چکیده

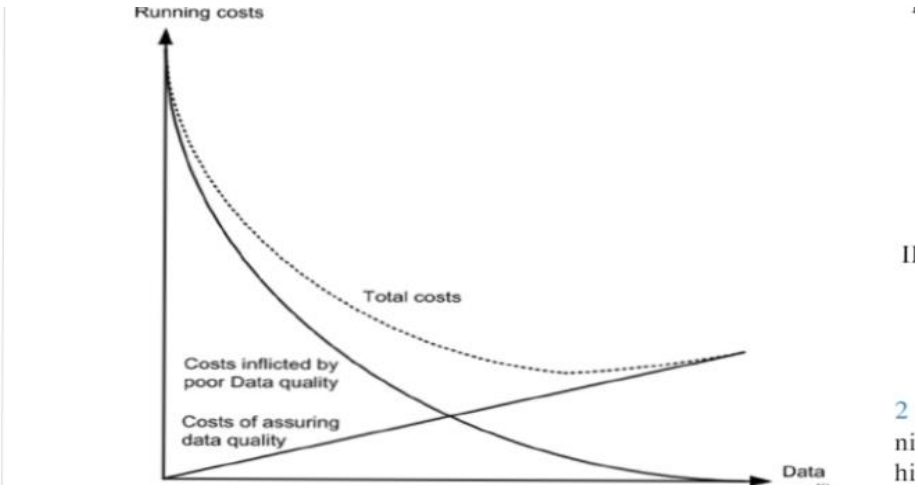
بیشتر بانکهای اطلاعاتی موجود از ناهماهنگی و ناسازگاری داده ها رنج می برند. تلاش برای افزایش کیفیت داده ها برای حل این مسئله ضروری است. در این مقاله ، دو روش برای استخراج دقیق قوانین توابع وابستگی شرطی، از چنین پایگاه داده هایی ، برای حل مشکل تمیز کردن داده ها، مطرح شده است. ایده تکنیکهای پیشنهادی این است که ابتدا الگوهای حداکثر (ماکزیمال) بسته بسته شده را استخراج کنند؛ سپس با کمک اندازه گیری پیش رفت ، مقررات وابسته به عملکردهای شرطی استخراج کنند. علاوه بر این ما یک الگوریتم تعمیر داده ها را برا رفع تاپل های متناقض موجود در بانک اطلاعاتی با بهره گیری از قوانین تولید شده پیشنهاد می کنیم. ما یک مطالعه تجربی گسترده هم برای تایید اثر بخشی تکنیکهای پیشنهادی در مقایسه با تکنیکهای موجود، در مجموعه داده های پزشکی واقعی و زندگی انجام داده ایم.

1. معرفی

امروزه سازمان ها ، با چالش های فزاینده ای در رشد داده های خود روبرو هستند. در حال حاضر تضمین کیفیت داده، در اکثر حوزه های کاربردی ، یک چالش اساسی است. سازمانها ی دولتی و خصوصی زیادی هستند که جنبه قابل توجهی از داده هایشان وجود دارد که برای اهداف مدیریت داده استفاده می شود. از این رو، وجود داده های متناقض ، ارزیابی آنها را کاهش میدهد (از بین می برد) . و باعث می شود آنها نادرست یا حتی مضر باشند. ارزش داده ها وابسته به کیفیت آن است که تعیین کردن آن دشوار است. داده ها برخلاف ، محصولات تولیدی ، ویژگی های فیزیکی ندارند که باعث می شود کیفیت آنها به آسانی ارزیابی شود . بنابراین کیفیت تابعی از خصوصیات نامشهود است. در این متن کیفیت به این معنی است که داده " مناسب برای استفاده " یا " پتانسیل استفاده" هستند.

"مناسب برای استفاده" نه تنها مفاهیم کیفی استاتیکی واریانس و بیاس (تعصب) را شامل می شود ، بلکه ویژگی های دیگری مانند قوام (استحکام) و تکثیر را نیز تعیین می کند که چگونگی استفاده از اطلاعات آماری را به طور موثر مشخص می کند. کیفیت داده ها یک ویژگی اساسی است که اعتبار اطلاعات برای تصمیم گیری در سازمان ها را مشخص می کند. کیفیت داده ها با گذشت زمان به سرعت در پایگاه داده های دنیای واقعی تخریب می شوند و بر نتایج معدن تاثیر می گذارند.

داده ای کثیف سالانه میلیاردها دلار برای تجارت آمریکا هزینه می برند، که منجر به تصمیم گیری های ضعیف می شود. شکل 1 ارتباط بین هزینه هایی که برای داده هایی با کیفیت ضعیف خرج می شود و هزینه هایی که برای داده هایی با کیفیت بالا خرج می شود را نشان می دهد. یک تجارت بین هزینه هایی که توسط کیفیت ضعیف داده ها و هزینه هایی که توسط کیفیت بالای داده ها است ، وجود دارد. افزایش داده های با کیفیت ضعیف در نتیجه گیری از داده های با کیفیت بالا تاثیر می گذارد که منجر به تفکر در مورد چگونگی اطمینان و دستیابی به کیفیت داده ها می شود.



بنابراین ، تمرکز بر تشخیص و اصلاح مشکلات ناسازگاری داده ها به عنوان فرایند تمیز کردن داده ها مهم است . به طور خاص ، تضمین داده های قابل اعتماد با کیفیت بالا یک مزیت رقابتی برای همه ی صنایع است؛ که به راه حل های دقیق اسکرابینگ داده نیاز دارد.

اصطلاح تمیز کردن داده ها (scrubing داده ها) مرحله مهم پیش پردازش داده ها محسوب می شود. این واقعا برای حفظ سوابق متناقض ، قبل از استخراج و بررسی داده ها ، کار می کند. تمیز کردن داده ها اساسا در چندین برنامه کاربردی مانند انبار کردن داده ها ، مدیریت کیفیت داده ها و کشف دانش در پایگاه های داده وجود دارد. داده کاوی یک حوزه تحقیق است که به معنای فرایند ایجاد الگوهای پنهان و ناشناخته در پایگاه داده ها است و از این واقعیت برای ساخت مدل استفاده می کند. داده کاوی روش ها و فناوریهایی را برای تبدیل مقادیر زیادی از داده ها برای پردازش و تجزیه و تحلیل به اطلاعات مفید ، برای اهداف تصمیم گیری فراهم می کند.

داده کاوی در الگوریتم های فعلی تمیز کردن داده ها، موضوعی محبوب تر و ضروری تر

می شود. زیرا پاک کردن داده های دستی نیز فرایندی طاقت فرسا و وقت گیر است و خود به تنهایی مستعد خطا است.

بدون شک، اکثر تکنیکهای تمیز کردن داده ها در نوشته ها از تطابق رکورد (مدارک) استفاده می کنند ، که احتمالا سوابق اشتباه را با سوابق درست (تمیز شده) اصلی مقایسه می کند. مقابله با مسئله ناسازگاری داده ها با بکار بردن داده های خود بدون نیاز به داده های اصلی اجباری است. درحین تثبیت ناسازگاری داده ها از جمله وابستگی های عملکردی و وابستگی های عملکردی شرطی را باید بجا آورد.

شناسایی مقادیر متناقض (ناسازگار) یک گام اساسی در فرایند تمیز کردن داده ها است.

در حال حاضر، افزایش داده ها در بیشتر حوزه های برنامه (اپلیکیشن) یک چالش مهم برای اطمینان از ثبات و صحت کیفیت داده ها است که برای اهداف مدیریت داده ها است که برای اهداف مدیریت داده استفاده می شود. داده های نادرست ضبط شده بصورت الکترونیکی ، منجر به داده های بی کیفیت می شود.

