Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 9, Number 3, Fall 2020, pp. 109-120
Received: 10/21/2017, Revised: 07/15/2018, Accepted: 04/22/2020

Proposing a process to integrate and identify repetitions to improve the quality of data

Niloofar Mollamohammad[1], Negin Daneshpour[2]
1- Faculty of Computer and Information Technology Engineering, Qazvin Branch,Islamic Azad University, Qazvin, Iran.
2*- Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.
[1]nlfrmhmd@gmail.com, [2*] ndaneshpour@sru.ac.ir

Abstract- Recently, information in the workplace and decision has a major role. Due to the importance of deciding, it is also necessary to ensure data quality. Data quality can be improved by data cleaning methods. In this research, we propose a process for discovering duplications and contradictory types of records, integrating and identifying duplications to improve the quality of data. Our proposed process consists of different activities. These activities are coding records, clustering by expectation maximization algorithm, making token for records, integrate coding records methods and making token for records methods, and extracting association rules by Fp-growth Algorithm. The results of the tests show that the proposed process has averaged 96% recall, 99% precision, 95% accuracy and 95% f-score. The proposed method is compared with a duplication and error detection method. The results indicate an increase of 13% for recall, 1% for accuracy and 6% for f-score in the proposed process.

Keywords- Data quality, Data quality factors, Data cleaning

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۹، شماره ۳، پائیز ۱۳۹۹ ، صفحه ۱۰۹-۱۲۰
دریافت: ۱۳۹۶/۰۷/۲۹، بازنگری: ۱۳۹۷/۰۴/۲۴، پذیرش: ۱۳۹۹/۰۲/۰۳

ارائه فرایندی جهت یکپارچه‌سازی و تشخیص تکرار برای بهبود کیفیت داده‌ها

نیلوفر ملامحمد[۱]، نگین دانشپور*[۲]
۱- دانشکده مهندسی برق و فناوری اطلاعات، واحد قزوین، دانشگاه آزاد اسلامی،قزوین، ایران.
*۲- دانشکده مهندسی کامپیوتر، دانشگاه تربیت دبیر شهید رجایی، تهران، ایران.
[1]nlfrmhmd@gmail.com, [2*]ndaneshpour@srttu.edu
تاریخ دریافت: تاریخ بازنگری: تاریخ پذیرش:
* نشانی نویسنده مسئول: نگین دانشپور، تهران، لویزان، خیابان شعبانلو، دانشگاه تربیت دبیر شهید رجایی، دانشکده مهندسی کامپیوتر، کد پستی: ۱۶۷۸۸-۱۵۸۱۱

چکیده- اطلاعات در محیط‌های کاری امروزی و تصمیم‌گیری‌ها نقشی اساسی دارند. با توجه به اهمیت تصمیم‌گیری، اطمینان از کیفیت داده‌های موجود ضروری است. با استفاده از روش‌های پاک‌سازی داده می‌توان کیفیت داده‌ها را بهبود بخشید. در این مقاله فرایندی در جهت کشف انواع رکوردهای تکراری و متناقض، یکپارچه‌سازی و تشخیص تکرار برای بهبود کیفیت داده‌ها ارائه می‌شود. فرایند پیشنهادی شامل بخش‌هایی ازجمله کد کردن داده‌ها و خوشه‌بندی با استفاده از الگوریتم امید ریاضی- بیشینه‌سازی، ساخت نشانه برای رکوردها، ادغام روش‌های کدکردن داده‌ها و ساخت نشانه و ایجاد قوانین انجمنی با استفاده از الگوریتم Fp-growth است. نتایج آزمایش‌ها نشان می‌دهد در فرایند پیشنهادی به‌طور متوسط معیار فراخوانی ۹۶٪، صحت ۹۹٪، دقت ۹۵٪ و امتیاز- اف ۹۵٪ شده است. روش پیشنهادی با یک روش شناسایی تکرار و خطا، مقایسه شده است که نتایج حاصل نشان‌دهنده‌ی افزایش ۱۳٪ فراخوانی، ۱٪ صحت و ۶٪ امتیاز- اف است.

واژه‌های کلیدی- کیفیت داده، عوامل کیفیت داده‌ها، پاک‌سازی داده‌ها

۱- مقدمه

امروزه اطلاعات با سرعت بالایی تولید می‌شوند. برای دستیابی به بهترین سطح تصمیم‌گیری در راستای سودآوری کسب‌وکار، دسترسی به داده‌ها در سطحی مناسب و به روشی تعاملی، اغلب به‌صورت یک رویا برای مدیران اجرایی کسب‌وکار و سایر مدیران است که با استفاده از پایگاه‌داده‌تحلیلی محقق می‌شود [۱]. اطمینان از کیفیت اطلاعات پایگاه‌داده‌تحلیلی ضروری است و با کیفیت داده‌های موجود در آن ارتباط مستقیم دارد. کیفیت داده‌ها اشاره به وجود داده‌های عاری از خطاهای تایپی، بدون تکرار، بدون گم‌شدگی و عدم وجود داده‌های ناقص دارد. در صورت عدم اعتماد به کیفیت داده‌ها، اطمینان از دانش استخراج شده از آن‌ها نیز امکان‌پذیر نیست [2]. عوامل متفاوتی از قبیل خطا در ورود داده‌ها، خطا در انتقال داده‌ها، ابزارهای استفاده‌شده خطادار برای جمع‌آوری داده‌ها و تناقض در قراردادهای نام‌گذاری موجب کاهش کیفیت داده‌ها می‌شود [۳]. کیفیت داده‌ها را می‌توان با استفاده از تکنیک‌های پاک‌سازی داده‌ها افزایش داد. پاک‌سازی داده‌ها ساختار و محتوایی ضروری و اولین گام در ایجاد پایگاه‌داده‌تحلیلی باکیفیت است [۴ و ۵]. پاک‌سازی داده، تنظیف یا حذف داده نیز نامیده می‌شود [۶] که به مشکلات داده‌ها پس از وقوع آن‌ها می‌پردازد. این فرایند شامل فعالیت اصلاح و یا تعدیل اطلاعات گم‌شده، نادرست، ناتمام و با فرمت‌های نامناسب [۷]؛ حذف رکوردهای تکراری و فیلتر کردن داده‌های بد است [۸]. همچنین مسئولیت خطایابی، رفع خطا و ناهمسانی‌های داده‌ها به‌منظور ارتقاء کیفیت داده را بر عهده دارد [۶] که منجر به بهبود صحت و استفاده موثر از داده‌های موجود می‌شود [۹]. پاک‌سازی داده‌ها دو

بخش تشخیص و تصحیح ناهنجاری‌های موجود در داده‌ها را در بر می‌گیرد.

راهکارهای متفاوتی برای تشخیص و تصحیح خطاهای موجود در داده‌ها ارائه شده است که از روش‌های قوانین [۱۰ و ۱۱ و ۱ و ۱۲]، الگوریتم‌های تجربی [۱۳]، خوشه‌بندی [۱۴]، مدل کاربر [۲۷]، الگوریتم‌های استنتاجی [۱۶]، رویکردهای آماری [۱۷]، جنگل تصادفی و نمونه‌برداری [۱۸]، پرس‌وجو محور [۱۹ و ۲۰] و پاک‌سازی تعاملی [۲۱] استفاده کرده‌اند. کاستی‌های موجود در برخی روش‌های ارائه‌شده، شامل محدودیت در تشخیص برخی از خطاها، پیچیدگی بالا، توسعه ناپذیری، عملکرد نامناسب، زمان اجرای طولانی، محدودیت در نوع داده و سطح بالای وابستگی به کاربر است.

هدف اصلی این مقاله ارائه فرایندی در جهت کشف انواع رکوردهای تکراری و متناقض در انواع داده‌های موجود است که با کاربر تعامل داشته باشد.

رکوردهای تکراری و متناقض مورد نظر عبارتند از:

- دو یا چند رکورد کاملا مشابه در مجموعه داده‌ها
- رکوردهایی که چند ویژگی مشابه دارند اما در برخی از ویژگی‌ها هم کاملا متفاوت هستند.
- رکوردهایی که چند ویژگی مشابه دارند اما در برخی از ویژگی‌ها مقداری متفاوت دارند، به عنوان مثال در یکی از رکوردها ویژگی نام به صورت Ali نوشته شده است و در دیگری به صورت Alee نوشته شده است.

