معرفی و مقایسه روش‌های پیش‌پردازش داده برای کاربردهای مختلف داده‌کاوی

مریم بابائی[۱]، زهرا صفار یزدی[۲]، محمدحسین سرایی[۳]

وجود مسائلی نظیر ناقص بودن داده‌ها، ناسازگاری آن‌ها و وجود ناخالصی‌هایی همچون خطاها، مقادیر تقریبی و مقادیر خارج از محدوده نرمال در پایگاه داده‌های واقعی، باعث کاهش کیفیت داده‌کاوی می‌شود. برای دستیابی به نتایج مطلوب‌تر، نیاز به داده‌های با کیفیت بالاتر وجود دارد. پیش‌پردازش، گامی مهم در راستای داده‌کاوی موفقیت‌آمیز است. اعمالی که در پیش‌پردازش انجام می‌شوند عبارت از حذف ناخالصی‌ها و اصلاح داده‌های نادرست، یکپارچه‌سازی داده‌ها، تغییر داده‌ها و کاهش داده‌ها می‌باشد. بر اساس نوع کاربردی که عمل داده کاوی باید روی آن انجام شود، تکنیک‌های مختلفی برای هر یک از این اعمال مورد استفاده قرار می‌گیرد. این مقاله، به معرفی تکنیک‌های به کار رفته برای انجام پیش‌پردازش در کاربردهای مختلف، ارزیابی و مقایسه نتایج حاصل از آن‌ها می‌پردازد؛ کارآمدترین تکنیک به کار رفته برای هر کاربرد را تعیین نموده و برای استفاده‌های آینده در کاربردهای مشابه، پیشنهاد می‌کند.

کلمات کلیدی : پیش‌پردازش، یکپارچه‌سازی، تغییر داده‌ها، کاهش داده‌ها

Pre-processing techniques for data mining applications

Maryam Babaie[1], Zahra Saffar Yazdi[2], Mohammad Hosein Saraee[3]
Advanced Database Syetems, Data Mining and Bioinformatics Research Laboratory
Department of Electrical and Computer Engineering
Isfahan University of Technology, Isfahan, 84156-83111

Incomplete and inconsistent data and existence of impurities like errors, noises and outliers in real datasets decrease quality of data mining. To obtain more desirable results, we need to have high quality data. Data pre-processing is an important step toward successful data mining. Major tasks in data pre-processing are data cleaning, data integration, data transformation and data reduction. On the basis of the type of application on which data mining is done, different techniques for each of the above tasks are used. This paper introduces some existing techniques for pre-processing data within different applications, evaluates and compares their results and finally presents the most efficient techniques for each of such applications.

Keywords: pre-processing, integration, data transformation, data reduction

۱. مقدمه

امروزه، در پایگاه‌داده‌های دنیای واقعی، امکان وجود داده‌های دارای نویز، فیلدهای فاقد مقدار و داده‌های ناسازگار بسیار بالاست. علت این امر، اندازه بسیار بزرگ و پراکندگی و ناهمگونی منابع استخراج داده‌ها می‌باشد. داده‌های با کیفیت پایین، منتج به داده‌کاوی با کیفیت پایین می‌شوند. بنابراین، به منظور افزایش کیفیت داده‌ها و متعاقباً نتایج فرایند داده‌کاوی، نیاز به انجام پیش‌پردازش روی داده‌ها وجود دارد.

تکنیک‌های پیش‌پردازش متعددی وجود دارد. از پاکسازی داده‌ها[۱] می‌توان برای حذف نویز و تصحیح ناسازگاری‌های موجود در داده‌ها استفاده کرد. یکپارچه‌سازی داده‌ها[۲]، داده‌ها را از منابع مختلف در یک محل منسجم مانند مخزن داده[۳]، ادغام می‌کند. از روش‌های تبدیل داده‌ها[۴] از جمله نرمال‌سازی نیز می‌توان استفاده کرد. کاهش داده‌ها[۵] ، می‌تواند از طریق روش‌هایی مثل به هم پیوستن[۶]، حذف ویژگی‌های اضافی، خوشه‌بندی[۷] و نمونه‌برداری[۸]، حجم داده‌ها را کاهش دهد. می‌توان از مجموعه‌ای از این تکنیک‌ها نیز، در کنار یکدیگر استفاده نمود.

[1] دانشجوی کارشناسی ارشد معماری کامپیوتر، دانشگاه صنعتی اصفهان، دانشکده مهندسی برق وکامپیوتر m.babaie@ec.iut.ac.ir
[2] دانشجوی کارشناسی ارشد معماری کامپیوتر، دانشگاه صنعتی اصفهان، دانشکده مهندسی برق وکامپیوتر z.saffaryazdi@ec.iut.ac.ir
[3] عضو هیأت علمی، دانشگاه صنعتی اصفهان، دانشکده مهندسی برق وکامپیوتر saraee@cc.iut.ac.ir

تکنیک‌های پیش‌پردازش اگر قبل از عمل داده‌کاوی انجام گیرد می‌تواند به میزان قابل توجهی باعث بهبود فرایند داده‌کاوی و کاهش زمان لازم برای اجرای آن شود. [۶]

در اغلب فعالیت‌هایی که در زمینه داده‌کاوی، روی پایگاه‌داده‌های مختلف صورت گرفته است، پیش‌پردازش داده‌ها یکی از اولین و مهمترین مراحل انجام شده می‌باشد. با توجه به کاربرد داده‌کاوی در پایگاه‌داده‌های مختلف حاوی داده‌هایی در زمینه‌های متفاوت، و ماهیت و خصوصیات مختلف هر یک از این انواع داده‌ها، در موارد گوناگون، تکنیک‌های پیش‌پردازش متفاوتی می‌توانند به‌کار گرفته شوند. تعیین مناسب‌ترین تکنیک‌های پیش‌پردازش برای هر دسته از انواع داده‌ها، نقش مهمی در بهبود کیفیت و کارایی داده‌کاوی ایفا می‌کند.

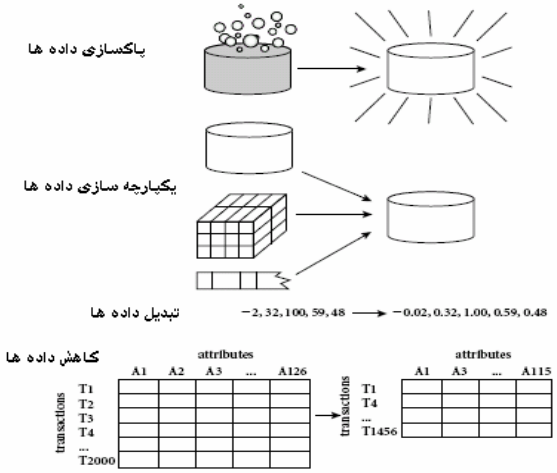
در این مقاله به بررسی تکنیک‌های مختلف پیش‌پردازش و کاربرد آن‌ها در برنامه‌های مختلف داده‌کاوی پرداخته شده است. نتایج ارائه شده از داده‌کاوی انواع مختلف داده‌ها و برنامه‌های کاربردی گوناگون، مطالعه شده و تکنیک‌های پیش‌پردازش به‌کار رفته در هریک از آن‌ها، مورد ارزیابی و مقایسه قرار گرفته اند. در نهایت با جمع‌بندی این بررسی‌ها، رایج‌ترین و مؤثرترین تکنیک‌های به‌کار رفته در هر زمینه استخراج شده و برای استفاده در کاربردهای آینده، معرفی شده است.