در اینجا، ما علاقه مند به ایجاد قوانین قابل اعتماد و دقیق برای کیفیت داده ها هستیم که برای بعدا از آن برای اصلاح ناسازگاری داده ها در چندین دامنه برنامه استفاده می شود.

مشارکتها:

مشارکت های اصلی این مقاله شامل ارائه تکنیک هایی برای تقویت فرایند تمیز کردن داده ها است. به طور خاص مشارکت های کاغذی شامل موارد زیر هستند:

پیشنهاد تکنیک ICCFD MINER: برای ایجاد قوانین تمیز کردن داده های قابل اعتماد و عمدتا بر اساس الگو های بسته شده مکرر .

پیشنهاد تکنیک MICCFD :برای افزایش عملکرد اولین تکنیکهای پیشنهادی MINER ICCFD در تولید قوانین نظافت داده ها . این پیش رفت در استخراج الگو های حداکثر مکرر و ژنراتور های مرتبط با آنها بجای استفاده از الگوهای مکرر بسته بدست می آید. این بعنوان مکانیسم استراتژی جست و جوی موثر برای کاهش اندازه دامنه فضای جست و جو عمل می کند.

الگوریتم T-REPAIR : این الگوریتم برای قوانین ایجاد شده پیشنهاد شده است و مقادیر ورودی t رابطه برای اعتبار سنجی خطاهای تعمیر از مجموعه داده شده است.

سازمان:

این مقاله به شرح زیر برگزار می شود:

بخش 2 در مورد کارهای مرتبط بحث می کند.

در بخش 3 پیش زمینه و تعریف مفاهیم کیفیت داده ارائه می شود.

سپس در بخش 4 تمیز کردن داده ها را در کاربردهای پزشکی برجسته می کند.

بخش 5 تکنیک های پیشنهادی ICCFD-MINER و MICCFD-MINERو

T-REAPAIR را ارائه می دهد.

بخش 6 به مطالعه تجربی و نتایج انجام شده برای مجموعه های مختلف پزشکی می پردازد.

سرانجام بخش 7 کار پیشنهادی را به پایان میرساند و روند های آینده را بررسی می کند.

2. کار مرتبط:

متاسفانه علی رغم نیاز مبرم به تکنیک های دقیق و قابل اعتماد برای افزایش کیفیت داده ها وتمیز کردن داده ها ، هنوز هیچ راه حل اساسی برای رفع این مشکلات ارائه نشده است. بحث و تحلیل کمی در مورد افزایش ثبات داده ها صورت گرفته است. با این حال ، بسیاری از کارهای اخیر بر تطابق رکورد و تشخیص تکراری بودن داده ها متمرکز شده است.محققان پایگاه داده و کیفیت داده ها انواع محدودیت های یکپارچگی را بر اساس وابستگی های عملکردی(FD) مورد بحث قرار دارند. در مرجع نویسندگان الگوریتم FD-MINE را پیشنهاد میکنند که وابستگی عملکردی را از رابطه داده شده کشف می کنند. یک بررسی و

مقایسه جامع از هفت الگوریتم برای کشف وابستگی های عملکردی در مرجع مطرح و بحث شده است.

الگوریتم های مورد بررسی شامل

TANE , FUN , FD-MINE , DFD , DEP-MINER, FASTFDS , FDEP

هستند که به طور گسترده در مرجع نشان داده شده است. با این وجود FD های سنتی عمدتا برای طراحی طرحواره توسعه می یابند اما اغلب قادر به تشخیص خطاهای ارزش معنایی داده ها نیستند. محققان دیگر بر گسترش FD تمرکز می کنند. آنها آنچه را که به اصطلاح وابستگی های عملکردی شرطی و وابستگی گنجاندن شرطی گفته می شود را برای ضبط کردن خطاها در داده ها ارائه داده اند.

الگوریتم هایی که برای کشف قوانین CFD از رابطه پیشنهاد شده اند عبارتند از:

الگوریتم MINER CFD : برای کشف وابستگی های عملکری شرطی ثابت .

2 . الگوریتم CTANE : که (ANE) را برای کشف (CFD)های عمومی گسترش میدهد و FCFD را برای کشف CFD ها عمومی با استفاده از یک استراتژی جست و جو ی عمیق به جای روش سطح بندی همانطورکه در الگوریتم (CTANE) استفاده می شود.

چندین تکنیک کیفیت داده برای تمیز کردن تاپلس های کثیف از پایگاه داده ها ارائه شده است. زیرا محققان هدف خود را برای دستیابی به اطلاعات مهم منتسب به پایگاه داده ها می دانند. در مرجع نویسندگان سه مدل را برای مشخص کردن کامل بودن اطلاعات نسبی در مورد داده های پیشنهادی که از طریق آنها مقادیر ممکن است از بین بروند، پیشنهاد می کنند. روش های استنتاج آماری در منابعی که شامل اطلاعات گمشده

و خطاهای صحیح به طور خودکار است، بررسی می شود. این رویکردها مقادیر گمشده را برای بهبود کیفیت داده ها نگه می دارد.

از بخش فناوری چندین ابزار منبع باز برای پردازش داده های کثیف تولید شده است.

OPEN RENE , DATA WRANGLER دو ابزار منبع باز برای کار با داده های از دست رفته برای تمیز کردن آنها است که توضیح مفصل آنها در منبع آمده است.

علاوه براین انواع روش های تبدیل داده ها مانند ابزارهای تجاری (ATL) (استخراج، تحول، بارگیری) وجود دارد. روش های استخراج بر روی استخراج داده ها بر روی منابع همگن یا ناهمگن تمرکز دارد. هدف از روش های تبدیل و دگرگونی ، ذخیره داده ها در قالب یا ساختار مناسب برای تحقیق و تحلیل است. روش بارگیری مربوط به بارگذاری داده ها به یک مخزن منبع داده، از جمله انبار داده ها یا سایر منبع داده های جداگانه وابسته به نیاز سازمان است. این ابزارها برای تمیز کردن داده ها ، برای پشتیبانی از هرگونه تغییر در ساختار، بازنمایی یا محتوای داده ها تهیه شده اند . استفاده از قوانین ویرایش به همراه داده های کارشناسی ارشد منبع شرح داده شده است. چنین قوانینی می توانند با بروز رسانی کپسول های ورودی با داده های اصلی برخی از موارد خاص را بدست می آورند. با توجه به محدودیت ها قوانین ویرایش دارای شناسایی پویا بوده و با داده های کارشناسی ارشد ارتباط دارد. با توجه به یک ورودی تاپلس که با یک الگو مطابقت دارد ، قوانین ویرایش به ما می گویند که کدام ویژگی ها تاپلس باید بروز شوند و چه مقادیری از داده های استاد (اصلی) باید به آنها اختصاص داده شود. این رویکرد مستلزم تعیین قوانین ویرایش به صورت دستی برای هردو رابطه است یعنی رابطه اصلی و رابطه ورودی که بسیار گران و پر هزینه است.

ترمیم از راه حل اکتشافی مبتنی بر تابع جداول هزینه ی دو بروز رسانی استفاده می کند که همیشه ترمیم و تعمیر قطعی ندارد. ویرایش قوانین

،کاربران را ، ملزم به بررسی هر نوع حلقه ای که پرهزینه است می کند. علاه بر این کار های زیادی در ادبیات با تکیه بر شباهت خاص دامنه و اپراتور های تطبیق ارائه شده است. این آثار شامل: تطابق رکورد ، پیوند رکورد ، تشخیص کپی و ادغام است . این رویکردها دو کارکرد را بنام مطابقت و ادغام تعریف می کنند. در حالیکه عملکرد تطابق کپی رکوردها را شناسایی می کند؛ عملکرد ادغام دو رکورد تکراری را به یکی ترکیب می کند سرانجام نتیجه گرفته می شود که روش های موجود تاثیرات قطعی و قایل اعتماد برای مسئله نا هماهنگی را تعیین نمی کند . چنین روش هایی هنگام تشخیص خطاها در داده های حساس مانند پرونده الکترونیکی پزشکی (EMR) سیستم های بهداشت و درمان به درستی کار نمی کنند.