فرایند پیشنهادی شامل هشت بخش اصلی انتخاب ویژگی، یکسان‌سازی داده‌ها، کد کردن داده‌ها، ساخت نشانه، ادغام بخش‌های کدکردن داده‌ها و ساخت نشانه، ساخت متغیر واحد، ایجاد قوانین انجمنی، شناسایی رکوردهای تکراری و متناقض و اطلاع به خبره است. در بخش کد کردن داده‌ها نوآوری‌هایی ازجمله خوشه‌بندی داده‌ها با استفاده از الگوریتم امید ریاضی-بیشینه‌سازی[۱] و همچنین روشی برای مشخص کردن رکوردهای نسبتاً مشابه و میزان شباهت میان آن‌ها پیشنهاد می‌شوند. نوآوری بخش نشانه سازی، ساخت نشانه برای هر رکورد است. سپس بخش کد کردن داده‌ها و ساخت نشانه با یکدیگر تلفیق می‌شوند. متغیر R-T برای شناسایی رکوردهای تکراری و متناقض پیشنهاد می‌شود. در فرایند پیشنهادی، قوانین انجمنی با استفاده از الگوریتم

Fp-growth ساخته می‌شوند. با توجه به قوانین ساخته شده کشف نمونه‌های تکراری صورت می‌گیرد. روش پیشنهادی بر روی سه مجموعه داده مشهور Restaurant[۲]، Employee[۳] و Adult[۴] با اندازه‌های متفاوت پیاده‌سازی شده است و پس از محاسبه معیارهای سنجش کارایی، مشاهده شد که به‌طور متوسط معیار فراخوانی[۵] ٪۹۶، دقت[۶] ٪۹۵، صحت[۷] ٪۹۹ و امتیاز-اف[۸] ٪۹۵ به دست می‌آید.

ساختار کلی مقاله بدین شرح است: در بخش دوم مقاله شرحی از کارهای گذشته و کاستی‌های آن‌ها بیان‌شده است؛ در بخش سوم فرایند پیشنهادی آورده شده است؛ در بخش چهارم پیاده‌سازی فرایند پیشنهادی و نهایتاً در بخش پنجم نتیجه‌گیری و پیشنهادی در جهت کارهای آتی آورده شده است.

۲- پیشینه تحقیق

در حوزه پاک‌سازی داده‌ها تحقیقاتی انجام شده و روش‌هایی جهت یکپارچه‌سازی، تشخیص و تصحیح خطای موجود در مجموعه داده‌ها ارائه شده است؛ که با استفاده از آن‌ها کیفیت داده‌ها بهبود می‌یابد.

این روش‌ها به چهار دسته تقسیم می‌شوند:

- شناسایی و حذف تکرار
- شناسایی و پرکردن داده‌های ناکامل
- شناسایی و رفع ناسازگاری داده‌ها
- شناسایی و حذف داده‌های پرت

وی و همکاران (۲۰۰۷) روش پاک‌سازی داده‌ها را بر اساس قوانین انجمنی ارائه دادند. در این روش از قوانین پایه کسب‌وکار، قوانین استخراج‌شده کسب‌وکار و قوانین کسب‌وکار پیشرفته استفاده شده است. این روش امکان جمع‌آوری داده‌ها از منابع چندگانه بزرگ را فراهم می‌کند و باعث افزایش دقت می‌شود [۲۸].

جو (۲۰۰۷)، ترکیبی از تکنیک‌هایی بر پایه جنگل تصادفی و روش‌های نمونه‌برداری ارائه کرد که شامل دو مرحله عمده پاک‌سازی داده‌ها و طبقه‌بندی بر اساس جنگل تصادفی است. روش پیشنهادی عملکرد قابل قبولی دارد [۱۸].

جباملار و همکاران (۲۰۰۸) چارچوبی را با شش مرحله ارائه دادند که شامل انتخاب ویژگی‌ها، ساخت نشانه، انتخاب الگوریتم خوشه‌بندی، محاسبه شباهت برای ویژگی‌های انتخاب‌شده، انتخاب تابع حذف و ادغام است. هدف اصلی چارچوب ارائه‌شده از پاک‌سازی داده‌ها، کاهش زمان و پیچیدگی فرایند کاوش و افزایش کیفیت داده‌ها در پایگاه‌داده‌تحلیلی است. ویژگی‌های این چارچوب تعامل با کاربر، درک آسان و عملکرد قابل‌قبول است. این چارچوب تنها به حذف رکوردهای تکراری محدود است. پیاده‌سازی این چارچوب دشوار است و توسعه‌پذیر نیست [۲۴].

آرورا و همکاران (۲۰۰۹) الگوریتمی ارائه کردند که ابزاری خودکار برای تشخیص و شناسایی داده‌های غیرقابل استفاده در مجموعه داده‌ها و پایگاه‌داده‌تحلیلی است. این الگوریتم بر اساس قوانین انجمنی ریاضی است. [۱۰].

یو و همکاران (۲۰۰۹) چارچوب پاک‌سازی داده‌ها را بر اساس مدل کاربر و با توجه به نیازهای سیستم‌های اطلاعاتی فعلی ارائه دادند. چارچوب دارای پیاده‌سازی آسان، توسعه‌پذیری و عملکرد قابل‌قبول است، اما اجازه انتخاب ویژگی‌ها را نمی‌دهد و حتماً لازم است تمام ویژگی‌ها تجزیه‌وتحلیل شوند؛ که درواقع باعث پردازش‌های طولانی‌مدت و اتلاف زمان و هزینه است [۲۷].

میفیلد و همکاران (۲۰۰۹)، روشی ارائه کردند که با استفاده از الگوریتمی استنتاجی تقریباً کارآمد، به‌راحتی در استانداردهای سیستم‌های مدیریت پایگاه داده اجرا شده است و مقیاس خوبی برای پایگاه داده‌ها با اندازه بزرگ است. روش پیشنهادی دقت و انعطاف‌پذیری قابل قبولی دارد [۱۶].

حماد و همکارانش (۲۰۱۱) الگوریتمی بهبودیافته برای پاک‌سازی داده‌ها پیشنهاد دادند که تعامل با کاربر را با انتخاب قوانین هر منبع و اهداف موردنظر فراهم می‌کند. الگوریتم پیشنهادی تعاملی، توسعه‌پذیر، دارای عملکرد قابل‌قبول، پیاده‌سازی آسان و سرعت اجرای مناسب است. این الگوریتم گزارشی از انواع خطاهای تشخیص داده‌شده با توجه به منابع داده ارائه می‌کند. این الگوریتم بر کیفیت داده‌ها توجه دارد با این وجود بر زمان انجام آن تمرکز نشده است [۱۱].

کاویتهاکومار و همکاران (۲۰۱۱) چارچوبی را ارائه کردند. این چارچوب از تکنیک‌های پیش‌پردازش در جهت انتخاب ویژگی‌ها و از روش‌های خوشه‌بندی برای انتخاب مجموعه‌ای از رکوردها با توجه به کلیدهای خوشه بهره می‌گیرد. همچنین میزان شباهت‌ها با استفاده از تکنیک‌های داده‌کاوی محاسبه می‌شود. [۲۶].

خان و همکاران (۲۰۱۲) الگوریتمی را برای شناسایی و حذف رکوردهای تکراری در پایگاه‌داده‌تحلیلی و پیرو آن بهبود کیفیت داده‌ها و عملکرد پایگاه‌داده‌تحلیلی ارائه کردند. این الگوریتم از سه مرحله تبدیل، خوشه‌بندی و تطبیق برای شناسایی رکوردهای کاملاً تکراری و نسبتاً تکراری استفاده می‌کند [۳۲].

ریپون و همکاران (۲۰۱۲) تکنیک جدید دامنه مستقل را برای تطبیق بهتر رکوردهای مشابه و تکراری ارائه کردند. آن‌ها الگوریتمی برای تشخیص موارد تکراری در محیطی مستقل از دامنه پیشنهاد دادند. الگوریتم پیشنهادی قادر به تشخیص تکرارهایی با معنای مشابه و ظاهر متفاوت نیست و نیازمند بهبود روش‌های مرتب‌سازی در جهت رفع این کاستی است [۲۳].

پائول و همکارانش (۲۰۱۲) روشی ترکیبی برای پاک‌سازی داده در پایگاه‌داده‌تحلیلی ارائه کردند که شامل دو بخش اعمال الگوریتم پی‌ان‌اراس بر برخی از ویژگی‌ها و خطاهای فاحش و حذف آن‌ها و اعمال الگوریتم بسته بودن متعدی اصلاح شده برای حذف رکوردهای تکراری و پر کردن رکوردهای خالی است [۳۱].

اوهنگوو و همکاران (۲۰۱۳) الگوریتم پاک‌سازی بر اساس نشانه تعریف کرده‌اند که از چهار مرحله انتخاب و رتبه‌بندی ویژگی‌ها، استخراج و تشکیل نشانه، مرتب‌سازی نشانه‌ها، تشخیص تکرار، حذف و ایجاد شناسه‌های پایگاه‌داده‌تحلیلی تشکیل‌شده است [۲۵].