روندی که در ادامه مقاله دنبال می‌شود به این صورت است که در بخش دوم به معرفی تکنیک‌های مختلف پیش‌پردازش پرداخته می‌شود. در بخش سوم تکنیک‌های به کار رفته در کاربردهای مختلف معرفی می‌شود. سپس بهترین تکنیک‌های پیشنهادی برای هر دسته از کاربردها در جدولی جمع‌بندی شده و ارائه می‌گردد. در آخرین بخش نیز، جمع‌بندی کلی از مباحث مطرح شده در مقاله بیان می‌شود.

۲. تکنیک‌های پیش پردازش

در اولین و کلی‌ترین دسته‌بندی، تکنیک‌های به کار رفته در پیش‌پردازش به چهار دسته کلی تقسیم می‌شوند. .نمایش کلی از روش‌های پیش‌پردازش داده در شکل ۱ نشان داده‌شده است. اولین دسته از فعالیت‌های انجام گرفته، پاک‌سازی داده‌ها می‌باشد. پاک‌سازی داده‌ها بر مقداردهی به ویژگی‌های فاقد مقدار، متعادل‌سازی مقادیر دارای نویز، شناسایی و حذف مقادیر خارج از محدوده[۹] و رفع ناسازگاری داده‌ها متمرکز است. پاکسازی داده، به ویژه هنگام مجتمع‌سازی داده‌های ناهمگن مورد نیاز است و باید همراه با تبدیلات داده‌ای مرتبط با الگو مورد استفاده قرار گیرد.

در مخازن داده، پایگاه‌داده‌های وابسته و سیستم‌های اطلاعاتی مبتنی بر وب عمومی، نیاز به پاکسازی داده به شدت افزایش می‌یابد. این افزایش به این علت است که منابع، به علت در بر گرفتن نمایش‌های مختلف از داده‌های یکسان، دارای اطلاعات افزونه می‌باشند. به منظور فراهم نمودن دسترسی به داده‌های صحیح و سازگار، ترکیب نمایش‌های مختلف داده و حذف اطلاعات تکراری، ضروری است. بسته به تعداد منابع داده، درجه ناهمگن بودن آنها و "کثیفی" داده‌ها، ممکن است اجرای چندین گام تبدیل و پاکسازی داده‌ها ضروری باشد. [۲۲]



شکل ۱- روش‌های اصلی پیش‌پردازش داده‌ها

با توجه به اینکه در فرایند داده‌کاوی، داده‌های مورد بررسی از منابع مختلفی به دست می‌آیند، نیاز به یکپارچه‌سازی داده‌های این منابع در یک مجموعه داده کلی وجود دارد که این عمل، دومین بخش از فعالیت‌های پیش‌پردازش را تشکیل می‌دهد. مسئله یکپارچه‌سازی داده، به سه دلیل پیچیده است: اول اینکه منابع گوناگون، داده‌های همپوشان و وابسته به یکدیگر را در بر می‌گیرند. دوم، داده‌ها در مدل‌ها و طرح‌های مختلفی ذخیره می‌شوند، و سومین دلیل این است که منابع مختلف داده، قابلیت پردازش انواع متفاوتی از درخواست‌ها را دارند.[۱۳]

یک مسئله مهم در یکپارچه‌سازی داده‌ها، زبان مورد استفاده در توصیف محتوا و قابلیت‌های منابع داده می‌باشد. زبان مورد استفاده، باید هم به اندازه کافی رسا باشد و هم بتواند مشکل ناهمخوانی پرس‌وجوهای کاربر با تقاضاهای قابل فهم برای الگوی ذخیره‌سازی داده را برطرف کند.

مهمترین مزیت سیستم یکپارچه‌سازی، فراهم نمودن این امکان است که کاربران بدون نگرانی در مورد چگونگی به دست آوردن پاسخ‌ها، بر بخشی از داده‌ها که مورد نیاز آنها می‌باشد متمرکز شوند. در نتیجه از جستجوی منابع مرتبط داده، تعامل با هر یک از این منابع به صورت مجزا و ترکیب داده‌های منابع مختلف بی‌نیاز می‌گردند.

ممکن است داده‌های موجود برای فرایند داده‌کاوی مناسب نباشند. در این‌حالت، باید عمل تبدیل داده‌ها، روی آنها انجام شود. عمل تبدیل داده‌ها از بخش‌های زیر تشکیل می‌شود:

- متعادل‌سازی[۱۰]، این‌کار به منظور حذف نویز از داده‌ها انجام می‌شود. تکنیک‌های موجود برای انجام این عمل عبارتند از دسته‌بندی[۱۱]، رگرسیون و خوشه‌بندی.
- به‌هم پیوستن داده‌ها با انجام عملیات جمع یا رگرسیون روی داده‌ها صورت می‌گیرد. به عنوان مثال فروش روزانه ممکن است جمع شده و در ویژگی جدیدی به نام فروش ماهانه یا سالانه نگهداری شود. این گام، معمولاً به‌منظور ساخت مکعب داده‌ای[۱۲] برای تحلیل داده‌ها در دانه‌بندی‌های مختلف به‌کار می‌رود.
- عمومیت دادن[۱۳] به داده‌ها، داده‌های اولیه سطح پایین را با استفاده از سلسله مراتب مفهومی به مفاهیم سطح بالا تبدیل می‌کند.
- نرمال‌سازی، تغییر مقدار داده‌ها به‌گونه‌ای است که در دامنه محدود و مشخصی قرار گیرند.
- تولید ویژگی، با استفاده از مجموعه ویژگی‌های داده شده، ویژگی‌های جدیدی تولید نموده و به مجموعه داده‌ها اضافه می‌کند.

همانطور که قبلاً نیز گفته شد، یکی از دلایل عمده نیاز به انجام پیش‌پردازش، حجم بسیار بالای مجموعه داده‌هایی است که باید عمل داده‌کاوی روی آن انجام شود. تحلیل این حجم بالای اطلاعات، نیازمند زمان زیادی است که عملاً داده‌کاوی را غیر ممکن می‌سازد.