در این مقاله (CFD ثابت) به عنوان یک مورد خاص از قوانین انجمن مجددا مورد استفاده قرار میگیرد و برای حل مشکل تشخیص و تعمیر خطاهای ناسازگاری در پایگاه داده استفاده می شود.

3 .مقدماتی

چندین اصطلاح از ادبیات وجود دارد که مربوط به تکنیکهای پیشنهادی است. تعریف این اصطلاحات شامل وابستگی های عملکردی – شرطی ، الگوهای مکرر بسته و مکرر maximal ، مشکل CFD ثابت و فضای جست و جوی هرس ارائه شده است.

وابستگی های عملکردی (FD) بعنوان محدودیت بین دو مجموعه از ویژگی ها در رابطه با یک پایگاه داده تعریف می شود.

FD روی رابطه R بعنوان $X \rightarrow Y$ که در آن هر مقدار X دقیقا برابر با یک مقدار Y است.

در اینجا X بعنوان تعیین کننده در نظر گرفته می شود و Y بعنوان وابسته در نظر گرفته می شود.

FD رابطه بین تمام ترکیبات ممکن از جفت های ارزش ویژگی را توصیف می کند و برای کار برروی طراحی طرحواره که معنایی از داده ها را برای تمییز کردن داده ها ضبط نمی کند، توصیف نمی کند.

وابستگی های عملکردی شرطی (cfd) توسعه یافته DF است که هدف آن تشخیص ناسازگاری داده ها بین تاپل ها در یک رابطه واحد است .

CFD Q در رابطه R ، یک جفت است (X→Y ,TP) که در آن X→Y یک جفت در R است و TP الگوی TUPPLE از Q با خواص X و Y است. برای هر ویژگی A در TP[A] ، X U Y یا ثابت در دامنه A است یک متغییر نامشخص '_' .

در این مثال، رکورد جدول روابط با مشتری مورد بررسی قرار می گیرد. اینگونه سوابق دارای خصوصیات زیر است:

کد کشور(CC) ، کد منطقه(AC) ، شماره تلفن (PN) ، نام (NM) ، خیابان (STR) ، شهر (CT) و کد پستی (ZIP) همانطور که در جدول 1 نشان داده شده است.

هدف از این مثال توضیح تفاوت بین FD و CFD است. FDهای سنتی که در جدول 1 وجود دارد بشرح زیر است:

F1 : [CC.AC] → CT

F2 : [CC.AC.PN] → STR

F1 بیان می کند که اگر دو مشتری کد کشوری و کد منطقه یکسانی داشته باشند آنها یک شهر مشترک هم دارند بطور مشابه برای F2 .

موارد زیر CFD است که در جدول شماره 1 وجود دارد.

Q0 : ([CC.ZIP]→STR.(32.-||-))

Q1 : ([CC.AC]→CT.(40.872 .-||-UN))

Q2 : ([CC.AC]→CT.(32.222.-||-VIZAG))

Q3 : ([CC.AC]→CT.(40.101.-||-EDI))

قوانین بعنوان متغییر CFD تا Q0 طبقه بندی می شوند جایی که Q1 و Q2 ، CFD ثابت هستند در این کار ما نگران چنین CFD ثابت هستیم. Q0 CFD ادعا می کند کد ZIP به طور منحصر بفرد STR را تعیین ی کند. این FD بیشتر از همه رابطه با الگوی CC=32 را در زیر مجموعه ای از تاپل ها نگهداری می کند.

Q1 CFD اطمینان می دهد در صورتی که اگر مشتری دارای کد کشوری CODE=40 و کد منطقه CDOE=872 باشد شهر مشتری در سازمان ملل متحداست (UN) بطور مشابه برای Q2 و Q3 داریم.

Q0,Q1,Q2,Q3 CFD نشان می دهد که این قوانین نمی توانند توسط FDها نمایان شوند .

الگوهای متداول بسته ، اگر در یک سوپرست مناسب با همان پشتیبانی کنجانده نشود، الگوی متناوب بسته می شود.

ژنراتور Y با یک الگوی غالبا بسته X ، یک محدودیت الگویی با داشتن همان پشتیبانی از X است و هیچ زیر مجموعه ای با پشتیبانی یکسان ندارد. مجموعه الگوهای مکرر بسته، بی ضرر است که اطلاعات کاملی را در مورد الگوهای مکرر مربوط به آن ارایه می دهد.

از مجموعه الگوهای مکرر بسته ، به آسانی میتوتنیم هویت و پشتیبانی از همه الگو های مکرر را بدون استخراج مجدد بانک اطلاعاتی استخراج کنیم. در عین حال در یک زمان یکسان ، الگوهای مکرر بسته ، به خودی خود می توانند سفارش ها با اندازه کوچکتر از اندازه الگو های مکرر قبول کنند؛ بخصوص در پایگاه داده متراکم.

برای مثال از جدول ([CC,AC,CT,ZIP] , ((40,827,UN,8422)الگوی بسته ای با پشتیبانی 3 است. این الگوی بسته دارای دو الگوی ژنراتور است ((40,827) , [CC,AC]) و ((08422) , [ZIP]) است که هر دو دارای پشتیبانی برابر با 3 هستند.

الگو های فرکانس حداکثر، MAXIMAL نامیده می شوند. زیرا هیچ گونه سوپرست مکرر ندارند.

به عبارت دیگر الگو های مکرر maximal هستند اگر هیچ یک ار سوپر ست های آنها مکرر نباشد. الگو مکرر حداکثر استفاده از معادن به کشف الگو های طولانی در بانکهای اطلاعاتی متراکم کمک می کند. استخراج الگو های ضربهای حداکثر به یک مسئله مهم تبدیل شده است. زیرا مجموعه الگوهای مکرر حداکثر نه تنها الگو های منحصر بفرد ، بلکه تعداد الگو های حداکثر مکرر نیز می تواند به طور قابل توجهی کمتر از تعداد بسته های مکرر بسته شود. مجموعه الگو های حداکثر مکرر بدین ترتیب زیر مجموعه ای از الگوهای مکرر بسته است که زیر مجموعه ای از همه الگو های متداول است. این الگو ها با استفاده از مکانیسم جست و جوی موثر برای کاهش اندازه دامنه ی فضای جست و جو استخراج می شوند.

علاوه بر این ، مجموعه این الگو ها یک مجموعه حداقل است ،یعنی کوچکترین مجموعه ای که از آن همه الگوهای مکرر استفاده می شود.

چکیده

بیشتر بانکهای اطلاعاتی موجود از ناهماهنگی و ناسازگاری داده ها رنج می برند. تلاش برای افزایش کیفیت داده ها برای حل این مسئله ضروری است. در این مقاله ، دو روش برای استخراج دقیق قوانین توابع وابستگی شرطی، از چنین پایگاه داده هایی ، برای حل مشکل تمیز کردن داده ها، مطرح شده است. ایده تکنیکهای پیشنهادی این است که ابتدا الگوهای حداکثر (ماکزیمال) بسته بسته شده را استخراج کنند؛ سپس با کمک اندازه گیری پیش رفت ، مقررات وابسته به عملکردهای شرطی استخراج کنند. علاوه بر این ما یک الگوریتم تعمیر داده ها را برا رفع تاپل های متناقض موجود در بانک اطلاعاتی با بهره گیری از قوانین تولید شده پیشنهاد می کنیم. ما یک مطالعه تجربی گسترده هم برای تایید اثر

بخشی تکنیکهای پیشنهادی در مقایسه با تکنیکهای موجود، در مجموعه داده های پزشکی واقعی و زندگی انجام داده ایم.