رحمان و همکاران (۲۰۰۹) الگوریتمی برای تشخیص رکوردهای تکراری بر پایه الگوریتم‌های تجربی ارائه کردند. الگوریتم پیشنهادی دقت و فراخوان مناسبی را فراهم کرده‌است و پیچیدگی بالایی دارد [۲۲].

هشینو و همکاران (۲۰۱۵) رویکردی آماری برای پاک‌سازی داده‌ها بر پایه محدودیت ارائه کردند. این رویکرد رکوردهایی را که با قوانین وابستگی تابعی یا وابستگی تابعی شرطی متضاد هستند شناسایی می‌کند. این رویکرد مقیاس‌پذیری خطی دارد و دقت تشخیص خطای مناسبی را فراهم کرده است [۱۷].

برگمن و همکارانش (۲۰۱۵) روش گیو او سی او، پاک‌سازی جستجو محوری را ارائه دادند. در این دیدگاه نمایش‌های محقق شده به‌عنوان محرکی برای شناسایی اطلاعات غلط یا گمشده استفاده‌شده است. پرس‌وجوهای کاربران، دیدگاه‌های مرتبط و متمرکز از اساس پایگاه‌داده را فراهم می‌کند، بنابراین کشف خطا را تسهیل می‌کند. این الگوریتم به‌طور مؤثر فضای جست‌وجو را هرس می‌کند و میزان تعامل و برهم‌کنش با جمعیت را به حداقل می‌رساند. این روش به وجود افراد خبره در مراحل مختلف اجرا وابستگی بسیاری دارد [۱۹ و ۲۰].

جدول ۱ فرایندهای بررسی شده و دسته قرارگیری هر یک را نشان می‌دهد.

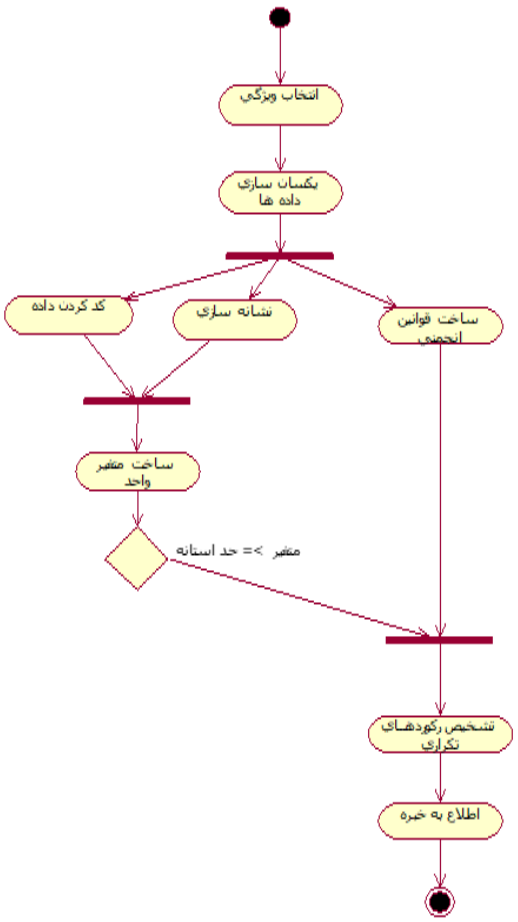
جدول ۱: فرایندها و کارایی آن‌ها

فرایند	شناسایی و حذف تکرار	شناسایی و پرکردن داده‌های ناکامل	شناسایی و رفع ناسازگاری داده‌ها	شناسایی و حذف داده‌های پرت
وی و همکاران (۲۰۰۷)			✓	
جو (۲۰۰۷)			✓	
جباملار و همکاران (۲۰۰۸)	✓			
آرورا و همکاران (۲۰۰۹)			✓	
یو و همکاران (۲۰۰۹)			✓	
میفیلد و همکاران (۲۰۰۹)		✓		
حماد و همکاران (۲۰۱۱)	✓	✓	✓	
کاویتهاکومار و همکاران (۲۰۱۱)		✓		
خان و همکاران (۲۰۱۲)	✓			
ریپون و همکاران (۲۰۱۲)	✓			
پائول و همکاران (۲۰۱۲)	✓	✓		
اوهنگوو و همکاران (۲۰۱۳)	✓			
رحمان و همکاران(۲۰۱۵)	✓			
هشینو و همکاران (۲۰۱۵)	✓	✓		
برگمن و همکاران (۲۰۱۵)		✓	✓	

این روش‌ها در تشخیص تعداد و نوع خطاها، توسعه‌پذیری، ارتباط با کاربر، سرعت اجرا و پیچیدگی متفاوت هستند. هدف این مقاله ارائه راهکاری با عملکرد مناسب است که به تشخیص تکرار در انواع داده‌های موجود می‌پردازد و در ارتباط با کاربر است.

۳- روش پیشنهادی

در این بخش فرایندی برای یکپارچه‌سازی، تشخیص تکرار و تشخیص خطا در مجموعه داده‌ها پیشنهاد می‌شود. فرایند پیشنهادی شامل هشت بخش اصلی است که نمودار فعالیت آن در شکل (۱) آمده است.



شکل ۱: نمودار فعالیت رویکرد پیشنهادی

۳-۱- انتخاب ویژگی‌های مورد نظر

سازمان‌ها منابع داده‌ای بسیاری دارند و مجموعه داده‌هایی با ویژگی‌های متفاوتی را دریافت می‌کنند. مجموعه داده‌های دریافتی می‌تواند شامل میلیون‌ها رکورد و صدها ویژگی باشد. انتخاب ویژگی‌ها در مقایسه میان رکوردها بسیار مهم است [۳۳]. مرحله شناسایی و انتخاب ویژگی‌ها یکی از مراحل موردنیاز در رویکرد پیشنهادی است. این مرحله به‌عنوان پایه و اساسی برای تمامی بخش‌های باقی‌مانده در نظر گرفته شده است. ویژگی‌های مجموعه داده خود ممکن است تناقضات و زوائدی را دربر داشته باشند. به‌عنوان‌مثال امکان استفاده از نام‌های متفاوت برای نشان دادن یک ویژگی و یا یک نام برای نمایش ویژگی‌های متفاوت وجود دارد؛ علاوه بر این در برخی موارد سازمان‌ها به تجزیه‌وتحلیل چند ویژگی از دیتاهای خود در جهت تصمیم‌گیری نیاز دارند. در این مرحله ویژگی‌های مجموعه داده بررسی و شناسایی شده‌اند و با توجه به نیاز کسب‌وکار، ویژگی‌های مورد نیاز برای تشخیص خطا توسط خبره انتخاب شده است. ویژگی‌های انتخاب شده به عنوان

ورودی وارد فرایند شده‌اند تا پس از پاک‌سازی، سازمان‌ها از طریق آن‌ها تصمیمات مؤثری در پیشبرد اهداف کسب‌وکارشان بگیرند.
انجام مرحله انتخاب ویژگی‌ها سبب کاهش زمان و تلاش مورد نیاز برای انجام سایر مراحل می‌شود.

۳-۲- یکسان‌سازی داده‌ها

منظور از یکسان‌سازی، انجام اصلاحات بر روی مجموعه داده است تا در مراحل بعدی بدون وجود مشکلی از آن بهره گرفته شود. اصلاحات انجام‌شده شامل از میان بردن کاراکترهای اضافی و نگارشی است. علاوه بر این شامل تبدیل تمام حروف به‌صورت حروف بزرگ و یا کوچک است. تبدیل حروف برای جلوگیری از ایجاد مشکل در مرحله کدسازی داده‌ها و تخصیص عدد به آن‌ها انجام شده است. داده‌های یکسان‌سازی شده ذخیره می‌شوند و از آن‌ها به‌عنوان ورودی بخش‌های کد کردن داده‌ها، نشانه سازی و ایجاد قوانین انجمنی استفاده می‌شود.