کاهش داده، مبحث مهمی در زمینه داده‌کاوی است. هدف تکنیک‌های کاهش داده در داده‌کاوی، استخراج زیرمجموعه‌ای کوچک از حجم انبوهی از مجموعه‌داده‌ها با حفظ خصوصیات داده‌های اصلی می‌باشد. اینکار باعث می‌شود عملیات سخت یا غیرممکن داده‌کاوی به صورت کارا و مؤثری انجام شوند. [۱۵]

استراتژی‌های کاهش داده‌ها، شامل موارد زیر است:

- پیوسته سازی مکعب داده، به منظور ساخت مکعب داده روی داده‌ها اجرا می‌شوند.
- انتخاب زیرمجموعه ویژگی‌ها، ویژگی‌ها و ابعاد افزونه، بی‌ربط یا دارای ارتباط ضعیف را شناسایی و حذف می‌کند. این تکنیک و روش‌های آن، به تفصیل در زیر بخش ۲-۱ شرح داده می‌شود.
- نمونه‌برداری، عبارت از انتخاب زیرمجموعه‌ای از داده‌هاست که ویژگی‌های کلی مجموعه داده‌ها را دارا می‌باشند. این استراتژی دارای تکنیک‌های مختلفی است که جزئیات آنها در بخش ۲-۲ بیان می‌گردد.
- گسسته سازی[۱۴] و تولید سلسله مراتب مفهوم، مقادیر سطری ویژگی‌ها را با بازه‌ها یا سطوح مفهومی بالاتر جایگزین می‌کند.

۲-۱ انتخاب ویژگی

انتخاب ویژگی، مسئله انتخاب زیر مجموعه کوچکی از ویژگی‌هاست که در حالت ایده‌آل برای توصیف مفهوم نهایی، لازم و کافی می‌باشد. هدف نهایی انتخاب ویژگی، به دست آوردن فضای ویژگی با این مشخصات می‌باشد : (۱) ابعاد کمتر، (۲) نگهداری اطلاعات کافی، (۳) ارتقاء قابلیت تجزیه فضای ویژگی به دسته‌های مختلف، به عنوان مثال از طریق حذف تأثیر ویژگی‌های دارای نویز و (۴) ارائه قابلیت مقایسه ویژگی‌ها روی مثال‌هایی از یک دسته.[۲۱]

برای بعضی داده‌های خاص، انتخاب ویژگی راه مناسبی نیست زیرا همه ویژگی‌ها برای تعیین دسته‌ای که هر نمونه به آن تعلق دارد، مورد نیاز هستند. محققین، انتخاب ویژگی را با استفاده از روش‌های مختلف، امتحان کرده‌اند. از جمله این روش‌ها می‌توان به روش‌های آماری، هندسی، معیارهای نظریه اطلاعات و برنامه‌نویسی ریاضی اشاره کرد.

Siedlecki و Sklansky از الگوریتم‌های ژنتیک برای انتخاب ویژگی استفاده کرده‌اند. در الگوریتم مورد استفاده آنها، مجموعه اولیه n ویژگی به رشته بیتی n عنصری از صفر و یک‌ها کدگذاری شده که وجود و فقدان ویژگی‌ها در مجموعه را بیان می‌کنند. آن‌ها از تابع دقت کلاسیک به عنوان تابع تطبیق[۱۵] (در الگوریتم‌های ژنتیک برای انتخاب ویژگی‌ها) استفاده کرده‌اند و در مقایسه با روش انشعاب و تحدید و جستجوی ترتیبی به شبکه‌های عصبی بهتری دست یافته‌اند.

Kohavi و Frasca از جستجوی اول بهترین استفاده کرده‌اند که پس از رسیدن به تعداد معینی از نقاط بسط داده شده بدون بهبود، متوقف می‌شود. آنها پیشنهاد کردند که ممکن است بهتر باشد از زیرمجموعه‌ای از ویژگی‌ها استفاده شود که کاهش ناپذیر باشد، یعنی زیرمجموعه‌ای که حذف یک ویژگی از آن، استقلال ویژگی‌ها را از بین می‌برد.

Battiti روشی برای استفاده از اطلاعات متقابل، به منظور ارزیابی محتوای اطلاعات هر ویژگی خاص با توجه به کلاس خروجی، توسعه داد. ویژگی‌هایی که به این روش انتخاب شده‌اند، به عنوان ورودی شبکه‌های عصبی استفاده شدند. نشان داده شده است که این روش، نسبت به روش‌های انتخاب ویژگی که از وابستگی‌های خطی استفاده می‌کنند، بهتر عمل می‌کند.

بسیاری از الگوریتم‌های کارای انتخاب ویژگی، ویژگی‌های مرتبط را بر اساس ارزیابی همبستگی کلاس و یک ویژگی (یا زیرمجموعه‌ای از ویژگی‌ها) تعیین می‌کنند. معیارهای مورد استفاده برای ارزیابی میزان ارتباط، عبارتند از: معیارهای مبتنی بر فاصله، معیارهای مبتنی بر اطلاعات و معیارهای سازگاری. با استفادها از این معیارها، الگوریتم‌های انتخاب ویژگی معمولاً با مجموعه‌ای تهی آغاز می کنند، و متوالیاً ویژگی‌های خوب را به زیرمجموعه ویژگی‌های انتخاب شده می‌افزایند. به این چارچوب انتخاب ویژگی، SFS گفته می‌شود.[۲۸]

می‌توان مسئله انتخاب ویژگی را به عنوان مسئله‌ای از دسته مسائل بهینه‌سازی ترکیبی در نظر گرفت. در آغاز، روش انتخاب ویژگی براساس بهینگی ارائه شده توسط Olafsson و Yang [۲۴] را در نظر می‌گیریم که از متاهیوریستیک بخش‌های تودرتو[۱۶] برای بهینه‌سازی ترکیبی استفاده می‌کند. کلید موفقیت این متاهیوریستیک، الگوی بخش‌بندی آن است که ساختاری مستقل از کاربرد را روی فضای جستجو اعمال می‌نماید. Olafsson و Yang چنین رویکردی را برای بخش‌بندی هوشمندانه فضای زیرمجموعه‌های ویژگی‌ها ارائه نموده‌اند. همچنین نشان داده‌اند که این رویکرد، در مقایسه با دیگر روش‌های انتخاب ویژگی بهتر عمل می‌کند.

۲-۲ نمونه برداری

با استفاده از نمونه‌برداری، الگوریتم‌های داده‌کاوی بر زیرمجموعه کوچکی از داده‌ها اعمال می‌شوند و نتایج تقریبی بدست آمده، برای استخراج نتایج کلی مورد استفاده قرار می‌گیرند. استفاده از تکنیک‌های نمونه‌برداری، موازنه‌ای بین تسریع فرایند داده کاوی و افزایش دقت آن می‌باشد. مزایای استفاده از نمونه‌برداری، هنگامی با وضوح بیشتری قابل تشخیص است که روی مسئله‌ای خاص اعمال شود.