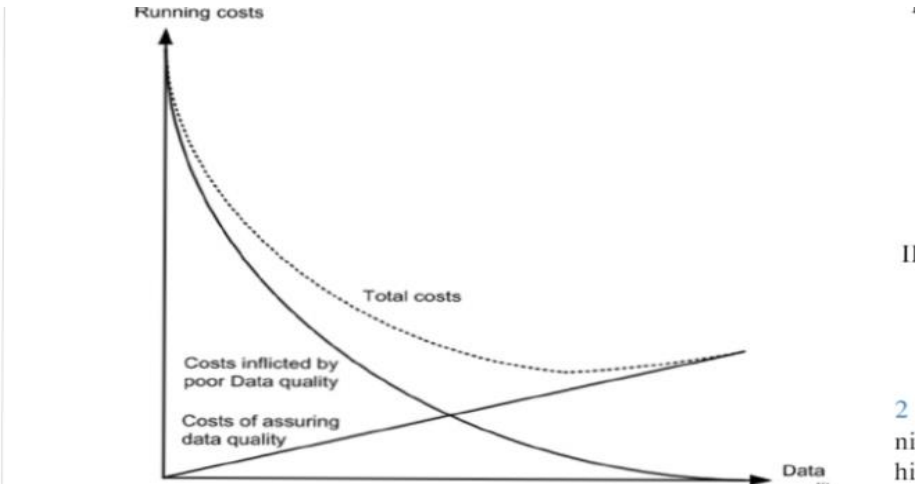
1. معرفی

امروزه سازمان ها ، با چالش های فزاینده ای در رشد داده های خود روبرو هستند. در حال حاضر تضمین کیفیت داده، در اکثر حوزه های کاربردی ، یک چالش اساسی است. سازمانها ی دولتی و خصوصی زیادی هستند که جنبه قابل توجهی از داده هایشان وجود دارد که برای اهداف مدیریت داده استفاده می شود. از این رو، وجود داده های متناقض ، ارزیابی آنها را کاهش میدهد (از بین می برد) . و باعث می شود آنها نادرست یا حتی مضر باشند. ارزش داده ها وابسته به کیفیت آن است که تعیین کردن آن دشوار است. داده ها برخلاف ، محصولات تولیدی ، ویژگی های فیزیکی ندارند که باعث می شود کیفیت آنها به آسانی ارزیابی شود . بنابراین کیفیت تابعی از خصوصیات نامشهود است. در این متن کیفیت به این معنی است که داده " مناسب برای استفاده " یا " پتانسیل استفاده" هستند.

"مناسب برای استفاده" نه تنها مفاهیم کیفی استاتیکی واریانس و بیاس (تعصب) را شامل می شود ، بلکه ویژگی های دیگری مانند قوام (استحکام) و تکثیر را نیز تعیین می کند که چگونگی استفاده از اطلاعات آماری را به طور موثر مشخص می کند. کیفیت داده ها یک ویژگی اساسی است که اعتبار اطلاعات برای تصمیم گیری در سازمان ها را مشخص می کند. کیفیت داده ها با گذشت زمان به سرعت در پایگاه داده های دنیای واقعی تخریب می شوند و بر نتایج معدن تاثیر می گذارند.

داده ای کثیف سالانه میلیاردها دلار برای تجارت آمریکا هزینه می برند، که منجر به تصمیم گیری های ضعیف می شود. شکل 1 ارتباط بین هزینه هایی که برای داده هایی با کیفیت ضعیف خرج می شود و هزینه هایی که برای داده هایی با کیفیت بالا خرج می شود را نشان می دهد.

یک تجارت بین هزینه هایی که توسط کیفیت ضعیف داده ها و هزینه هایی که توسط کیفیت بالای داده ها است ، وجود دارد. افزایش داده های با کیفیت ضعیف در نتیجه گیری از داده های با کیفیت بالا تاثیر می گذارد که منجر به تفکر در مورد چگونگی اطمینان و دستیابی به کیفیت داده ها می شود.



بنابراین ، تمرکز بر تشخیص و اصلاح مشکلات ناسازگاری داده ها به عنوان فرایند تمیز کردن داده ها مهم است . به طور خاص ، تضمین داده های قابل اعتماد با کیفیت بالا یک مزیت رقابتی برای همه ی صنایع است؛ که به راه حل های دقیق اسکرابینگ داده نیاز دارد.

اصطلاح تمیز کردن داده ها (scrubing داده ها) مرحله مهم پیش پردازش داده ها محسوب می شود. این واقعا برای حفظ سوابق متناقض ، قبل از استخراج و بررسی داده ها ، کار می کند. تمیز کردن داده ها اساسا در چندین برنامه کاربردی مانند انبار کردن داده ها ، مدیریت کیفیت داده ها و کشف دانش در پایگاه های داده وجود دارد. داده کاوی یک حوزه تحقیق است که به معنای فرایند ایجاد الگوهای پنهان و ناشناخته در پایگاه داده ها است و از این واقعیت برای ساخت مدل استفاده می کند. داده کاوی روش ها و فناوریهایی را برای تبدیل مقادیر

زیادی از داده ها برای پردازش و تجزیه و تحلیل به اطلاغات مفید ، برای اهداف تصمیم گیری فراهم می کند.

داده کاوی در الگوریتم های فعلی تمیز کردن داده ها، موضوعی محبوب تر و ضروری تر

می شود. زیرا پاک کردن داده های دستی نیز فرایندی طاقت فرسا و وقت گیر است و خود به تنهایی مستعد خطا است.

بدون شک، اکثر تکنیکهای تمیز کردن داده ها در نوشته ها از تطابق رکورد (مدارک) استفاده می کنند ، که احتمالا سوابق اشتباه را با سوابق درست (تمیز شده) اصلی مقایسه می کند. مقابله با مسئله ناسازگاری داده ها با بکار بردن داده های خود بدون نیاز به داده های اصلی اجباری است. درحین تثبیت ناسازگاری داده ها از جمله وابستگی های عملکردی و وابستگی های عملکردی شرطی را باید بجا آورد.

شناسایی مقادیر متناقض (ناسازگار) یک گام اساسی در فرایند تمیز کردن داده ها است.

در حال حاضر، افزایش داده ها در بیشتر حوزه های برنامه (اپلیکیشن) یک چالش مهم برای اطمینان از ثبات و صحت کیفیت داده ها است که برای اهداف مدیریت داده ها است که برای اهداف مدیریت داده استفاده می شود. داده های نادرست ضبط شده بصورت الکترونیکی ، منجر به داده های بی کیفیت می شود.

در اینجا، ما علاقه مند به ایجاد قوانین قابل اعتماد و دقیق برای کیفیت داده ها هستیم که برای بعدا از آن برای اصلاح ناسازگاری داده ها در چندین دامنه برنامه استفاده می شود.

مشارکتها:

مشارکت های اصلی این مقاله شامل ارائه تکنیک هایی برای تقویت فرایند تمیز کردن داده ها است. به طور خاص مشارکت های کاغذی شامل موارد زیر هستند:

پیشنهاد تکنیک ICCFD MINER: برای ایجاد قوانین تمیز کردن داده های قابل اعتماد و عمدتا بر اساس الگو های بسته شده مکرر .