۳-۳- کد کردن داده‌ها

در این بخش رکوردهای کاملاً تکراری مشخص می‌شود و جدولی نیز از رکوردهای نسبتاً تکراری و میزان شباهت میان آن‌ها ایجاد می‌شود. در بخش کد کردن داده‌ها نوآوری‌هایی ازجمله خوشه‌بندی داده‌ها با استفاده از الگوریتم امید ریاضی-بیشینه‌سازی انجام می‌شود. علاوه بر این روشی برای مشخص کردن رکوردهای نسبتاً مشابه و میزان شباهت میان آن‌ها پیشنهاد می‌شود. بخش کد کردن داده‌ها شامل سه مرحله است که در ادامه شرح داده می‌شوند

- تبدیل رکوردها به فرمت عددی

هدف از انجام این مرحله تبدیل تمام ویژگی‌های داده‌ای موجود با فرمت‌های متفاوت متنی و عددی به مقادیر عددی و سپس استفاده از آن مقادیر و تولید عددی برای هر کدام از رکوردهای موجود در مجموعه داده است. سپس عملیات خوشه‌بندی بر روی اعداد هررکورد انجام می‌شود. شرح فعالیت این مرحله بدین‌صورت است: ابتدا داده‌های یکسان‌سازی شده از مرحله قبل به عنوان ورودی این مرحله دریافت می‌شوند. سپس فرمول مبنا که در پژوهش‌های گذشته مورد استفاده قرار گرفته است و نتایج مناسبی ایجاد کرده است [۳۲]، بر روی تمام ویژگی‌های داده اعمال می‌شود. نتیجه‌ی حاصل، مقادیر عددی منحصربه‌فرد برای تمام ویژگی‌های مجموعه داده است. با توجه به مقادیر به‌دست‌آمده به هرکدام از ویژگی‌های هر رکورد، ستونی اختصاص داده می‌شود که در آن مقادیر عددی مرتبط با تمام ویژگی‌های هر رکورد در کنار یکدیگر آورده می‌شوند. بار دیگر فرمول مبنا بر هر یک از ویژگی‌های ستون ایجاد شده اعمال می‌شود و در نتیجه برای هر رکورد عددی ساخته می‌شود. فرمول مبنا و متغیرهایش در (۱) آمده است.

$$\text{alphvalue} = \begin{cases} 0-9 \\ \text{aA}=10, \text{bB}=11, \dots, \text{zZ}=35 \end{cases}$$

$$\text{m} = \text{Large prime number}$$

$$Radix => 36$$

$$\sum[((\text{radix})^{\text{position}} * \text{alphvalue}) \bmod \text{m}] \qquad (1)$$

به ارقام ۰-۹ مقادیر صفر تا نه و به حروف الفبا aA تا zZ مقادیر ۱۰ تا ۳۵ اختصاص داده می‌شود. m به عنوان یک عدد اول بزرگ ۷۳۱ در نظر گرفته می‌شود. در فرمول، position نشان‌دهنده شماره کاراکتر مورد محاسبه است که ارزش آن از راست به چپ و از صفر تا تعداد کاراکترها، شماره‌گذاری می‌شود. به عنوان مثال اگر ویژگی نام شامل داده ASIM بوده باشد نحوه تبدیل آن به مقدار عددی در ادامه آمده است.

$$\{(((36)^3 * 10) \bmod 731) + (((36)^2 * 28) \bmod 731) + ((36)^1 * 18) \bmod 731 + ((36^0) * 22) \bmod 731)\}$$
$$= 182 + 469 + 648 + 22$$
$$= 1321$$

- خوشه‌بندی

هدف از مرحله خوشه‌بندی، تقسیم مجموعه داده به گروه‌هایی از داده‌ها است و این کار بر روی مقادیر ایجاد شده برای هر رکورد صورت می‌گیرد. در این مرحله الگوریتم خوشه‌بندی بر روی ستونی شامل اعداد تولید شده برای هر رکورد اعمال می‌شود. با توجه به مطالعات صورت گرفته در حوزه الگوریتم‌های خوشه‌بندی موجود، از الگوریتم امید ریاضی- بیشینه‌سازی استفاده می‌شود. این الگوریتم شامل ویژگی‌هایی از جمله پایه آماری قوی، نیرومند در مواجهه با داده‌های نویزی، قبول کردن تعداد زیاد خوشه به‌عنوان ورودی، مدیریت داده‌ها با ابعاد بالا، همگرایی سریع با مقداردهی اولیه مناسب و کیفیت مناسب همراه با مجموعه داده‌های بزرگ است [۱۵]. در انتهای مرحله خوشه‌بندی، هر کدام از رکوردها با توجه به عددشان در خوشه‌ای قرار می‌گیرند. خوشه‌های تولید شده، با توجه به تعدادشان برچسب‌گذاری می‌شوند. خوشه‌بندی داده‌ها باعث کاهش تعداد مقایسه‌های موردنیاز می‌شود زیرا هر رکورد به‌جای مقایسه با تمام رکوردهای موجود در مجموعه داده، تنها با رکوردهای موجود در خوشه خود مقایسه می‌شود.

- تطبیق

پس از اجرای مرحله خوشه‌بندی، با توجه به مجموعه داده مورد نظر، تعدادی خوشه ایجاد می‌شود. هر خوشه مورد بررسی قرار می‌گیرد، بدین ترتیب که رکوردهای موجود در هر خوشه دوبه‌دو مقایسه می‌شود. برای انجام مقایسه میان رکوردهای هر خوشه، تعداد ویژگی‌های موجود در مجموعه داده‌ها در نظر گرفته می‌شود و سپس رکوردهای هر خوشه دوبه‌دو برای یافتن تعداد ویژگی‌های یکسان بررسی می‌شوند. مقایسه میان ویژگی‌های رکوردها از طریق مقایسه میان اعداد منحصربه‌فرد تولید شده برای هر ویژگی رکورد در مرحله تبدیل رکوردها به فرمت عددی، انجام می‌شود و تعداد اعداد برابر ذخیره می‌شود. به عدد ذخیره شده، عدد تشابه گفته می‌شود. عدد تشابه در بیشترین حالت برابر با تمام ویژگی‌های مجموعه داده است، به این معنا که تمام ویژگی‌های دو رکورد، یکسان بوده و دو رکورد کاملاً با یکدیگر مشابه هستند. برای پیدا کردن میزان تشابه میان رکوردها حد آستانه‌ای تعریف شده است و با مقایسه میان عدد تشابه و حد آستانه، رکوردهای مشابه و میزان شباهت میان آن‌ها مشخص شده و در جدولی به نام CR (Coding Records) ذخیره می‌شود.

۳-۴- نشانه سازی

هدف از پیشنهاد این بخش پیدا کردن رکوردهای مشابه، خطادار و ساخت مجدد جدولی نشان دهنده‌ی میزان این شباهت به روشی متفاوت است. در این قسمت برای هر یک از ویژگی‌های موجود در مجموعه داده نشانه‌ای ساخته می‌شود و سپس با استفاده از نشانه‌های ساخته شده میزان شباهت میان رکوردها مشخص می‌شود. نتایج حاصل از آن در جدولی ذخیره شده و در بخش‌های بعدی از آن استفاده می‌شود. بخش نشانه سازی شامل سه مرحله است که در ادامه توضیح داده می‌شود.

- **ساخت نشانه**

در این مرحله، بر روی داده‌های یکسان‌سازی شده و ویژگی‌های انتخاب شده از مجموعه داده، نشانه‌های مناسب ساخته می‌شود. هرکدام از ویژگی‌های مجموعه داده شامل نشانه هستند. با توجه به نوع و ساختار ویژگی‌های داده، نشانه‌ها می‌توانند تک جزئی و چند جزئی باشند. سه نوع نشانه به همراه حدودشان تعریف می‌شود.

نشانه‌های الفبایی: این نشانه‌ها اغلب برای نمایش ویژگی‌های نام مانند نام، نام خانوادگی، نام مشتری، عنوان کتاب و غیره مورد استفاده قرار می‌گیرد. مقادیر دربرگیرنده‌ی این نوع نشانه Aa-Zz است [۲۵ و ۳۴].

نشانه‌های عددی: این نشانه‌ها اغلب برای نمایش ویژگی‌های شماره مانند شماره تلفن، شماره خیابان، شماره آپارتمان و غیره مورد استفاده قرار می‌گیرد. مقادیر دربرگیرنده‌ی این نوع نشانه ۰-۹ است [۲۵ و ۳۴].

نشانه‌های الفبایی عددی: این نشانه‌ها اغلب برای نمایش ویژگی‌هایی مانند آدرس، کد محصول و نظایر آن مورد استفاده قرار می‌گیرد. مقادیر این نوع نشانه شامل هم مقادیر الفبایی است و هم عددی [۲۵ و ۳۴].