متدهای مورد استفاده برای نمونه‌برداری، با توجه به اینکه هدف از فرایند داده‌کاوی، به دست آوردن قواعد انجمنی[۱۷] است یا کلاس‌بندی داده‌ها، متفاوتند. از جمله الگوریتم‌های نمونه‌برداری به منظور به دست آوردن قواعد انجمنی، می‌توان به الگوریتم SRS [۱۸]، FAST [۱۹] و EA [۲۰] اشاره کرد. ساده‌ترین الگوریتم SRS است که در آن، پایگاه داده یکبار مرور می‌شود و هر تراکنش با احتمال معینی که براساس پارامترهای خاصی تعیین می‌شود، انتخاب می‌گردد. در نتیجه اندازه نمونه انتخاب شده در مرورهای مختلف، متغیر خواهد بود. [۳]

دو الگوریتم FAST و EA ، از این نظر که هر دو تلاش می‌کنند نمونه‌ای تولید کنند که فاصله با پایگاه داده اصلی را حداقل نماید، به یکدیگر شباهت دارند. FAST کار را با مجموعه بزرگی از نمونه‌های تصادفی شروع می‌کند و تراکنش‌های خارج از محدوده را، یعنی تراکنش‌هایی که حذف آن‌ها اختلاف support نمونه و پایگاه داده اصلی را تا حد زیادی کاهش می‌دهد، حذف می‌کند تا نمونه‌های کوچکتری تولید کند که خصوصیات پایگاه داده اصلی را با دقت بیشتری نشان می‌دهند.

الگوریتم EA بطور تکراری و تخمینی داده‌ها را نصف می کند تا نمونه نهایی را بدست آورد. این الگوریتم برخلاف FAST، حد بالای ضمانت شده‌ای برای فاصله با پایگاه داده اصلی ارائه می‌دهد. در این دو روش، اندازه نمونه نهایی بر اساس حافظه موجود تخمین زده می‌شود. بعضی از نمونه‌های تغییر یافته EA ، عمل نصف کردن در مراحل مختلف را بطور موازی انجام می‌دهند. این روش در حالتی که حافظه کم است، مناسب‌تر می‌باشد، زیرا نمونه‌های میانی در حافظه نگهداری نمی‌شوند. زمان اجرای EA نسبت به FAST بیشتر است ولی دقت نتایج نهایی آن بهتر می‌باشد. [۳و۴]

متدهای مورد استفاده نمونه‌برداری به منظور کلاس‌بندی داده‌ها، به سه شاخه کلی ایستا، پویا و فعال تقسیم‌بندی می‌شود. نمونه‌برداری ایستا، به نمونه‌برداری‌هایی اشاره دارد که بدون هیچ آگاهی قبلی از محتوای پایگاه داده، روی آن اجرا می‌شوند. نمونه‌برداری تصادفی که معروف‌ترین و ساده‌ترین الگوریتم نمونه‌برداری است، متعلق به این دسته می‌باشد. نمونه‌برداری پویا مانند نمونه‌برداری ایستا عمل می‌کند. تنها تفاوت این دو روش، در فرایند اعتبار سنجی نمونه به دست آمده می‌باشد[۱]

G. H. John و P. Langley [۹] ، دقت دو روش نمونه‌برداری ایستا و پویا را روی یازده مجموعه داده از مخزن پایگاه‌داده‌های UCI با یکدیگر مقایسه نموده‌اند. نتایج حاصل، نشان می‌دهد که نمونه برداری ایستا، با وجود اینکه روی نمونه‌های کوچک، نتایج بهتری دارد، برای کل پایگاه داده، به دقت پایین‌تری در نتایج دست می‌یابد.

هرچند نمونه‌برداری، یکی از بهترین و رایج‌ترین تکنیک‌های پیش‌پردازش می‌باشد، اما استفاده از آن برای همه کاربردهای داده‌کاوی، توصیه نمی‌شود. از جمله کاربردهای مفید نمونه‌برداری، پردازش بسیاری از انواع داده‌ها از جمله داده‌های تجاری و سیاست‌گذاری عمومی می‌باشد. در مقابل، در برخی موارد، نمونه‌برداری باعث حصول نتایج نامعتبر و ناقص می‌شود، که در این موارد نباید عمل نمونه‌برداری از داده‌ها را انجام داد. [۱۸] این شرایط عبارتند از:

- هنگامی‌که حساب دقیق تمام مبالغ مورد نیاز است. مثلاً در سیستم‌هایی که حساب بدهی - سرمایه را نگهداری می‌کنند. در چنین سیستم‌هایی تمامی تراکنش‌ها باید مورد پردازش قرار گیرد.
- وقتی از کل داده‌ها باید استفاده شود. برای مثال در فرایند سرشماری جمعیت یک کشور، کل داده‌های موجود، مورد نیاز است.
- زمانی‌که پردازش، نیازمند نظارت پیوسته است. مثلاً برای اطلاعات بیماران در بیمارستان‌ها، فرایندهای تولیدی دقیق و گزارشات آب و هوایی در فرودگاه‌ها.
- هنگام انجام عملیات ممیزی که هر یک از رکوردها باید جداگانه بررسی شوند. به عنوان مثال هنگام بازرسی ادعانامه‌های بیمه به منظور کشف ادعاهای غیرواقعی و تولید گزارشات موارد خاص.

۳. تکنیک‌های پیش‌پردازش پیشنهادی برای کاربردهای مختلف داده‌کاوی

همانطور که در بخش‌های پیشین نیز اشاره شد، می‌توان از چندین تکنیک پیش‌پردازش در یک برنامه داده‌کاوی استفاده کرد؛ اما تمام تکنیک‌های معرفی شده در بخش قبل، برای همه کاربردها مناسب نیستند. بر اساس نوع کاربرد و ماهیت داده‌هایی که در آن مورد بررسی قرار می‌گیرند، تکنیک‌های مناسب کاربردهای مختلف نیز، متفاوتند. در این بخش، به معرفی تکنیک‌های پیش‌پردازش به‌کار رفته در کاربردهای مختلف داده‌کاوی و مقایسه نتایج آنها پرداخته می‌شود.

اولین دسته از برنامه‌های مورد بررسی، برنامه‌های کاوش متن می‌باشد. متون ورودی برنامه کاوش متن، شامل تعداد زیادی stopword بی‌اهمیت هستند. همچنین ممکن است قالب این متون، برای داده‌کاوی مناسب نباشد. بنابراین، پیش‌پردازش، یعنی استفاده از روش‌هایی به‌منظور پاکسازی و ساختاردهی متن ورودی برای تحلیل‌های آینده، بخش مهمی از مطالعات عملی کاوش متن را تشکیل می‌دهد. [۷] بخش‌هایی از فرایند کاوش متن که در بخش پیش‌پردازش دسته‌بندی می‌شوند، شامل موارد زیر است:

- ریشه‌یابی، عبارت از فرایند حذف پسوندهای کلمه به منظور رسیدن به ریشه آنهاست. این کار، تکنیک رایجی است که در تحقیقات کاوش متن، مورد استفاده قرار می‌گیرد. این عمل، بدون از دست دادن اطلاعات برای کاربردهای خاص، پیچیدگی را کاهش می‌دهد.
- حذف جاهای خالی و تبدیل حروف کوچک و بزرگ
- حذف stopword، تکنیک بعدی پیش‌پردازش می‌باشد. stopwordها، کلماتی هستند که آنقدر در زبان تکرار می‌شوند که ارزش اطلاعاتی آنها تقریباً صفر است. به عبارت دیگر، انتروپی آن‌ها خیلی پایین است. بنابراین، معمولاً قبل از تحلیل متن، حذف می‌شوند. تعدادی از این stopwpordها، عبارتند از "و"، "یا"، "آن"، "به"، "که" و ... .
- تعیین مترادف‌ها: در برخی موارد، بهتر است مترادف‌های عبارت داده شده را بدانیم، چرا که این کلمات متفاوت معنی یکسانی را می‌رسانند.