پیشنهاد تکنیک MICCFD :برای افزایش عملکرد اولین تکنیکهای پیشنهادی ICCFD MINER در تولید قوانین نظافت داده ها . این پیش رفت در استخراج الگو های حداکثر مکرر و ژنراتور های مرتبط با آنها بجای استفاده از الگوهای مکرر بسته بدست می آید. این بعنوان مکانیسم استراتژی جست و جوی موثر برای کاهش اندازه دامنه فضای جست و جو عمل می کند.

الگوریتم T-REPAIR : این الگوریتم برای قوانین ایجاد شده پیشنهاد شده است و مقادیر ورودی t رابطه برای اعتبار سنجی خطاهای تعمیر از مجموعه داده شده است.

سازمان:

این مقاله به شرح زیر برگزار می شود:

بخش 2 در مورد کارهای مرتبط بحث می کند.

در بخش 3 پیش زمینه و تعریف مفاهیم کیفیت داده ارائه می شود.

سپس در بخش 4 تمیز کردن داده ها را در کاربردهای پزشکی برجسته می کند.

بخش 5 تکنیک های پیشنهادی ICCFD-MINER و MICCFD-MINERو

T-REAPAIR را ارائه می دهد.

بخش 6 به مطالعه تجربی و نتایج انجام شده برای مجموعه های مختلف پزشکی می پردازد.

سرانجام بخش 7 کار پیشنهادی را به پایان میرساند و روند های آینده را بررسی می کند.

2. کار مرتبط:

متاسفانه علی رغم نیاز مبرم به تکنیک های دقیق و قابل اعتماد برای افزایش کیفیت داده ها وتمیز کردن داده ها ، هنوز هیچ راه حل اساسی برای رفع این مشکلات ارائه نشده است. بحث و تحلیل کمی در مورد افزایش ثبات داده ها صورت گرفته است. با این حال ، بسیاری از کارهای اخیر بر تطابق رکورد و تشخیص تکراری بودن داده ها متمرکز شده است.محققان پایگاه داده و کیفیت داده ها انواع محدودیت های یکپارچگی را بر اساس وابستگی های عملکردی(FD) مورد بحث قرار دارند. در مرجع نویسندگان الگوریتم FD-MINE را پیشنهاد میکنند که وابستگی عملکردی را از رابطه داده شده کشف می کنند. یک بررسی و مقایسه جامع از هفت الگوریتم برای کشف وابستگی های عملکردی در مرجع مطرح و بحث شده است.

الگوریتم های مورد بررسی شامل

TANE , FUN , FD-MINE , DFD , DEP-MINER, FASTFDS , FDEP

هستند که به طور گسترده در مرجع نشان داده شده است. با این وجود FD های سنتی عمدتا برای طراحی طرحواره توسعه می یابند اما اغلب قادر به تشخیص خطاهای ارزش معنایی داده ها نیستند. محققان دیگر بر گسترش FD تمرکز می کنند. آنها آنچه را که به اصطلاح وابستگی های عملکردی شرطی و وابستگی گنجاندن شرطی گفته می شود را برای ضبط کردن خطاها در داده ها ارائه داده اند.

الگوریتم هایی که برای کشف قوانین CFD از رابطه پیشنهاد شده اند عبارتند از:

الگوریتم MINER CFD : برای کشف وابستگی های عملکری شرطی ثابت .

2 . الگوریتم CTANE : که (ANE) را برای کشف (CFD)های عمومی گسترش میدهد و FCFD را برای کشف CFD ها عمومی با استفاده از یک استراتژی جست و جو ی عمیق به جای روش سطح بندی همانطورکه در الگوریتم (CTANE) استفاده می شود.

چندین تکنیک کیفیت داده برای تمیز کردن تاپلس های کثیف از پایگاه داده ها ارائه شده است. زیرا محققان هدف خود را برای دستیابی به اطلاعات مهم منتسب به پایگاه داده ها می دانند. در مرجع نویسندگان سه مدل را برای مشخص کردن کامل بودن اطلاعات نسبی در مورد داده های پیشنهادی که از طریق آنها مقادیر ممکن است از بین بروند، پیشنهاد می کنند. روش های استنتاج آماری در منابعی که شامل اطلاعات گمشده و خطاهای صحیح به طور خودکار است، بررسی می شود. این رویکردها مقادیر گمشده را برای بهبود کیفیت داده ها نگه می دارد.

از بخش فناوری چندین ابزار منبع باز برای پردازش داده های کثیف تولید شده است.

OPEN RENE , DATA WRANGLER دو ابزار منبع باز برای کار با داده های از دست رفته برای تمیز کردن آنها است که توضیح مفصل آنها در منبع آمده است.

علاوه براین انواع روش های تبدیل داده ها مانند ابزارهای تجاری (ATL) (استخراج، تحول، بارگیری) وجود دارد. روش های استخراج بر روی استخراج داده ها بر روی منابع همگن یا ناهمگن تمرکز دارد. هدف از روش های تبدیل و دگرگونی ، ذخیره داده ها در قالب یا ساختار مناسب برای تحقیق و تحلیل است. روش بارگیری مربوط به بارگذاری داده ها به یک مخزن منبع داده، از جمله انبار داده ها یا سایر منبع داده های جداگانه وابسته به نیاز سازمان است. این ابزارها برای تمیز کردن داده ها ، برای پشتیبانی از هرگونه تغییر در ساختار، بازنمایی یا

محتوای داده ها تهیه شده اند . استفاده از قوانین ویرایش به همراه داده های کارشناسی ارشد منبع شرح داده شده است. چنین قوانینی می توانند با بروز رسانی کپسول های ورودی با داده های اصلی برخی از موارد خاص را بدست می آورند. با توجه به محدودیت ها قوانین ویرایش دارای شناسایی پویا بوده و با داده های کارشناسی ارشد ارتباط دارد. با توجه به یک ورودی تاپلس که با یک الگو مطابقت دارد ، قوانین ویرایش به ما می گویند که کدام ویژگی ها تاپلس باید بروز شوند و چه مقادیری از داده های استاد (اصلی) باید به آنها اختصاص داده شود. این رویکرد مستلزم تعیین قوانین ویرایش به صورت دستی برای هردو رابطه است یعنی رابطه اصلی و رابطه ورودی که بسیار گران و پر هزینه است.

ترمیم از راه حل اکتشافی مبتنی بر تابع جداول هزینه ی دو بروز رسانی استفاده می کند که همیشه ترمیم و تعمیر قطعی ندارد. ویرایش قوانین ،کاربران را ، ملزم به بررسی هر نوع حلقه ای که پرهزینه است می کند. علاه بر این کار های زیادی در ادبیات با تکیه بر شباهت خاص دامنه و اپراتور های تطبیق ارائه شده است. این آثار شامل: تطابق رکورد ، پیوند رکورد ، تشخیص کپی و ادغام است . این رویکردها دو کارکرد را بنام مطابقت و ادغام تعریف می کنند. در حالیکه عملکرد تطابق کپی رکوردها را شناسایی می کند؛ عملکرد ادغام دو رکورد تکراری را به یکی ترکیب می کند سرانجام نتیجه گرفته می شود که روش های موجود تاثیرات قطعی و قایل اعتماد برای مسئله نا هماهنگی را تعیین نمی کند . چنین روش هایی هنگام تشخیص خطاها در داده های حساس مانند پرونده الکترونیکی پزشکی (EMR) سیستم های بهداشت و درمان به درستی کار نمی کنند.

