویسی	فیبا
یادداشت	واژه‌نامه
یادداشت	کتابنامه: ص: [۱۸۶] - ۱۸۸
یادداشت	نمایه
موضوع	داده‌پردازی
موضوع	Electronic data processing
موضوع	داده‌کاوی
موضوع	Data mining
شناسه افزوده	عطاییان، مهدیه، ۱۳۷۱ -
شناسه افزوده	دانشگاه تربیت دبیر شهید رجایی
شناسه افزوده	Shahid Rajaei Teacher Training University
زده بندی کنگره	Q۵۷۶
زده بندی دیوبی	۱۵۱۰۸۸
شماره کتابشناسی ملی	۱۵۱۰۸۸
اطلاعات رکورد کتابشناسی	فیبا



سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

عنوان	پاک‌سازی داده‌ها: رفع ناسازگاری، تکرار، مقدار از دست‌رفته و اغتشاش
تألیف	دکتر نگین دانشپور، عضو هیأت علمی دانشگاه تربیت دبیر شهید رجایی / مهدیه عطاییان
ویراستار ادبی	دکتر ساغر سلمانی‌نژاد
نوبت چاپ	اول - زمستان ۱۴۰۰
انتشارات	دانشگاه تربیت دبیر شهید رجایی
لیتوگرافی، چاپ	رجاء نقشبند، شریف
طراح جلد	محمد معتمدی‌نژاد
ناظر چاپ	محمد معتمدی‌نژاد
صفحه‌آرا	نیره فیروزی
شمارگان	۱۰۰ جلد
قیمت	۶۰۰,۰۰۰ ریال
شابک	۹۷۸-۶۲۲-۶۵۸۹-۲۰-۸
	ISBN: 978-622-6589-20-8

کلیه حقوق این اثر برای مؤلفان و دانشگاه تربیت دبیر شهیدرجانی محفوظ است.

نشانی: تهران، لویزان، کد پستی ۱۶۷۸۸-۱۵۸۱۱، صندوق پستی ۱۶۳ - ۱۶۷۸۵، تلفن: ۹۰ (۲۶۳۲) - ۲۲۹۷۰۰۶۰

۲۲۹۷۰۰۷۰، تلفکس: ۲۲۹۷۰۰۴۲، پست الکترونیکی: publish@sru.ac.ir، وب سایت: <http://publish.sru.ac.ir>

پیش‌گفتار

داده‌های جمع‌آوری شده از منابع مختلف و توزیع شده برای اتخاذ تصمیمات مفید و سودمند باید به اطلاعات و دانش تبدیل شوند. به‌طور سنتی استخراج دانش را تحلیل‌گران داده انجام می‌دهند اما با توجه به رشد روزافزون داده‌ها، نیازمند روش‌هایی مبتنی بر رایانه هستیم. فرآیندهای مبتنی بر رایانه برای کشف دانش را «داده‌کاوی» می‌گویند. کیفیت داده‌های مورد بررسی در هنگام استخراج دانش، اهمیت بسزایی دارد. صحت، کامل بودن و سازگاری از جمله معیارهای مورد بررسی در بحث کیفیت داده‌ها هستند. داده‌های دنیای واقعی ممکن است به دلایل مختلفی دچار مشکل عدم کیفیت شوند. مقادیر جاافتاده، داده‌های مغشوش، ناسازگاری و مقادیر تکراری از جمله مشکلات عمده داده‌ها هستند. از این رو پیش‌پردازش که اغلب به منظور رفع این مشکلات انجام می‌شود، یکی از مهم‌ترین مراحل کشف دانش است.

پیش‌پردازش فرایندی زمان‌بر است اما با توجه به هزینه‌هایی که داده‌های فاقد کیفیت برای سازمان‌ها ایجاد می‌کنند، فرایندی اجتناب‌ناپذیر است. در صورتی که دانش مورد نیاز برای اتخاذ تصمیمات، از داده‌های فاقد کیفیت استخراج شود، منجر به شکست تصمیمات می‌شود. پاک‌سازی و یکپارچه‌سازی از مراحل پیش‌پردازش هستند. در این کتاب تمرکز بر روی روش‌های جدید پاک‌سازی داده‌ها است. روش‌های مورد بررسی بر روی داده‌های عددی، ترتیبی و اسمی متمرکز هستند. در سال‌های اخیر پاک‌سازی داده‌ها از اهمیت ویژه‌ای برخوردار شده است، زیرا الگوریتم‌های داده‌کاوی برای عملکرد صحیح و ارائه دانش مفید، پیش‌بینی‌ها یا توصیف‌ها، نیاز به داده‌های صحیح، کامل و سازگار دارند. در فرایند پاک‌سازی داده‌ها هدف، تشخیص داده‌های مغشوش، تکراری، پرت و مقادیر جاافتاده است.