یکی از روش‌های رایج در کاوش متون این است که پس از اینکه تمام مترادف‌های کلمات تعیین شدند، به جای همه آنها، یک کلمه را جایگزین کنیم. عمل تعیین مترادف کلمات، و جایگزینی همه مترادف‌ها با یک کلمه، در دسته فعالیت‌های به‌هم پیوستن داده‌ها قرار می‌گیرد. [۲۰]

کاربرد بعدی پیش‌پردازش، کاوش داده‌های مکانی[۲۱] می‌باشد. با افزایش داده‌های مکانی، خوشه‌بندی مکانی و پیدا کردن داده‌های خارج از محدوده در کاوش این نوع داده‌ها، مورد توجه بسیاری قرار گرفته است. در یکی از برجسته‌ترین روش‌ها، آماره پویش مکانی[۲۲]، ناحیه‌ای را پیدا می‌کند که نسبت به مجموعه کلی داده‌ها، بیشترین انحراف را دارد.[۸]

مجموعه داده‌های مکانی، معمولاً حاوی حجم زیادی از داده هستند که در لایه‌های مختلف قرار گرفته است. این داده‌ها ممکن است حاوی خطا باشند یا به صورت مجموعه‌ای با مختصات هماهنگ، جمع‌آوری نشده باشند. بنابراین، معمولاً مجموعه‌ای از مراحل پیش‌پردازش برای آماده‌سازی داده‌ها برای گام‌های بعدی مدلسازی مورد نیاز است. این ماژول‌ها شامل توابع زیر می‌باشد.[۱۱]

- الحاق داده‌ها: استفاده از تابع الحاق روی داده‌ها برای تغییر دقت و محاسبه مقادیر در مجموعه مشترکی از داده‌ها مورد نیاز است. می‌توان از تکنیک‌های معین الحاق، نظیر معکوس فاصله و تقسیم ناحیه به مثلث‌های مجاور یکدیگر، استفاده کرد، اما این تکنیک‌ها مدل فرایند مکانی یا variogram ها را مورد توجه قرار نمی‌دهند. اغلب استفاده از تکنیک‌های الحاق مناسب برای داده‌های مکانی نظیر krigging و الحاق با استفاده از انحناء ترجیح داده می‌شود.
- نرمال‌سازی داده‌ها: از دو روش نرمال‌سازی برای داده‌های مکانی استفاده می‌شود: تبدیل داده‌ها به توزیع نرمال و بردن داده‌ها به محدوده‌ای معین.
- گسسته‌سازی داده‌ها: این گام، برای برخی تکنیک‌های مدل‌سازی (مانند قوانین انجمنی، آموزش درخت‌های تصمیم و مسایل کلاس‌بندی) مورد نیاز است و شامل ضوابط مختلفی برای گسسته‌سازی هدف و ویژگی می‌باشد.

- تولید ویژگی‌های جدید: کاربران می‌توانند با استفاده از عملگرهای پشتبانی شده در سیستم روی مجموعه ویژگی‌های موجود، ویژگی‌های جدیدی تولید کنند.
- انتخاب ویزگی: در دامنه‌هایی که ویژگی‌های زیادی دارند، این گام برای کاهش حجم فضای ویژگی به‌وسیله حذف ویژگی‌های بی‌ربط، مفید است.

دسته بعدی از کاربردهای مهم داده‌کاوی، کاوش کاربرد وب[۲۳] می‌باشد. این برنامه‌ها، بر استفاده از تکنیک‌های داده‌کاوی در رویدادنگاری کاربرد انباره‌های بزرگ داده‌های وب تمرکز دارند. هدف داده‌کاوی در این نوع داده‌ها، تولید نتایجی است که بتوان براساس آن‌ها وب‌سایت را برای سرویس‌دهی بهتر به نیازهای کاربران آن، ساختاردهی کرد. قبل از اعمال الگوریتم‌های داده‌کاوی، باید چندین تکنیک پیش‌پردازش روی داده‌های جمع‌آوری شده از رویداد نگاری سرور اجرا شود.

پیش‌پردازش داده‌های وب، شامل چندین وظیفه مستقل از دامنه می‌باشد که عبارتند از پاک‌سازی داده، شناسایی کاربر، شناسایی نشست، تکمیل مسیر، بازسازی نشست ، شناسایی تراکنش و شکل‌دهی. پاکسازی داده‌ها، عمل حذف رویدادهایی است که مورد نیاز فرایند داده‌کاوی نیستند. شناسایی کاربر، عمل وابسته سازی ارجاعات صفحه به کاربران مختلف است، حتی اگر این ارجاعات از آدرس IP واحدی صورت گرفته باشد. برای شناسایی کاربر علاوه بر رویدادنگاری سرور، توپولوژی وب‌سایت نیز مورد نیاز است. شناسایی نشست، تمام ارجاعات صورت گرفته به صفحات توسط کاربر را در رویدادنگاری دریافت می‌کند و آن‌ها را به نشست‌های کاربر می‌شکند. در اینجا نیز مانند شناسایی کاربر، نیازمند آگاهی از توپولوژی وب‌سایت هستیم. تکمیل مسیر، ارجاعاتی را که به علت caching حذف شده است، به رویدادنگاری اضافه می‌کند. بازسازی نشست‌های کاربر، بازسازی مسیر حرکت کاربران دائمی وب‌سایت در نشست‌های شناسایی شده می‌باشد. در حالات خاص، بروز خطا در بازسازی نشست‌ها و پیگیری ناقص فعالیت‌های کاربر در وب‌سایت، به سادگی منجر به الگوهای نامعتبر و نتایج نادرست می‌شود. شناسایی تراکنش، به این علت مورد نیاز است که قبل از انجام هرگونه کاوش در داده‌های کاربرد وب، دنباله‌های ارجاع صفحات، باید در واحدهایی منطقی که تراکنش‌های وب را مشخص می‌کنند، گروه‌بندی شوند. تفاوت تراکنش با نشست کاربر این است که اندازه تراکنش می‌تواند از یک ارجاع صفحه تا تمام ارجاع صفحات در جلسه کاربر، متغیر باشد. پس از اینکه مراحل پیش‌پردازش روی رویدادنگاری‌های سرور انجام شد، می‌توان به‌منظور شکل‌دهی نشست‌ها یا تراکنش ها برای نوعی از داده‌کاوی که انجام خواهد شد، از آخرین ماژول پیش‌پردازش استفاده نمود.