برای ساخت نشانه، برای ویژگی‌هایی که داده‌های یک جزئی را شامل شده‌اند؛ اولین کاراکتر موجود در ویژگی در نظر گرفته می‌شود. برای ویژگی‌های چند جزئی ابتدا کاراکتر اول در نظر گرفته شده و سایر کاراکترها بررسی می‌شوند تا به کاراکتر فاصله رسیده شود، سپس کاراکتر پس از فاصله در کنار کاراکتر اول قرار داده شده و به همین پس از فاصله در کنار کاراکتر اول قرار داده شده و به همین صورت ادامه ویژگی بررسی می‌شود. به‌عنوان مثال ویژگی نام(ASIM) به‌صورت A و ویژگی آدرس (4161 NORTH SIXTH STREET) به‌صورت 4NSS نشانه‌گذاری شده است. پس از ساخت نشانه برای هر ویژگی، نشانه‌های ویژگی‌های هر رکورد در کنار هم قرار می‌گیرند و برای هر رکورد نشانه‌ای ساخته می‌شود؛ به‌عنوان مثال اگر رکورد مورد نظر شامل ویژگی‌های ذکر شده باشد، نشانه رکورد به‌صورت A4NSS ساخته می‌شود. با توجه به وجود نشانه برای هر رکورد، سرعت محاسبه فاصله میان رکوردها و تطابق آن‌ها کاهش یافته است.

- **محاسبه فاصله میان رکوردها و تطابق**

پس از ساخت نشانه‌ی هر رکورد، برای مقایسه شباهت میان رکوردها، نشانه‌های ساخته‌شده برای دو رکورد موردنظر، بررسی شده و بر روی آن‌ها تابع لون‌اشتاین اعمال می‌شود. در روش لون‌اشتاین، فاصله بین دو رشته، به‌وسیله کمترین تعداد عملیات مورد نیاز برای تبدیل یک رشته به رشته دیگر معین می‌شود. عملیات مجاز در این روش یکی از عملیات درج، حذف یا جایگزینی است [۳۵]. قدرت لون‌اشتاین در توانایی انتخاب مناسب‌ترین عملیات برای تبدیل یک رشته به رشته دیگر است به‌گونه‌ای که تعداد کل عملیات، کمترین مقدار شود. هر چه عدد فاصله لون‌اشتاین کمتر باشد به معنی تلاش کمتر برای تبدیل رکوردها به یکدیگر است. این به معنای امکان وجود رکوردهای مشابه و خطادار است. نتایج مقایسه رکوردها و میزان شباهت میان

آن‌ها که از طریق فاصله لون‌اشتاین به‌دست‌آمده است، در جدولی به نام MT (Making Token) ذخیره می‌شود.

۳-۵- ترکیب نتایج و معرفی متغیر جدید

در این بخش متغیری با توجه به نتایج به‌دست‌آمده از مراحل کدسازی داده‌ها و نشانه سازی پیشنهاد می‌شود. با توجه به اینکه هر کدام از مراحل کد کردن داده‌ها و نشانه‌سازی به تنهایی قادر به شناسایی بخشی از رکوردهای تکراری بوده‌اند، ترکیب نتایج حاصل از آن‌ها باعث شناسایی دقیق‌تر تکرارها می‌شود.

در بخش کد کردن داده‌ها جدولی با نام CR ایجاد شد. جدول CR بر اساس تشابه میان رکوردها برمبنای عدد تشابه میان آن‌ها به‌دست‌آمد. هر چه عددی که در جدول CR موجود است بیشتر باشد، اختلاف میان رکوردها کمتر است و شباهت میان رکوردها بیشتر است. در بخش نشانه سازی نیز جدولی با نام MT ایجاد شد. جدول MT بر اساس تشابه میان رکوردها بر اساس فاصله لون‌اشتاین میان آن‌ها به دست آمد. هر چه عددی که در جدول MT موجود است کمتر باشد اختلاف میان رکوردها کمتر است و رکوردها شبیه‌تر هستند. در این بخش متغیری با نام R-T معرفی می‌شود که از تقسیم اعداد موجود در CR بر اعداد موجود در MT به‌دست می‌آید و در فرمول (۲) ارائه می‌شود.

$$R - T = \frac{CR}{MT} \qquad (2)$$

با توجه به توضیحات ذکر شده و اینکه هر چه عدد جدول CR بیشتر و عدد جدول MT کمتر باشد مطلوب‌تر است، می‌توان نتیجه گرفت که هرچه متغیر R-T بزرگ‌تر باشد تشابه میان رکوردها نیز بیشتر است.

۳-۶- ایجاد قوانین انجمنی

با توجه به پیشنهاد متغیر R-T در مرحله پیش، توانایی شناسایی رکوردهای مشابه فراهم می‌شود. هدف از انجام این بخش پیدا کردن الگوی تکرار رکوردها و قوانین موجود میان آن‌ها است. با کمک قوانین انجمنی الگوهای ناهنجار موجود در مجموعه داده‌ها شناسایی و بررسی می‌شوند. در ادامه با استفاده از ادغام نتایج به‌دست‌آمده از بخش‌های متفاوت فرایند، رکوردهای خطادار و ناهنجار به گونه‌ای مطلوب مشخص می‌شود. الگوریتم‌های بسیاری جهت کاوش قواعد انجمنی ارائه شده است. پس از بررسی این الگوریتم‌ها، برای اولین بار از الگوریتم Fp-growth که یکی از موفق‌ترین الگوریتم‌ها در این حوزه است [۳۶]، استفاده شده است. این الگوریتم در حال حاضر یکی از سریع‌ترین الگوریتم‌ها جهت کشف نمونه داده‌های تکراری می‌باشد [۳۰]. علاوه بر این Fp-growth بر اساس درختی پیشوندی از تراکنش‌های پایگاه داده نمایش داده می‌شود و باعث صرفه‌جویی قابل‌توجهی در حافظه ذخیره‌سازی تراکنش‌ها می‌شود [۳۶].

پس از پیاده‌سازی الگوریتم مورد استفاده بر روی مجموعه داده‌ها و بررسی ناهنجاری‌های موجود، نتایج حاصل ذخیره می‌شود.

۳-۷- تشخیص رکوردهای تکراری

نتایج به‌دست‌آمده از مراحل ترکیب نتایج و معرفی متغیر جدید و ایجاد قوانین انجمنی، در این قسمت در کنار یکدیگر جمع شده و مورد بررسی مجدد قرار می‌گیرد و دید یکپارچه‌ای نسبت به مجموعه داده‌ها، رکوردهای مشابه و تناقضات موجود فراهم می‌شود.

۳-۸- اطلاع به خبره

در این بخش نتایج حاصل شده از طریق فرایند، ازجمله رکوردهای مشابه، نیمه مشابه و خطادار به فرد خبره اعلام می‌شود تا تصمیماتی در جهت برطرف کردن این اشتباهات و نواقص گرفته شود.

۴- پیاده‌سازی فرایند

به‌منظور مشاهده کاربرد فرایند پیشنهادی، از محیط Matlab R2014b استفاده می‌شود و آزمایش‌ها بر روی سه مجموعه داده انجام می‌شود. فرایند پیشنهادی با [۲۲] مقایسه می‌شود.

۴-۱- مجموعه داده‌ها

پیاده‌سازی رویکرد پیشنهادی بر روی سه مجموعه داده شناخته‌شده و واقعی Restaurant، Employee و Adult انجام شده است.

برای ایجاد تکرار بر روی هر کدام از مجموعه داده‌ها، رکوردهای کاملاً تکراری و رکوردهایی نسبتاً تکراری، با اضافه کردن یک یا چند ویژگی تکراری در رکورد و اشتباهات تایپی در ویژگی‌ها ایجاد شده‌اند. در پنج، ده و پانزده درصد داده‌های هر یک از مجموعه داده‌ها خطا ایجاد شده است. مشخصات آماری مجموعه داده‌های مورد استفاده درجداول ۲- الف و ۲- ب نشان داده می‌شود.

جدول ۲-الف: ویژگی‌های آماری مجموعه داده‌ها

مجموعه داده	تعداد کل رکوردها
Restaurant	۸۶۴
Employee	۵۰۰۰
Adult	۴۸۸۴۲

جدول ۲-ب: ویژگی‌های آماری مجموعه داده‌ها

مجموعه داده	تعداد رکوردهای خطادار با نرخ خطای ۵ درصد	تعداد رکوردهای خطادار با نرخ خطای ۱۰ درصد	تعداد رکوردهای خطادار با نرخ خطای ۱۵ درصد
Restaurant	۴۳	۸۶	۱۱۲
Employee	۲۵۰	۵۰۰	۷۵۰
Adult	۲۴۴۲	۴۸۸۴	۷۳۲۶

هر کدام از مجموعه داده‌ها شامل ویژگی‌های متفاوتی هستند که در ادامه بیان می‌شود.

مجموعه داده Restaurant: مجموعه داده‌های دو Restaurant با ویژگی‌هایی از قبیل نام، آدرس و شهر است.

مجموعه داده Employee: مجموعه داده‌ای از کارمندان شامل ویژگی‌هایی از قبیل نام، نام خانوادگی، محل کار و میزان حقوق است.