در این کتاب به معرفی مفاهیم اولیه پیش‌پردازش داده‌ها و ابعاد مختلف آن می‌پردازیم. همچنین مشکلات کیفیت داده‌ها به صورت اجمالی بیان می‌شوند و راهکارهای ابتدایی برای پاک‌سازی داده‌ها به‌طور مختصر معرفی می‌شوند. تمرکز اصلی در این کتاب بر روی شرح الگوریتم‌های پیشنهادی نویسندگان در حوزه پاک‌سازی داده‌ها است. کارایی این الگوریتم‌ها به دقت ارزیابی شده و برتری آن‌ها در مقایسه با کارهای معتبر جدید و قابل قیاس در آن حوزه تأیید شده است.

به این منظور بر روی این الگوریتم‌ها آزمایش‌های متعدد در شرایط مختلف انجام داده‌ایم. بر اساس آزمایش‌ها و ارزیابی‌های انجام‌شده، این نتیجه حاصل شد که این الگوریتم‌ها نسبت به الگوریتم‌های مشابه عملکرد بهتری دارند و بر این اساس در این کتاب ارائه کردیم.

با توجه به این که پاکسازی داده‌ها امری اجتناب‌ناپذیر در تحلیل داده‌ها است و در سال‌های اخیر این امر از اهمیت ویژه‌ای برخوردار شده است، هدف اصلی از ویرایش این کتاب ارائه یک نسخه فارسی متمرکز در این حوزه است. اگرچه منابع انگلیسی متعددی در زمینه پیش‌پردازش و کیفیت داده‌ها موجود است، اما منبع فارسی متمرکزی در حوزه پاکسازی داده‌ها موجود نیست. این کتاب می‌تواند برای محققان یا تحلیلگران داده در دانشگاه‌ها و مراکز تحقیقاتی، مفید واقع شود.

فهرست مطالب

پیشگفتار

فصل ۱

پیش‌پردازش داده‌ها

۱.۱. کیفیت داده‌ها و ابعاد آن

۱.۱.۱. صحت

۲.۱.۱. کامل بودن

۳.۱.۱. سازگاری

۴.۱.۱. ابعاد مرتبط با زمان

۵.۱.۱. تفسیرپذیری و قابل اعتماد بودن

۲.۱. پاکسازی داده‌ها

۱.۲.۱. مقادیر از دست‌رفته

۲.۲.۱. تصحیح داده‌های مغشوش

۳.۲.۱. شناسایی داده‌های پرت

۴.۲.۱. شناسایی داده‌های تکراری

۵.۲.۱. شناسایی تناقضات محدودیت‌های یکپارچگی

۶.۲.۱. معیارهای ارزیابی الگوریتم‌های پاکسازی داده

فصل ۲

شناسایی تناقضات محدودیت‌های یکپارچگی و تعمیر داده‌ها

۱.۲. تصحیح خودکار مبتنی بر وابستگی تابعی و سیستم یادگیری مرکب

۱.۱.۲. مرحله اول: شناسایی تناقضات وابستگی‌های تابعی با رکوردها

- ۴۴ ۲.۱.۲. مرحله دوم: استخراج مقادیر تکرارشونده
- ۴۵ ۳.۱.۲. مرحله سوم: شناسایی خطا
- ۵۰ ۴.۱.۲. مرحله دوم: تصحیح خطا
- ۵۲ ۲.۲. تصحیح خودکار داده و وابستگی تابعی با الگوریتم اکتشافی
- ۵۴ ۱.۲.۲. تعمیر داده‌ها
- ۵۷ ۲.۲.۲. تعمیر محدودیت‌ها
- ۵۹ ۳.۲.۲. الگوریتم تجربی انتخاب محدودیت‌ها
- ۶۳ ۳.۲. خلاصه