در این مقاله (CFD ثابت) به عنوان یک مورد خاص از قوانین انجمن مجددا مورد استفاده قرار میگیرد و برای حل مشکل تشخیص و تعمیر خطاهای ناسازگاری در پایگاه داده استفاده می شود.

3 .مقدماتی

چندین اصطلاح از ادبیات وجود دارد که مربوط به تکنیکهای پیشنهادی است. تعریف این اصطلاحات شامل وابستگی های عملکردی – شرطی ، الگوهای مکرر بسته و مکرر maximal ، مشکل CFD ثابت و فضای جست و جوی هرس ارائه شده است.

وابستگی های عملکردی (FD) بعنوان محدودیت بین دو مجموعه از ویژگی ها در رابطه با یک پایگاه داده تعریف می شود.

FD روی رابطه R بعنوان $X \rightarrow Y$ که در آن هر مقدار X دقیقا برابر با یک مقدار Y است.

در اینجا X بعنوان تعیین کننده در نظر گرفته می شود و Y بعنوان وابسته در نظر گرفته می شود.

FD رابطه بین تمام ترکیبات ممکن از جفت های ارزش ویژگی را توصیف می کند و برای کار برروی طراحی طرحواره که معنایی از داده ها را برای تمییز کردن داده ها ضبط نمی کند، توصیف نمی کند.

وابستگی های عملکردی شرطی (cfd) توسعه یافته DF است که هدف آن تشخیص ناسازگاری داده ها بین تاپل ها در یک رابطه واحد است .

CFD Q در رابطه R ، یک جفت است $(X \rightarrow Y, TP)$ که در آن $X \rightarrow Y$ یک جفت در R است و TP الگوی TUPPLE از Q با خواص X و Y است. برای هر ویژگی A در TP[A] ، X U Y یا ثابت در دامنه A است یک متغییر نامشخص '_' .

در این مثال، رکورد جدول روابط با مشتری مورد بررسی قرار می گیرد. اینگونه سوابق دارای خصوصیات زیر است:

کد کشور(CC) ، کد منطقه(AC) ، شماره تلفن (PN) ، نام (NM) ، خیابان (STR) ، شهر (CT) و کد پستی (ZIP) همانطور که در جدول 1 نشان داده شده است.

هدف از این مثال توضیح تفاوت بین FD و CFD است. FDهای سنتی که در جدول 1 وجود دارد بشرح زیر است:

F1 : [CC.AC] → CT

F2 : [CC.AC.PN] → STR

F1 بیان می کند که اگر دو مشتری کد کشوری و کد منطقه یکسانی داشته باشند آنها یک شهر مشترک هم دارند بطور مشابه برای F2 .

موارد زیر CFD است که در جدول شماره 1 وجود دارد.

Q0 : ([CC.ZIP]→STR.(32.-||-))

Q1 : ([CC.AC]→CT.(40.872 .-||-UN))

Q2 : ([CC.AC]→CT.(32.222.-||-VIZAG))

Q3 : ([CC.AC]→CT.(40.101.-||-EDI))

قوانین بعنوان متغییر CFD تا Q0 طبقه بندی می شوند جایی که Q1 و Q2 ، CFD ثابت هستند در این کار ما نگران چنین CFD ثابت هستیم. CFD Q0 ادعا می کند کدzip به طور منحصر بفرد STR را تعیین ی کند. این FD بیشتر از همه رابطه با الگوی CC=32 را در زیر مجموعه ای از تاپل ها نگهداری می کند.

CFD Q1 اطمینان می دهد در صورتی که اگر مشتری دارای کد کشوری CODE=40 و کد منطقه CDOE=872 باشد شهر مشتری در سازمان ملل متحداست (UN) بطور مشابه برای Q2 و Q3 داریم.

CFD Q0,Q1,Q2,Q3 نشان می دهد که این قوانین نمی توانند توسط FDها نمایان شوند .

الگوهای متداول بسته ، اگر در یک سوپرست مناسب با همان پشتیبانی کنجانده نشود، الگوی متناوب بسته می شود.

ژنراتور Y با یک الگوی غالبا بسته X ، یک محدودیت الگویی با داشتن همان پشتیبانی از X است و هیچ زیر مجموعه ای با پشتیبانی یکسان ندارد. مجموعه الگوهای مکرر بسته، بی ضرر است که اطلاعات کاملی را در مورد الگوهای مکرر مربوط به آن ارایه می دهد.

از مجموعه الگوهای مکرر بسته ، به آسانی میتوتنیم هویت و پشتیبانی از همه الگو های مکرر را بدون استخراج مجدد بانک اطلاعاتی استخراج کنیم. در عین حال در یک زمان یکسان ، الگوهای مکرر بسته ، به خودی خود می توانند سفارش ها با اندازه کوچکتر از اندازه الگو های مکرر قبول کنند؛ بخصوص در پایگاه داده متراکم.

برای مثال از جدول ((40,827,UN,8422) , [CC,AC,CT,ZIP]) الگوی بسته ای با پشتیبانی 3 است. این الگوی بسته دارای دو الگوی ژنراتور است ((40,827) , [CC,AC]) و ((08422) , [ZIP]) است که هر دو دارای پشتیبانی برابر با 3 هستند. الگو های فرکانس حداکثر، MAXIMAL نامیده می شوند. زیرا هیچ گونه سوپرست مکرر ندارند.

به عبارت دیگر الگو های مکرر maximal هستند اگر هیچ یک ار سوپر ست های آنها مکرر نباشد. الگو مکرر حداکثر استفاده از معادن به کشف الگو های طولانی در بانکهای اطلاعاتی متراکم کمک می کند. استخراج الگو های ضربهای حداکثر به یک مسئله مهم تبدیل شده است. زیرا مجموعه الگوهای مکرر حداکثر نه تنها الگو های منحصر بفرد ، بلکه تعداد الگو های حداکثر مکرر نیز می تواند به طور قابل توجهی کمتر از تعداد بسته های مکرر بسته شود. مجموعه الگو های حداکثر مکرر بدین ترتیب زیر مجموعه ای از الگوهای مکرر بسته است که زیر مجموعه ای از همه الگو های متداول است. این الگو ها با استفاده از مکانیسم جست و جوی موثر برای کاهش اندازه دامنه ی فضای جست و جو استخراج می شوند.

علاوه بر این ، مجموعه این الگو ها یک مجموعه حداقل است ،یعنی کوچکترین مجموعه ای که از آن همه الگوهای مکرر استفاده می شود.

با توجه به مجموعه ای از قوانین MICCFD و چند تایی در رابطه با R ، الگوریتم هر کدام را چند بار تکرار می کند تا در برابر هر قاعده آزمایش شود.

سپس شمارنده را برای تمامی فعالیت هاپاک میکند (ارزش انرا false میگذارد)وچند جایگاه را برای تعمیر اشتباهات در نظر میگیرد . بعد برای هر قانون ازمخزن MICCFD تکرار میشود و شمارنده قوانین را افزایش میدهد. این الگوریتم چک میکند تا اگر سمت چپ قانون با چند تایی از رشته ها مطابقت داشته باشد سمت راست قانون را نیز با چند تایی ازرشته مقایسه میکند ودر صورت مطابقت ان نیز, با اطلاعات معتبر و واقعی موجود دررشته ادامه می دهد. در غیر اینصورت رشته بروز رسانی میشود و نشانگر بروز رسانی ان به true تغییر میکند. در نهایت چند تایی های بروز رسانی شده را برمی گرداند. و این چرخه تا تمام شدن چند تایی ها در رابط ادامه پیدا میکند.

6- مطالعه تجربی

در این بخش یک مطالعه تجربی برای اعتبارسنجی تکنیکهای پیشنهادی ارائه شده است. مطالعه تجربی در هر دو مجموعه داده های واقعی و زندگی مصنوعی از یکی از حوزه های حیاتی پزشکی یعنی دامنه پزشکی انجام شده است.