مجموعه داده Adult: مجموعه داده‌ای با ویژگی‌هایی از قبیل سن، سطح شغلی، تحصیلات، وضعیت تاهل، نژاد، جنس، ملیت، نقش خانوادگی است.

۴-۲- معیارهای ارزیابی

برای ارزیابی مدل پیشنهادی از چهار معیار متعارف فراخوانی، دقت، صحت و امتیاز-اف استفاده می‌شود که فرمول آن‌ها در (۳)(۴)(۵) و (۶) آمده است. پارامترهای موجود در فرمول‌های استفاده شده در ادامه توضیح داده می‌شوند.

TP نشان‌دهنده تعداد رکوردهایی است که تکراری بوده‌اند و به‌عنوان موارد تکراری شناسایی می‌شوند.FP تعداد رکوردهایی است که تکراری نبوده‌اند اما به‌عنوان موارد تکراری شناسایی می‌شوند.TN تعداد رکوردهایی است که تکراری نبوده‌اند و به‌عنوان تکراری نیز شناسایی نمی‌شوند. FN تعداد رکوردهایی است که تکراری بوده‌اند اما به‌عنوان تکراری شناسایی نمی‌شوند.

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

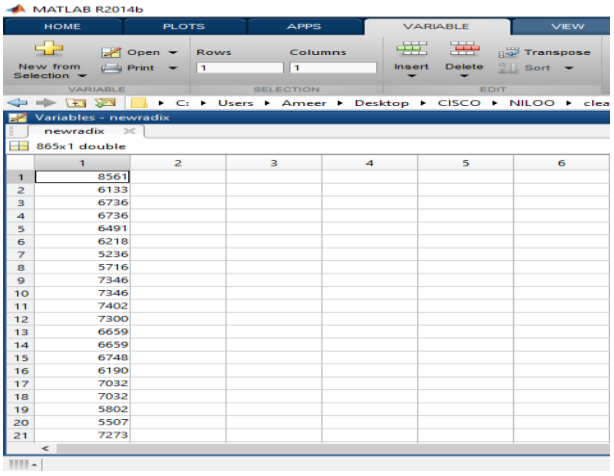
$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (5)$$

$$\text{F-score}=\frac{2*(Recall*Permidion)}{Recall+Permision} \quad (6)$$

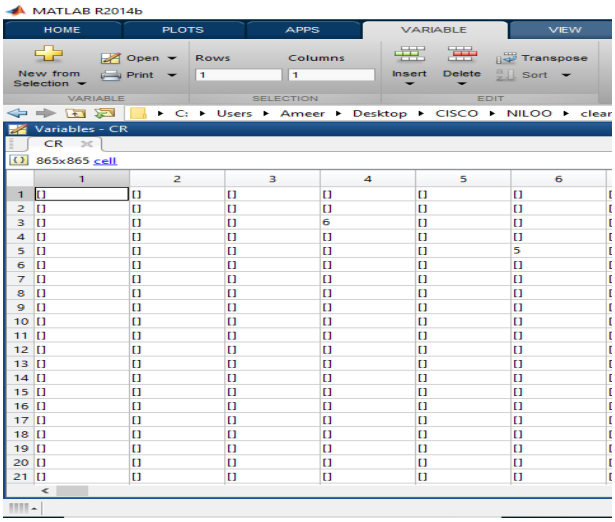
۴-۳- نتایج آزمایش‌ها

در این بخش، نمونه‌ای از انجام کار بر روی مجموعه داده Restaurant با ۵ درصد خطا آورده شده است. در ابتدا مراحل انتخاب ویژگی و یکسان‌سازی داده‌ها انجام می‌شود. در ادامه مرحله کدکردن داده‌ها است که نتایج حاصل از آن در نرم‌افزار متلب پوشه Newradix شکل (۲) آمده است که نشان‌دهنده عدد هر رکورد است.



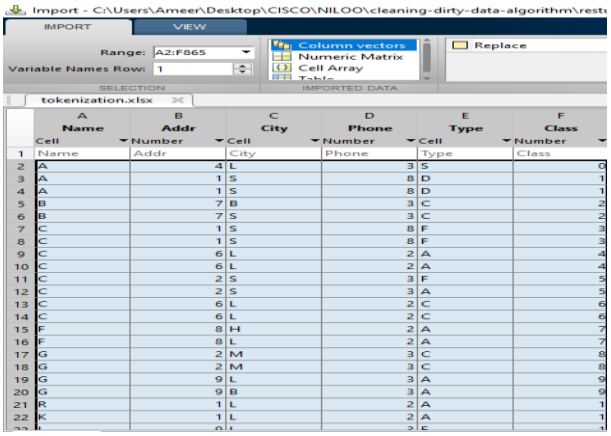
شکل ۲: رکوردهای کد شده

سپس خوشه‌بندی با الگوریتم Em انجام می‌شود که نتایج آن در پوشه Clusterdata ذخیره می‌شود. سپس رکوردهای موجود در هر خوشه مقایسه می‌شوند و رکوردهای مشابه و عدد تشابه به‌دست‌آمده از مرحله کدکردن داده‌ها در پوشه CR شکل (۳) ذخیره می‌شود.

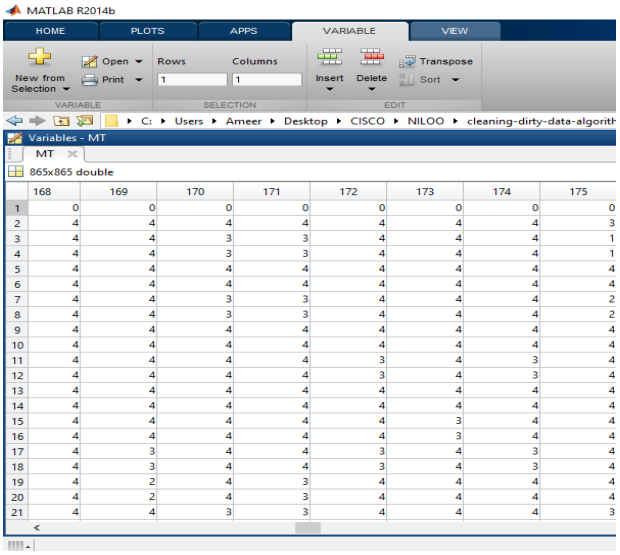


شکل ۳: جدول CR رکوردهای مشابه حاصل از مرحله کد کردن داده‌ها

نتایج حاصل از نشانه‌سازی در شکل (۴) فایل Tokenization آورده شده‌است. پوشه MT شکل (۵) نتایج را بعد از محاسبه فاصله لون‌اشتاین مشخص می‌کند و نشان‌دهنده‌ی رکوردهای مشابه و فاصله لون‌اشتاین میان رکوردها است.

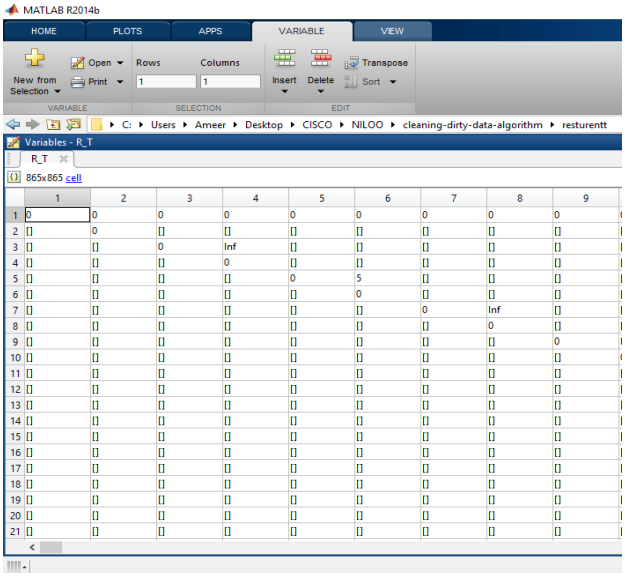


شکل ۴: نشانه‌های رکوردها



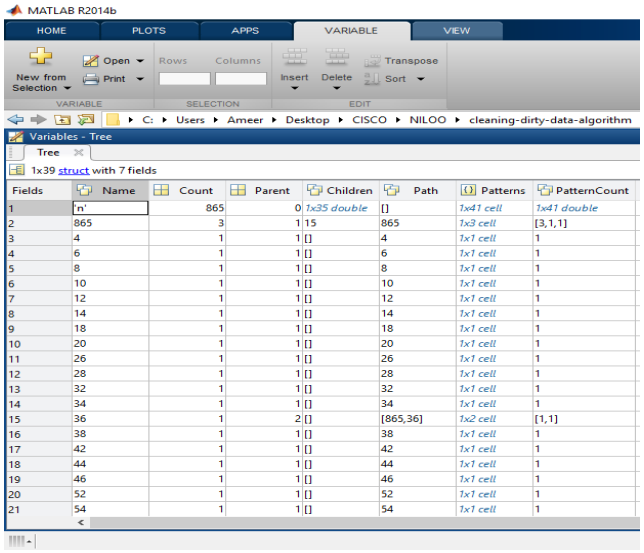
شکل ۵: جدول MT رکوردهای مشابه حاصل از مرحله نشانه‌سازی

سپس با توجه به شکل (۶)، جدول R-T بر اساس نتایج مراحل کد کردن داده‌ها و نشانه‌سازی ایجاد می‌شود.