یکی دیگر از کاربردهای عمده داده‌کاوی در بررسی داده‌های پزشکی می‌باشد. به علت خطا در ورود داده‌ها، داده‌های خارج از محدوده، به‌وفور در این پایگاه داده‌ها دیده می‌شوند. همچنین احتمال وجود نمایش‌های ناسازگار داده، به‌ویژه زمانی‌که بیش از یک مدل برای نمایش معانی خاص استفاده می‌شود، وجود دارد. برخی عناصر، به علت از قلم افتادن، بی‌ارتباط بودن، ریسک بالا و عملیاتی نبودن در زمینه پزشکی جمع‌آوری نشده‌اند.

گاهی اوقات ابهام بین عبارات پزشکی و عناصر داده‌ای، به‌سختی قابل رفع است. مشکل دیگر، این است که مجموعه متغیرهای مرتبط ممکن، توسط دانش دامنه قابل دسترسی، محدود می‌شود. اگر یک مدل پیش‌بینی مفید، از متغیری استفاده کند که با مسئله نهایی ارتباط واضحی نداشته باشد، این مدل با استفاده از پردازش دستی قابل دستیابی نیست.

در چارچوبی که برای آماده‌سازی مجموعه داده‌های پزشکی برای داده‌کاوی ارائه شده است، روش‌های زیر برای آماده‌سازی داده معرفی شده‌اند[۱۶]

۱. سرند کلی داده‌ها[۲۴]
۲. سرند داده‌ها بر اساس بیماری
۳. انتخاب دستی با استفاده از دانش پزشکی
۴. تبدیل عناصر داده‌ای انتخاب شده از "زوج متغیر-مقدار" به شکل جدول مسطح، که در آن اگر برای متغیری بیش از یک مقدار مشاهده شود، اولین مقدار مشاهده شده مورد استفاده قرار می‌گیرد.

Y. Qu و B. L. Adam در اولین گام، ابعاد داده را کاهش دادند. این کار با استفاده از کشف قله و چیدمان، با تخمین قابلیت منمایز سازی قله‌های منفرد انجام شده است، سپس از درخت‌های تصمیم و decision stump استفاده نموده‌اند. H. Tang ، Y. Mukomel و E. Fink از انتخاب خصوصیت آماری، بر اساس اختلاف بین مقادیر میانه طیف برای نمونه‌های سالم و سرطانی استفاده کرده‌اند. پس از انتخاب نقاط طیف با بالاترین اختلاف، شبکه‌های عصبی، ماشین‌های برداری پشتیبان[۲۵] و درخت‌های تصمیم با استفاده از داده‌های سرطان تخمدان ارزیابی شدند.

H. LiuJ. Li و L. Wong ، داده‌های متنوع سرطان تخمدان را برای ارزیابی برخی رویکردهای آماری انتخاب ویژگی، مورد استفاده قرار داده‌اند. پس از اعمال این روش‌های آماری، ویژگی‌های انتخاب شده با یکی از پنج روش زیر به منظور ارزیابی میزان تأثیر انتخاب ویژگی، دسته‌بندی شده‌اند: k-نزدیکترین همسایه، درختهای تصمیم، naïve bayes یا ماشین‌های برداری پشتیبان.[۲] در کل، در پیش‌پردازش داده‌های مربوط به سرطان تخمدان، از سه روش استفاده شده است: نرمال‌سازی، کاهش بعد، و گسسته‌سازی. علت استفاده از گسسته‌سازی، تأثیر آن در دو مورد زیر می‌باشد: (۱) حذف تغییرات کمّی اندک حاصل از نویز و (۲) انتقال از داده‌های با ابعاد زیاد به سمت انتخاب ویژگی‌های مهم که می‌توانند نمونه‌ها را از کلاس‌های مختلف انتخاب کنند.[۲۵]

در حالتی که برخی اندازه‌گیری‌های مهم صورت نگرفته باشند، نیازمند پرکردن مقادیرخالی آنها در محموعه داده‌ها هستیم. دو رویکرد کلی در برخورد با فقدان مقادیر وجود دارد. رویکرد اول، ذخیره‌سازی مستقیم پارامترهای فاقد مقدار است. ذخیره‌سازی آماری پارامترهای بدون مقدار معمولاً بر اساس پارامترهای معادل آنها در رکوردهایی که مقدار دارند صورت می‌پذیرد. این پارامتر، توسط تابعی بر اساس مقادیر سایر پارامترها محاسبه می‌شود. دومین رویکرد، روش‌هایی را پیشنهاد می‌کند که فقدان برخی داده‌ها را می‌پذیرند. این راه‌حل‌ها ممکن است از مواردی که رکوردهای دارای فیلدهای بدون مقدار را حذف می‌کنند پیشرفته‌تر باشند، اما از مدل‌های آماری ساده‌تر هستند. در دومین روش، پارامترهای فاقد مقدار، با مقادیر متناظر آن‌ها در شبیه‌ترین موارد دریافتی جایگزین می‌شوند.

علت اینکه تنها از روش‌های آماری استفاده نمی‌شود، این است که اولاً، این روش‌ها نیازمند متجانس بودن فضای نمونه هستند. اما نمی‌توان انتظار داشت که تمام بیماران، یک مجموعه متجانس تشکیل بدهند. در عوض، می‌خواهیم جهت اصلی داده‌ها را پیدا کنیم، موارد استثنا را برطرف سازیم و هر یک از این موارد استثنا را به طور مجزا مطالعه نماییم. دومین دلیل عدم اکتفا به روش‌های آماری، این است که این روش‌ها برای یک سیستم اطلاعاتی بسته طراحی شده‌اند. یعنی هیچ دانش خارجی پس از محاسبه مدل آماری روی داده‌هایی که در محاسبات در نظر گرفته شده‌اند، استفاده نمی‌شود. [۲۳]

دسته دیگری از کاربردهای داده‌کاوی، مربوط به اطلاعات تجاری می‌باشد. S. Lessman ،S. F. Crone و R. Stahlbock [۵] بررسی جامعی روی تأثیر استفاده از تکنیک‌های پیش‌پردازش بر داده‌کاوی این نوع اطلاعات انجام داده‌اند. آنها مقالات موجود در این زمینه را از این نقطه نظر که هر کدام از چه روشی برای پیش‌پردازش استفاده می‌کنند تحلیل نمودند. بررسی آن‌ها بر ارزیابی و مقایسه الگوریتم‌های کلاس‌بندی مختلفی که روی مجموعه‌ای از داده‌ها یا وظیفه خاص داده کاوی عمل می‌کنند، تأکید دارد. تنها ۴۷٪ از کل موارد مورد مطالعه از رویکردهای کاهش داده استفاده کرده‌اند، در حالیکه ۶۴٪ آن‌ها از تبدیل داده‌ها استفاده نموده‌اند. تنها یک مورد، اطلاعات را در قالب ویژگی‌های دسته‌ای نمایش داده است، اما متغیرهای دسته‌ای در ۷۱٪ از کل مطالعات، استفاده شده‌اند. عجیب‌ترین نکته اینکه، در کل بررسی‌های انجام شده، بدون ارزیابی گزینه‌های ممکن، تنها یکی از تکنیک‌های پیش‌پردازش مورد استفاده قرار گرفته است.