مجموعه داده های بهره برداری شده اطلاعات زیادی در مورد بیماران و وضعیت آنها دارند .

این مجموعه داده ها برای ارزیابی عملکرد سه روش پیشنهادی i.e. ICCFD _miner ، MICCFD _miner و T _Repair استفاده می شود. از تکنیک های پیشنهادی برای تولید قوانین قابل اعتماد از این مجموعه داده ها استفاده می شود ، و همچنین برای رفع خطاهایی که به طور خودکار یافت می شوند برای دستیابی به داده های یک حالت سازگار استفاده می شوند.

داده های تمیز شده برای دسترسی به تصمیم گیری و اهداف مدیریت ، به دسترسی به داده های تقاضا تبدیل می شوند و تصمیم گیری های دقیقی را بر اساس کیفیت دقیق داده های آنها انجام می دهند.

تکنیکهای پیشنهادی با استفاده از چندین اقدامات ارزیابی می شوند از جمله:

-دقت: ایجاد قوانین قابل اعتماد
-اندازه گیری زمان پاسخ

-کارآیی: استخراج الگوهای مکرر حداکثر
-توصیه فضای حافظه
-مقیاس پذیری با اندازه دیتابیس و ریاضت رابطه

-عامل قطعی(CF): صحت قوانین تولید شده در تولید را بررسی کنید
-اعتبارسنجی: اثربخشی الگوریتم تعمیر _T با استفاده از مخزن قوانین MICCFD-miner

1-6 تنظیم آزمایشی

آزمایش های گسترده ای با استفاده از مجموعه داده های مصنوعی و واقعی انجام می شود.

چنین داده هایی از مخزن یادگیری ماشین UCI (http://achive.ics.uci.edu/ml/)یعنی تیروئید ، تومور اولیه ، کاردیوتوگرافی ، قارچ ، بزرگسالان ، شنوایی شناسی ، پیما دیابتی و شفافیت قلب بارگیری میشود .

جدول4 تعداد اشتراکات و تعداد نمونه ها در هر مجموعه داده را نشان می دهد.

تمام آزمایشات با استفاده از جاوا (JDK1.7) انجام می شود .
اجرای این برنامه بر روی ماشینی با پردازنده دو پنتیوم T3400 و 2 گیگابایت حافظه اجرا شده در ویندوز 7 اس اس انجام می شود .
هر آزمایش بیش از 5 بار تکرار می شود و میانگین آن گزارش می شود.

نتایج تجربی6-2

در این بخش نتایج حاصل از مطالعه تجربی بیش از سه روش پیشنهادی نشان

داده شده موردبحث و بررسی قرار گرفته است

چندین آزمایش بر روی مجموعه داده ها از تکنیک های دامنه کاربرد پزشکی با استفاده از یکی از اقدامات ذکر شده انجام شده است

1-2-6 آزمایش

هدف از این آزمایش ، سنجش صحت تولید قوانین دقیق و قابل اعتماد ، و همچنین اندازه‌گیری زمان پاسخ است.

ارائه و اندازه گیری دقیق و زمان اندازه گیری پاسخ تولید شده از اولین تکنیک پیشنهادی ICCFD-Miner ، که حداقل و غیر زائد است با CCFD-ZartMNR مقایسه می شود.
الگوریتم [17] مجموعه داده های بیش از داده ها ، به عنوان مثال ، پرتقال تو ، کاردیوتوگرافی ، تومور اولیه ، کلیولند 14 قلب و دیابت پیما به شرح زیر است:

الف)
آزمایش 1 قوانین قابل اعتماد را از مجموعه داده های تیروئید تولید می کند. به یاد بیاورید که مجموعه داده های تیروئید شامل 30 عدد و

3772 رشته داده های بیمار در توصیف تشخیص های تیروئید است. نتایج انجام شده در این مجموعه داده در نشان داده شده است.

در شکل 9 با تغییر مقادیر آستانه حداقل پشتیبانی و حداقل اطمینان ، توجه می شود که تکنیک پیشنهادیCCFD_Miner از -CCFD ZartMNRalgorithm در تولید علاقه دقیق حداقل قوانین غیر زائد استفاده می کند. به عنوان مثال ، با توجه به حداقل پشتیبانی ..97 - andmin andنتیجه گیری .0.99 ، تعداد قوانین تولید شده به ترتیب 85 و 220 توسط الگوریتم های پیشنهادی ICCFD-Miner و theCCFD_ZartMNR ارائه شده است.
در شکل 10 ، نتایج نشان می دهد که ICCFD_Miner همچنین الگوریتم CCFD-ZartMNR را در تولید قوانین با مقادیر مختلف پشتیبانی و مقادیر مختلف با توجه به زمان پاسخگویی انجام می دهد.

ب)

نتایج تجربی انجام شده بر روی مجموعه داده اولیه تومورورودیزازها در شکل نشان داده شده است. 11 و 12. به یاد بیاورید که این مجموعه داده شامل 18 عدد و 339 رکورد است. اثربخشی تکنیک پیشنهادیICCFD-Miner را در برابر الگوریتم CCFD-ZartMNR الگوریتم مجموعه داده اولیه تومور با توجه به تعداد قوانین تولید شده و زمان پاسخ تأیید می کند.

پ)
اثربخشی تکنیک پیشنهادی فنی ICCFD-Miner در برابر الگوریتم CCFD-ZartMNRبر داده های بیمار مبتلا به بیماری دیوتوکوگرافی Carبررسی شده است. به یاد بیاورید که این مجموعه داده شامل 23 ویژگی و 2126 رکورد است .
نتایج اثبات می کند که روش پیشنهادی همانطور که در شکل نشان داده شده ، مؤثرتر از الگوریتم دیگر است.

6.2.2. Experiment-2

The second experiment is conducted to evaluate the effi-
ciency of extracting maximal frequent patterns from the sec-
ond proposed MICCFD-Miner technique compared to the
existing CCFD-ZartMNR [17] technique. Also, memory space
consumption from second proposed MICCFD-Miner tech-
nique compared to existing CCFD-ZartMNR technique is
measured.

6.2.2. Experiment-2

The second experiment is conducted to evaluate the effi-
ciency of extracting maximal frequent patterns from the sec-
ond proposed MICCFD-Miner technique compared to the
existing CCFD-ZartMNR [17] technique. Also, memory space

consumption from second proposed MICCFD-Miner tech-

nique compared to existing CCFD-ZartMNR technique is

measured.

6.2.2. آزمایش -2

آزمایش دوم به منظور ارزیابی اثربخشی استخراج الگوهای حداکثر مکرر از تکنیک پیشنهادی MICCFD-Miner با استفاده از روش دوم موجود در مقایسه با تکنیک موجود [17] CCFD-ZartMNRانجام شده است. همچنین ، فاصله فضا حافظه از فناوری پیشنهادی دوم MICCFD-Miner در مقایسه با تکنیک موجود CCFD-ZartMNR اندازه گیری شده است.

این آزمایش بر روی مجموعه داده های قارچ انجام شده است. به یاد داشته باشید که این مجموعه داده شامل 23 صفت و 8124 رکورد است. مجموعه داده قارچ توصیف کننده های فیزیکی قارچ است ، در حالی که هر گونه قارچ به عنوان تعریف قابل تشخیص ، کاملاً مسموم ، یا از نظر ناشناخته قابل شناسایی است. نتایج حاصل از MICCFD-Miner با استفاده از تکنیک موجود -CCFD نسبت به قارچ نشان داده شده در شکل 15 و ZartMNR16

اثربخشی در کاهش الگوی تعداد ایجاد شده توسط دومین پیشنهادی MICCFD-Minertechnique نسبت به یک روش دیگر(-CCFD ZartMNR) را بدون اطلاع از اطلاعات نشان می دهد. محور x مقادیر پشتیبانی متفاوتی است و محور y تعداد کل الگوهای استخراج شده است و مقدار ثابت ثابت تعیین شده برابر با 1. را تعیین می کند.