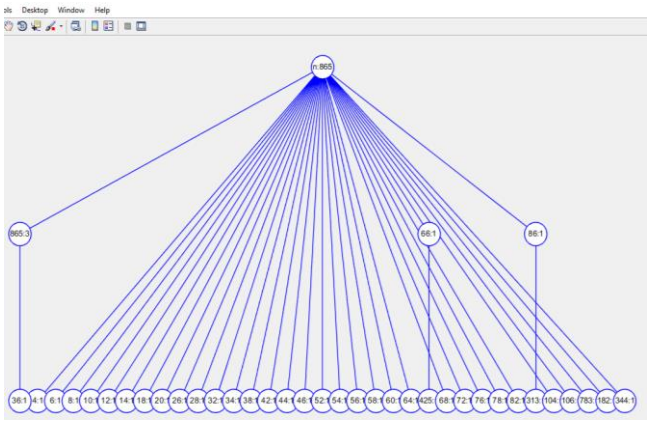


شکل ۶: جدول R-T

در ادامه با پیاده‌سازی الگوریتمFp-growth به الگوهای پرتکرار دست می‌یابیم و ناهنجاری‌های موجود در مجموعه داده‌ها شناسایی می‌شود شکل (۷ و ۸).



شکل ۷: الگوی تکرار داده‌ها



شکل ۸: درخت حاصل از پیاده‌سازی Fp-growth

نتایج حاصل از پیاده‌سازی شناسایی ۴۰ رکورد تکراری به‌درستی و با نام TP، سه رکورد غیرتکراری به‌عنوان تکراری با نام FP، سه رکورد تکراری به عنوان غیرتکراری با نام منفی FN و ۸۱۸ رکورد غیرتکراری به عنوان غیرتکراری با نام TN است. سپس معیارهای ارزیابی با توجه به فرمول‌های (۳)، (۴)، (۵) و (۶) محاسبه شده است. نتایج حاصل از محاسبه معیارهای مورد ارزیابی بر روی مجموعه داده‌های متفاوت در جدول ۳ نشان داده می‌شود.

جدول ۳: مقادیر معیارهای ارزیابی در مجموعه داده‌ها با نرخ خطای متفاوت

مجموعه داده	درصد خطا	فراخوانی	صحت	دقت
Restaurant	۵	۰.۹۳	۰.۹۸	۰.۹۳
Restaurant	۱۰	۰.۹۳	۰.۹۸	۰.۹۳
Restaurant	۱۵	۰.۹۳۷	۰.۹۸۴	۰.۹۴۵
Employee	۵	۰.۹۸	۰.۹۹۴	۰.۹۰
Employee	۱۰	۰.۹۹	۰.۹۹۴	۰.۹۵۷
Employee	۱۵	۰.۹۹	۰.۹۹۵	۰.۹۷
Adult	۵	۰.۹۸	۰.۹۹۷	۰.۹۶
Adult	۱۰	۰.۹۹	۰.۹۹۷	۰.۹۸
Adult	۱۵	۰.۹۹	۰.۹۹۷	۰.۹۹۲

با توجه به بررسی‌های انجام شده بر روی معیارهای ارزیابی به دست آمده، مشاهده می‌شود که در هر مجموعه داده با افزایش درصد خطای موجود، معیارهای ارزیابی موردنظر ازجمله صحت، دقت، فراخوانی و امتیاز-اف افزایش یافته است.

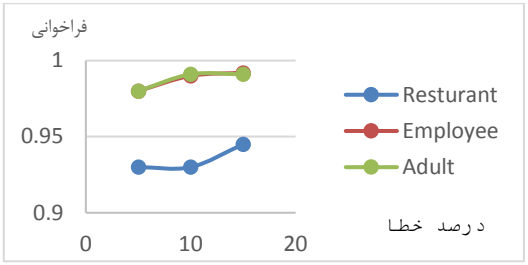
در ادامه به مقایسه مقادیر معیارهای به دست آمده بر روی سه مجموعه با یکدیگر پرداخته می‌شود که نمودارهای آن در شکل ۹ به صورت الف- ب- ج و دال نشان داده می‌شود.

شکل ۹- الف نشان دهنده‌ی نمودار ارتباط میان اندازه مجموعه دادها و معیار فراخوانی است و مشاهده می‌شود با افزایش حجم داده‌ها معیار فراخوانی افزایش یافته است.

شکل ۹- ب نشان دهنده‌ی نمودار ارتباط میان اندازه مجموعه داده‌ها و معیار صحت است و مشاهده می شود با افزایش حجم داده‌ها معیار دقت افزایش یافته است.

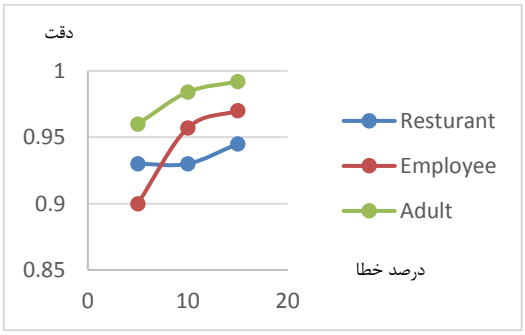
شکل ۹- ج نشان دهنده‌ی نمودار ارتباط میان اندازه مجموعه داده‌ها و معیار دقت است و مشاهده می‌شود با افزایش حجم داده-ها معیار صحت افزایش یافته است.

شکل ۹- دال نشان‌دهنده‌ی نمودار ارتباط میان اندازه مجموعه داده‌ها و معیار امتیاز- اف است و مشاهده می‌شود با افزایش حجم داده‌ها معیار امتیاز- اف افزایش یافته است.

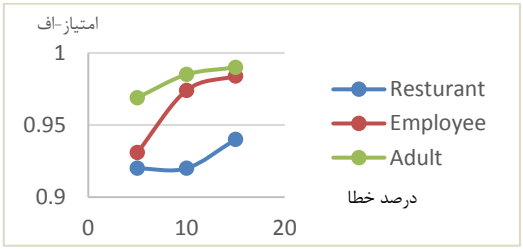


شکل ۹- الف: معیار فراخوانی در سه مجموعه داده

شکل ۹- ب: معیار دقت در سه مجموعه داده



شکل ۹- ج: معیار صحت در سه مجموعه داده



شکل ۹- د: معیار امتیاز-اف در سه مجموعه داده

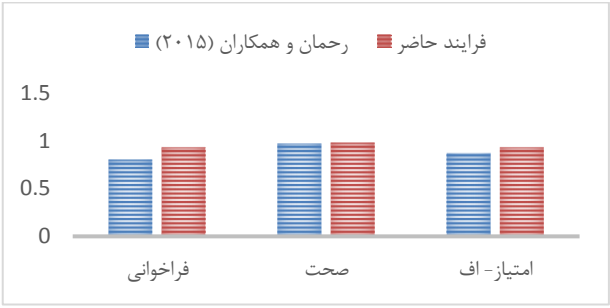
مشاهده می‌شود با افزایش حجم مجموعه داده‌ها مقادیر معیارهای ارزیابی مورد نظر نیز افزایش می‌یابد.

فرایند پیشنهادی با [۲۲] که از طریق رویکردی تجربی به تشخیص رکوردهای تکراری موجود در مجموعه داده‌ها پرداخته است، مقایسه می‌شود. مقایسه بر روی مجموعه داده Restaurant با نرخ خطای ۵٪ صورت می‌گیرد؛ که نتایج آن در جدول ۴ آمده است.

جدول ۴- مقایسه فرایند پیشنهادی و [۲۲]

مقاله	فراخوانی	صحت	امتیاز- اف
رحمان و همکاران (۲۰۱۵)	۰.۸۰	۰.۹۷	۰.۸۷
فرایند حاضر	۰.۹۳	۰.۹۸	۰.۹۳

نمودار ۱۰ نشان دهنده‌ی مقایسه میان دو فرایند است؛ با برر سی نتایج به‌دست آ مده م شاهده می شود که معیار های ارز یابی در فرایند پیشنهادی با مقادیر ۱۳٪ فراخوانی، ۱٪ صحت و ۶٪ امتیاز- اف افزایش یافته‌اند.