پایگاه‌داده‌های چندرسانه‌ای، یکی دیگر از زمینه‌های کاربرد فرایند داده‌کاوی می‌باشند. در این پایگاه‌داده‌ها، اشیاء بسیاری وجود دارد که هریک از آنها دارای ابعاد مختلفی هستند. به عنوان مثال، تنها ویژگی رنگ، می‌تواند ۲۵۶ بعد داشته باشد که هربعد، تناوب تکرار یک رنگ داده شده را در تصویر، نشان می‌دهد. این درحالی است‌که تصویر، هنوز می‌تواند ابعاد مختلف دیگری نیز داشته باشد.

انتخاب زیرمجموعه‌ای از ویژگی‌ها، تکنیکی برای کاهش حجم مسئله می‌باشد. همچنین، می‌توان از مجموعه ویژگی‌های موجود، ویژگی‌های جدیدی به‌دست آورد که در حل مسئله داده‌کاوی مفیدتر باشند. به این تکنیک، ساخت/تبدیل ویژگی گفته می‌شود. گسسته‌سازی نیز می‌تواند تعداد مقادیر ممکن برای ویژگی‌های پیوسته را تا حد زیادی کاهش دهد. علت مفید بودن این تکنیک، این است که هر چه تعداد مقادیر ممکن برای ویژگی‌ها بیشتر باشد، فرایند یادگیری، کندتر و با کارایی کمتری صورت خواهد گرفت. همچنین از آنجا که اغلب اختلاف بسیاری بین حداقل و حداکثر مقدار ویژگی‌ها وجود دارد، نرمال‌سازی نیز برای این نوع داده‌ها مفید می‌باشد. [۱۰]

V. Megalokonomou و همکارانش [۱۷] ، که روی داده‌کاوی تصاویر مغز، تحقیق نموده‌اند، از قطعه‌بندی و نرمال‌سازی مکانی تصاویر، استفاده نموده‌اند.آن‌ها عمل کاهش نویز و هموار سازی را نیز روی داده‌ها انجام داده‌اند، اما این‌کار به صورت مجزا قبل از شروع فرایند داده‌کاوی صورت نگرفته است، بلکه به عنوان بخشی از فرایند کلی داده‌کاوی در آن گنجانده شده است.

دسته بعدی از برنامه‌های مورد بررسی، برنامه‌های تشخیص نفوذ[۲۶] هستند. ماهیت زمانی دنباله رویدادها در سیستم‌های کامپیوتری مبتنی بر شبکه، موجب می‌شود معیارهای آماری و زمانی ویژگی‌ها اهمیت زیادی پیدا کنند. حتی ممکن است نیاز باشد ویژگی‌های دیگری با این ماهیت، به داده‌ها اضافه شوند. تکنیک‌های معمول انتخاب ویژگی ممکن است برای این نوع داده‌ها مستقیماً قابل استفاده نباشند، چرا که ماهیت ترتیبی ویژگی‌ها در آن‌ها لحاظ نشده است. Fawcett و Provost ، ایده‌های جالبی در مورد انتخاب خودکار ویژگی برای سیستم کشف تقلب سلولی ارائه نموده‌اند. متد ارائه شده در شناسایی "تقلب تحمیل فوق زائد"[۲۷] که در آن فعالیت‌های تقلب‌آمیز با استفاده از حساب‌های قانونی انجام می‌شوند، بسیار خوب عمل می‌کند. [۱۲]

با توجه به مباحثی که تا کنون به آنها پرداخته شد و مطالعات صورت گرفته، تکنیک‌های پیش‌پردازش مناسب برای هر کاربرد در جدول ۱ بطور خلاصه جمع‌بندی شده‌اند. با تحلیل این جدول می‌توان محور اصلی بخش پیش‌پردازش در فرایند داده‌کاوی را برای هر کاربرد به سادگی مشخص کرد. به عنوان مثال با توجه به جدول، در کاربردهای پزشکی، پاکسازی داده‌ها مهمترین بخش پیش‌پردازش را تشکیل می‌دهد. علت این موضوع همانطور که در ابتدای این بخش بیان شد، حساسیت داده‌های پزشکی و اهمیت بالای دقت در تحلیل این نوع داده‌ها می‌باشد، لذا با ارائه الگوریتم‌های کارا برای پاک‌سازی داده‌ها، می‌توان کیفیت و دقت نتایج حاصل از فرایند داده‌کاوی را در کاربردهای پزشکی تا حد قابل توجهی ارتقا داد.

مثال دیگری از کاربردهای داده‌کاوی که در جدول مشاهده می‌شود، کاربردهای تجاری است. با بررسی جدول، مشاهده می شود مهم‌ترین قسمت فرایند پیش‌پردازش برای این کاربردها، کاهش داده‌هاست. با توجه به ماهیت سیستم‌های اطلاعات تجاری و حجم زیاد داده‌های این

سیستم‌ها، بدیهی است امکان تحلیل سریع و کارای این داده‌ها، وجود ندارد. لذا نیاز به تکنیک‌هایی که بتوانند این حجم عظیم داده‌ها را به‌گونه‌ای مناسب کاهش دهند به شدت احساس می‌شود.

جدول۱-خلاصه تکنیک‌های پیش‌پردازش مناسب برای هر کاربرد

تکنیک / کاربرد	پاکسازی داده‌ها				یکپارچه‌سازی	تبدیل داده‌ها					کاهش داده‌ها		
	MV	NR	OR	I		S.	A.	G.	N.	FC	FS	Sa.	D.
متن		✓						✓			✓		
مکانی					✓				✓	✓	✓		✓
کاربرد وب	✓							✓					
پزشکی	✓		✓	✓					✓		✓		✓
تجاری		✓	✓				✓	✓			✓	✓	✓
چندرسانه‌ای		✓				✓	✓		✓	✓	✓	✓	✓
سیستم تشخیص نفوذ											✓		✓

راهنمای جدول:

I: inconsistency	OL: outlier removal	NR.: noise removal	MV: missing value
N.: normalization	G. : generalization	A.: aggregation	S.: smoothing
D. : discretization	Sa.: sampling	F.S.: feature selection	FC: feature construction

با تعیین مهم‌ترین بخش پیش‌پردازش در هر کاربرد، می‌توان بر ارائه الگوریتم‌های مناسبی برای آن بخش به روش مبتنی بر کاربرد، تمرکز کرد. چنین الگوریتم‌هایی می‌توانند برای کلیه کاربردهای داده‌کاوی از یک نوع، مورد استفاده قرار گیرند.