فصل ۳

جایگذاری مقادیر جافتاده

- ۶۹ ۱.۳. جایگذاری مقادیر جافتاده مبتنی بر خوشه‌بندی و رویکرد ترکیبی
- ۶۹ ۱.۱.۳. مرحله اول: خوشه‌بندی
- ۷۱ ۲.۱.۳. مرحله دوم: جایگذاری مقادیر جافتاده مبتنی بر رگرسیون خطی و نزدیکترین همسایگی
- ۷۴ ۲.۳. جایگذاری مقادیر جافتاده با استفاده از روش‌های مبتنی بر همبستگی
- ۷۵ ۱.۲.۳. مرحله اول: محاسبه اولویت صفات جافتاده
- ۷۵ ۲.۲.۳. مرحله دوم: انتخاب مجموعه داده پایه
- ۷۷ ۳.۲.۳. مرحله سوم: حداکثر نمودن همبستگی بین ویژگی‌ها با بسط دادن مجموعه پایه
- ۸۱ ۴.۲.۳. مرحله چهارم: تخمین مقادیر جافتاده با استفاده از زیرمجموعه‌های یافته شده
- ۸۳ ۵.۲.۳. مرحله پنجم: ترکیب گام‌های حداکثرسازی و تخمین
- ۸۴ ۳.۳. جایگذاری مقادیر جافتاده با استفاده از خوشه‌بندی فازی و معیار اطلاعات متقابل
- ۸۵ ۱.۳.۳. مرحله اول: محاسبه اولویت صفات جافتاده
- ۸۶ ۲.۳.۳. مرحله دوم: خوشه‌بندی
- ۹۱ ۳.۳.۳. مرحله سوم: انتخاب ویژگی در هر خوشه انتخاب شده
- ۹۲ ۴.۳.۳. مرحله چهارم: اعمال مدل رگرسیون بر هریک از خوشه‌ها و صفات انتخاب شده
- ۹۳ ۵.۳.۳. مرحله پنجم: انجام نظرخواهی وزن دار برای تخمین نهایی
- ۹۳ ۴.۳. جایگذاری مقادیر جافتاده با استفاده از درونیابی معکوس فاصله وزن دار و رویکردی ترکیبی

- ۹۴ ۱.۴.۳. مرحله اول: تعیین شعاع جستجو با استفاده از خوشه‌بندی k -means
- ۱۰۶ ۲.۴.۳. مرحله دوم: تعیین توان تأثیر همسایگان با الگوریتم جستجوی فاخته
- ۱۰۸ ۳.۴.۳. مرحله سوم: تخمین مقادیر جافتاده
- ۵.۳ جایگذاری مقادیر جافتاده با استفاده از شبکه عصبی مدلسازی گروهی داده‌ها و الگوریتم هوش هیجانی
- ۱۱۰
- ۱۱۱ ۱.۵.۳. شبکه عصبی مدلسازی گروهی داده‌ها
- ۱۱۵ ۲.۵.۳. الگوریتم هوش هیجانی
- ۱۱۷ ۳.۵.۳. ترکیب شبکه عصبی مدلسازی گروهی داده‌ها و الگوریتم هوش هیجانی
- ۱۲۰ ۶.۳. خلاصه
- ۱۲۳ **فصل ۴**
- ۱۲۳ **شناسایی رکوردهای تکراری**
- ۱۲۶ ۱.۴. شناسایی رکوردهای تکراری مبتنی بر خوشه‌بندی و نشانه‌سازی
- ۱۲۶ ۱.۱.۴. مرحله اول: انتخاب ویژگی و بخش‌بندی داده‌ها
- ۱۲۷ ۲.۱.۴. مرحله دوم: محاسبه شباهت نمونه‌ها مبتنی بر یادکردن و خوشه‌بندی داده‌ها
- ۱۲۹ ۳.۱.۴. مرحله سوم: محاسبه فاصله نمونه‌ها مبتنی بر یادکردن
- ۱۳۰ ۴.۱.۴. مرحله چهارم: شناسایی رکوردهای مشابه
- ۱۳۲ ۱.۲.۴. مرحله اول: انتخاب ویژگی
- ۱۳۳ ۲.۲.۴. مرحله دوم: خوشه‌بندی سلسله مراتبی
- ۱۳۷ ۳.۲.۴. مرحله سوم: مقایسه رکوردها
- ۱۴۱ ۳.۴. خلاصه
- ۱۴۳ **فصل ۵**
- ۱۴۳ **تشخیص داده‌های مغشوش**
- ۱۴۶ ۱.۵. تشخیص اغتشاش با استفاده از فیلترها
- ۱۴۷ ۱.۱.۵. فیلتر ترکیبی
- ۱۴۷ ۲.۱.۵. فیلتر بخش‌بندی تکرار شونده
- ۱۴۸ ۲.۵. تشخیص اغتشاش مبتنی بر خوشه‌بندی و نزدیک‌ترین همسایگی