نمودار ۱۰: مقایسه فرایند پیشنهادی با فرایند رحمان و همکاران (۲۰۱۵)

۵- نتیجه‌گیری و پیشنهادهای کارهای آینده

در این مقاله فرایندی جدید برای تشخیص تکرار در مجموعه داده‌ها پیشنهاد شده است که هدف اصلی آن پاک‌سازی دقیق و افزایش کیفیت داده‌ها است. این فرایند ویژگی‌هایی از قبیل تعامل با کاربر، درک آسان و عملکرد قابل قبول را در برمی‌گیرد.

فرایند پیشنهادی مبتنی بر کد کردن رکوردها و خوشه‌بندی بر اساس الگوریتم امید ریاضی- بیشینه‌سازی، ساخت نشانه برای رکوردها، ادغام روش‌های کد کردن داده‌ها و نشانه سازی و ایجاد قوانین انجمنی با الگوریتم Fp-growth است. با پیاده‌سازی فرایند پیشنهادی بر روی مجموعه داده‌های متفاوت و محاسبه معیارهای ارزیابی، افزایش معیارهای ارزیابی مورد نظر با افزایش اندازه مجموعه داده‌ها مشاهده شد. همچنین مشاهده شد که معیارهای اندازه‌گیری با افزایش درصد خطای موجود در مجموعه داده‌ها افزایش یافته‌اند. فرایند پیشنهادی می‌تواند برای بهبود کیفیت داده‌ها مورداستفاده قرار گیرد. از ضعف‌های فرایند پیشنهادی، زمان اجرای آن در مواجهه با مجموعه داده‌های حجیم است که می‌توان با مطالعات بیشتر در جهت برطرف کردن آن گام برداشت.

مراجع

[1] K. Ali and M. Warraich," A framework to implement Data Cleaning in Enterprise Data Warehouse for Robust Data Quality," in Information and Emerging Technologies ., Karachi., ICIET.2010.

[2] D.Luebbers, U. Grimmer, and M.Jarke," Systematic Development of Data Mining-Based Data Quality Tools," in 29th international conference on Very large data bases., Berlin., VLDB,pp.548-559,2003.

[3] J.Han and M.Kamber,"Data Mining: Concepts and Techniques," 3th ed. Morgan kaufman publisher is an imprint of Elsevier,pp.1-673,2006.

[4] K.Sarpong, K. Adu-Manu, and J. Kingsley Arthur,"A Review of Data Cleansing Concepts – Achievable Goals and Limitations," International Journal of Computer Science and Information Technologies.,vol.3,pp. ۵۲۱۴ - ۵۲۱۲,2013.

[5] B.Pinar," A Comparison of Data Warehouse Design Models," M.S. thesis, Dept. Computer.Eng., Atilim University.,2005.

[6] E. Rahm and H. Hai Do," Data Cleaning: Problems and Current Approaches," Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,2000.

[7] J.Broeck, S.A. Cunningham, R.Eeckels and K.Herbst,"Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities,"vol. 2, pp. 966-970,2005.

[8] L.Ettinger,"Improving the DataWarehouse with Selected Data Quality Techniques: Metadata Management, Data Cleansing and Information Stewardship," University of Oregon Applied Information Management Program,2008.

[9] H.Muller and J.Christoph Freytag," Problems, Methods, and Challenges in Comprehensive Data Cleansing," Humboldt-Universität zu Berlin, Berlin,2005.

[10] R.Arora, P.Pahwa, and S.Bansal," Alliance Rules for Data Warehouse Cleansing",in International Conference on Signal Processing Systems., PP743-747,2009.

[11] M.Hamad and A.Jihad," An Enhanced Technique to Clean Data in the Data Warehouse," E-systems Engineering, pp.301-311,2011.

[12] Y.Hao,D.Xing-chun, and L.Kai-qi,"Research on Information Quality Driven Data Cleaning Framework,"in International Seminar on Future Information Technology and Management Engineering., PP537-539,2008.

[13] M.Rehman and V.Esichaikul,"Duplicate Record Detection For Database Cleansing.in Machine Vision,"in Second International Conference., Dubai, pp. 333 – 338,2009.

[14] R.Kavitha Kumar and RM.CHADRASEKARAN,"Attribute Correction-Data Cleaning Using Association Rule and Clustering Methods," International Journal of Data Mining & Knowledge Management Process.,vol.1,pp.22-32,2011.

[15] O.Abbas ," Comparisons between data clustering algorithm", Computer science Department, Yarmouk University, Jordan, pp320-325,2007.

[16] C.Mayfield, J.Neville, and S. Prabhakar,"A Statistical Method for Integrated Data Cleaning and Imputation", Purdue University,Computer Science Technical Reports,2009.

[17] A. Hoshino,H.Nakayama,K. Kanno, and K.Nishimura ,"Leveraging the Common Cause of Errors for Constraint-Based Data Cleansing," Springer International Publishing Switzerland,pp 164-176,2015.

[18] J.Gu," Random Forest Based Imbalanced Data Cleaning and Classification," 2007.

[19] M.Bergman, T.Milo, S.Novgorodov and W.Tan,"QOCO: A Query Oriented Data Cleaning System with Oracles,"in International Conference on Very Large Data Bases., Hawaii., VLDB.pp.1900-1903,2015.

[20] M.Bergman, T.Milo, S.Novgorodov and W.Tan,"Query-oriented data cleaning with oracles," in 41 st International Conference on Management of Data., Melbourne., pp. 1199-1214,2015.

[21] V.Raman and M.Hellerstein," Potter'sWheel: An Interactive Data Cleaning System,"in 27th Conference on Very Large Data Bases., Roma., VLDB.2001.

[22] M.Rehman andV. Esichaikul ,"Duplicate Record Detection For Database Cleansing".in Machine Vision, Second International Conference on , Dubai, pp. 333 – 338,2009.

[23] K.Ripon and A.Rahman,"A Domain-Independent Data Cleaning Algorithm for Detecting Similar-Duplicates," Journal of Computers,vol.5,pp.1800-180,2012.

[24] J.Tamilselvi and V.Saravanan,"A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse", IJCSNS International Journal of Computer Science and Network Security, PP117-121,2008.

[25] E.Ohanekwu and C.I. Ezeife," A Token-Based Data Cleaning Technique for DataWarehouse Systems," Natural Science and Engineering Research Council (NSERC).,2013.

[26] R.KavithaKumar and RM.Chadrasekaran, "Attribute Correction-Data Cleaning Using Association Rule and Clustering Methods,"International Journal of Data Mining & Knowledge Management Process (IJDKP),vol.1,2011.

[27] H.Yu, Z.Xiao-yi, and Y.Zhen, "A Universal Data Cleaning Framework Based on User Model," ISECS International Colloquium on Computing, Communication, Control, and Management,2009.

[28] W.Wei and M.Zhang, and B, Zhang and X.Tang,"A Data Cleaning Method Based on Association Rules,"in24 th International Conference on Intelligent System and Khnowledge Engineering., 2007.

[30] J.Han and H. Pei H, and Y. Yin," Mining Frequent Patterns without Candidate Generation," in ACM SIGMOD international conference on Management of data, pp.1-12,2000.

[31] A.Paul, V.Ganesan, JS. Challa and Y.Sharma,"HADCLEAN: A Hybrid Approach to Data Cleaning in DataWarehouses,"IEEE.pp.136-142,2012.

[32] B.Khan, A.Rauf, H.Javed and S.Khusro,"Removing Fully and Partially Duplicated Records throughK-Means Clustering," IACSIT International Journal of Engineering and Technology,PP.750-754,2012.

[33] I.Kononenko and S. Hong," Attribute Selection for Modeling", Future Generation Computer Systems, pp.181 – 195,1997.

[34] J.Wang,"Data warehousing and mining: concept, methodologies, tools, and appplications," in Information ScienceRefrence,2008.

[35] G.Venkatesh and A. Sarma," Data Cleaning A Practical Perspective", Morgan & Claypool, pp.1-85,2013.

[36] C. Borgelt, "An Implementation of the FP-growth Algorithm," in 1st international workshop on open source data, pp.1-5,2005.

پاورقی‌ها:

[1] Expectation-maximization(Em)

[2] http://www.cs.utexas.edu/users/ml/riddle

[3] http://data.cityofchicago.org/Administration-Finance /Currant-Employee-Names-Salaries-and-Position-Title/xzkq-xp2w

[4] https://archive.ics.uci.edu/ml/datasets/adult

[5] Recall

[6] Precision

[7]Accuracy

[8] F-score