۴. نتیجه‌گیری

به علت نامناسب بودن داده‌های پایگاه‌داده‌های دنیای واقعی، نمی‌توان به نتایج حاصل از داده‌کاوی روی این داده‌ها اعتماد کرد. بنابراین، برای دستیابی به نتایج مطلوب، نیازمند داده‌های با کیفیت بالاتر هستیم. پیش‌پردازش، گامی مهم در راستای داده‌کاوی موفقیت‌آمیز است. در این مقاله، تکنیک‌های مختلف پیش‌پردازش در داده‌کاوی معرفی شده و خلاصه‌ای از الگوریتم‌های موجود برای هر یک از این تکنیک‌ها بیان شدند. سپس کاربرد هر تکنیک در چندین نوع از پایگاه‌داده‌های دنیای واقعی و تأثیر این تکنیک‌ها بر نتایج به‌دست آمده بررسی شدند. براساس نوع داده‌های این پایگاه‌داده‌ها، میزان اهمیت هر یک از تکنیک‌های پیش‌پردازش برای آن نوع داده مشخص شد. در نهایت، با جمع‌بندی بررسی‌های صورت گرفته و ارزیابی مفیدترین تکنیک‌های پیش‌پردازش در هر کاربرد داده‌کاوی، کارآمدترین تکنیک‌ها برای هر کاربرد تعیین شده و برای استفاده‌های آینده در کاربردهای مشابه، پیشنهاد شدند.

نتایج مطالعات صورت گرفته، بطور خلاصه در جدولی نمایش داده شدند که بر اساس آن، می‌توان برای انتخاب تکنیک‌های پیش‌پردازش در فعالیت‌های داده‌کاوی مختلف، تصمیم‌گیری کرد. از جمله مواردی که به‌عنوان ادامه کار پیشنهاد می‌شود، تولید الگوریتم‌هایی کارایی برای مهم‌ترین تکنیک‌های مناسب در هر نوع داده است که در تمام کاربردهای داده‌کاوی از آن نوع، قابل استفاده باشد.

۵. مراجع

[1] Aounallah M., Quirion S., Mineau G. W., Distributed Data Mining vs. Sampling Techniques: A Comparison, Canadian Conference on AI 2004, 454-460

[2] Arodz T., Training Set Size in Ensemble Feature Selection for Clinical Proteomics Bio-Algorithms and Med-Systems, 1(1/2):107-110, CM UJ, 2005

[3] Astashyn A., Deterministic data reduction methods for transactional data sets, Master thesis, Polytechnic Univ., June 2004

[4] Brönnimann, H., Chen, B., Dash, M., Haas, P., Qiao, Y., Scheuerman, P. a., Efficient data-reduction methods for on-line association rule discovery. In Data Mining: Next Generation Challenges and Future Directions, Selected papers from the NSF Workshop on Next-Generation Data Mining (NGDM '02). MIT Press, Cambridges, MA. 190-208, 2003

[5] Crone S. F., Lessmann S., Stahlbock R., The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing, European Journal of Operational Research, Volume 173, Issue 3, 16 September 2006, P. 781-800

[6] Han j. , Camber A. ,Data Mining: Concepts and Techniques, Second Edition, Morgan Kauffmann publications,2006

[7] Feinerer I., Hornik K., Meyer D. , Text Mining Infrastructure in R, Journal of Statistical Software, March 2008, Volume 25, Issue 5

[8] Frank R., Jin W., Ester M., Efficiently Mining Regional Outliers in Spatial Data, LECTURE NOTES IN COMPUTER SCIENCE, fas.sfu.ca, 2007

[9] John GH., Langley P., Static versus dynamic sampling for data mining, Conference on Knowledge Discovery and Data Mining, cll.stanford.edu, 1996
[10] Kotsiantis S., Kanellopoulos D., Pintelas P., Multimedia mining, WSEAS Transactions on Systems, 2004
[11] Lazarevic A., Fiez T., Obradovic Z., A Software System for Spatial Data Analysis and Modeling, System Sciences, 2000
[12] Lee W., Stolfo S. J., Mok K. W., Adaptive Intrusion Detection: A Data Mining Approach, Artificial Intelligence Review, v.14 n.6, p.533-567, December 1, 2000
[13] Levy A., Logic-Based Techniques in Data Integration, Survey chapter in book "Logic Based Artificial Intelligence", J. Minker (ed.), Kluwer Publishers, 2000
[14] Li X., Data reduction via adaptive sampling, Communications in Information and Systems, intlpress.com, 2002
[15] Li X. B., Jacob V. S., Adaptive data reduction for large-scale transaction data, 2005
[16] Lin J., Haug P. J., Data Preparation Framework for Preprocessing Clinical Data in Data Mining, , AMIA Annual Symposium Proceedings, 2006
[17] Megalooikonomou V., Ford J., Shen L., Makedon F., Saykin, A., Data mining in brain imaging, Statistical Methods in Medical Research, 9, 4, 359-394, 2000
[18] Milley A. H., Seabolt J. D., Williams J. S., Data mining and the case for sampling ,White paper, SAS Institute, Carey, 1998
[19] Mobasher B., Srivastava J., Data Preparation for Mining World Wide Web Browsing Patterns, Journal of Knowledge and Information Systems, 1999
[20] Mooney R. J., Bunescu R., Mining knowledge from text using information extraction. SigKDD Explorations on Text Mining and Natural Language Processing, 2005
[21] Piramuthu S., Evaluating Feature Selection Methods for Learning in Data Mining Applications, Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences-Volume 5, p.294, January 06-09, 1998
[22] Rahm E., Do H. H., Data Cleaning: Problems and Current Approaches, IEEE Bulletin on Data Engineering 23:4, 2000
[23] Vorobieva O., Rumyantsev A., Schmidt R., A CBR Solution for Missing Medical Data, in proceeding of International Conference on Case-Based Reasoning(ICCBR-07), August 15, 2007
[24] Yang J., Olafsson S., Optimization-based feature selection with adaptive instance sampling Computers & Operations Research, Volume 33, Issue 11, November 2006, Pages 3088-3106
[25] Yap G. E., Tan A. H., Pang H., Learning Causal Models for Noisy Biological Data Mining: An Application to Ovarian Cancer Detection. AAAI 2007, p. 354-359
[26] Yiping k., A Survey on Preprocessing Techniques in Web Usage Mining , Computer Science Department The Hong Kong University of Science and Technology , 2003
[27] Yu, L., Liu H., Feature selection for high-dimensional data: a fast correlation-based filter solution. ICML-03, 2003, pp 856-863
[28] Zhao Z., Liu H., Searching for interacting features. In Proceedings of the 20th International Joint Conference on AI (IJCAI-07), 2007

---

[1] Data cleaning
[2] Data integration
[3] Data warehouse
[4] Data transformation
[5] Data reduction
[6] Aggregation
[7] Clustering
[8] Sampling
[9] Outlier
[10] Smoothing
[11] Binning
[12] Data cube
[13] Generalization
[14] Discretization
[15] Fitness
[16] Nested partitions metaheuristic
[17] Association rule
[18] Simple Random Sampling
[19] Finding Association rules from Sampled Transactions
[20] Epsilon Approximation
[21] Spatial mining
[22] Spatial scan statistic
[23] Web usage mining
[24] Screening
[25] Support vector machine
[26] Intrusion detection
[27] Superimposition fraud