حداکثر الگوهای مکرر استخراج شده از تکنیک پیشنهادی MICCFD_Miner چهل و یک الگو است ، اما تعدادالگوهای بسته ی تولید شده از تکنیک CCFD-ZartMNR 140تا هستند.
این نتایج اثربخشی استخراج الگوهای حداکثر مکرر با استفاده از MICCFD_Minerرا به عنوان اولین گام در تولید قوانین دقیق و قابل اعتماد بودن به جای استخراج الگوهای بسته مکرر از تکنیک های موجود فعلی نشان می دهد .
با این حال ، شکل 16 نشانگر احتمالی فضای حافظه تعدادی از الگوهای استخراج شده از روش MICCFD_Miner پیشنهادی نسبت به روش CCFD-ZartMNRاست. محور x مقادیر پشتیبانی است ، و محور y میزان مصرف حافظه در مگابایت است ، در مقدار ثابت ثابت fixed برابر با 1 است.

6.2.3. آزمایش -3

این آزمایش با هدف ارزیابی میزان مقیاس پذیری تکنیک پیشنهادی MICCFD_Miner در ثانویه در مقایسه با روش -CCFD ZartMNR [17]، با تغییر اندازه مجموعه داده و اساس رابطه انجام گرفته است. این آزمایش بر روی Twodatasets ، یعنی بزرگسالان و شنوایی شناسی انجام می شود.

الف)

در این آزمایش ، مقیاس پذیری دومین تکنیک پیشنهادی MICCFD_Minerدر هنگام افزایش تعداد لپ تاپ ها در تعداد fixedویژگی ها در مقایسه با تکنیک CCFD-ZartMNRارزیابی می شود. نتایج حاصل از Adultdataset در شکل 17 نشان داده شده است. بیاد بیاورید که این مجموعه داده شامل 15 ویژگی و 32.561 رکورد است. این پیشگویی را توضیح می دهد که آیا درآمد بیش از 50 کیلو دلار در سال است یا بر اساس داده های استنادی نیست. در حالی که

سرخپوشان به 15 پوند ، حداقل اطمینان به 0/95 و حداقل پشتیبانی از 5/0 ، اندازه مجموعه داده (تعداد رشته ها) از 5 تا 30 کیلوگرم افزایش می یابد. زمان پاسخ دومین پیشنهاد شدهMICCFD_Miner در مقایسه با تکنیک های موجود CCFD-ZartMNR موجود با تعداد متنوعی از قلاب ها در شکل 17 گزارش شده است ، که نشان می دهد مقیاس بندی زمان خطی روش پیشنهادی MICCFD_Miner ما نسبت به -CCFD ZartMNRtechnique موجود. محور x تعداد رشته ها * 1000 و محور y اندازه گیری زمان پاسخ در میلی ثانیه(ms) است.
این آزمایش نتیجه گیری می کند که تکنیک پیشنهادی _MICCFD Minerاز یک روش دیگر موجود CCFD-ZartMNR در مقیاس پذیری در افزایش تعداد مشبک با زمان پاسخ کمتر فراتر می رود.

ب) ارزیابی قابلیت مقیاس پذیری با افزایش تعداد بعدی صفات در تعداد fi xed توپل برای تکنیک پیشنهادی MICCFD_Miner در رابطه با تکنیک CCFD-ZartMNR انجام شده است. نتایج به دست آمده در مجموعه داده های شنوایی شناسی ، که شامل 70 ویژگی و 226 نمونه است. نتایج این آزمایش به نمایش درآمده است در حالی که اندازه بانک اطلاعاتی به 226 رشته ، حداقل اطمینان به 1 ، و حداقل پشتیبانی 0.99 fi ، زاویه thefiing از 10 به 70 ویژگی در حال افزایش است.
شکل 18 اندازه گیری پاسخ دومین تکنیک فنی پیشنهادی MICCFD_Miner از تکنیک موجود CCFD-ZartMNR موجود را نشان می دهد که تعداد صفات را با اختلاف زیاد می کند.

6.2.4. آزمایش -4

در این آزمایش ، درستی از خروجی تولید شده از دو روش پیشنهادی ، یعنی ICCFD-Miner وMICCFD-Miner ، در مقایسه با روش CCFD-ZartMNR با استفاده از اندازه گیری ضریب اطمینان بررسی می شود. ضریب اطمینان ((CFبه عنوان معیاری برای سنجش قدرت قواعد تولید شده و تأیید حذف افزونگی قوانین با استفاده از کیفیت سایر قواعد تعریف شده است.

ضریب اظمینان (x-A)= p(A|x)-p(A) \ 1- p(A)

نتایج فاصله اندازه گیری ضریب اطمینان عجیب است [-1: 1] ، در حالی که مقادیر به شرح زیر است:

(i)مقادیر CF برابر 0 است ، به معنای عدم وجود شواهد از قوانین.

(ii)مقادیر CF کمتر از 0 ، که به معنای قوانین اثبات منفی است.

(iii) مقادیر CF بیشتر از 0 ، به معنای نفع برای صحیح بودن مقررات تولید شده است.

ضریب اطمینان وی به شش مجموعه داده پزشکی شامل تیروئید ، تومور اولیه ،شفافیت14- قلب ، اتاق ماس ، قلب و عروق و مجموعه داده های شنوایی شناسی اعمال می شود. نتایج حاصل از دو روش پیشنهادی در مورد این مجموعه داده ها در جدول 5 نشان داده شده است. این نتایج خروجی قوانین تولید شده از دو روش پیشنهادی را تضمین و اعتبار می کند. این نشان می دهد که چنین تکنیک هایی همیشه در مقایسه با CCFD-ZartMNRtechnique قوانین دقیقی را با ضریب خستگی بیشتر از 0 ایجاد می کنند. CCFD-ZartMNR قوانینی را ایجاد می کند که شواهد منفی دارند (کمتر از صفر)

6.2.5. آزمایش -5

در این آزمایش ، ICCFD-Miner و MICCFD-Minertechniques در مقابل تکنیک CCFD-ZartMNR با مقایسه فضای مصرفی حافظه مقایسه شده اند. این آزمایش در مجموعه داده اولیه تومور انجام شده است ، که شامل 18 ویژگی و 339 پرونده است. نتایج حاصل از این آزمایش در شکل موجود است. 19 و 20. فیگ. 19 نشان می دهد که MICCFD-Minertechnique پیشنهادی دوم در صرفه جویی در فضای حافظه نسبت به اولین تکنیک پیشنهادی -ICCFD

Miner و .CCFD-ZartMNRtechniqueشکل 20 همچنین نشان می دهد که روش دوم پیشنهاد شده MICCFD-Miner از زمان پاسخ دو روش دیگر بهتر است. همچنین مشاهده می شود که هیچ تفاوت معنی داری بین هر دو روش پیشنهادی وجود ندارد.

6.2.6. آزمایش 6

در این آزمایش ، کارایی الگوریتم T_Repair با استفاده از خروجی MICCFD-Minerبا استفاده از قوانین بازآزمایی ، ارزیابی می شود. الگوریتم T_Repair در مجموعه داده های پزشکی قبلاً ذکر شده تأیید شده است. نتایج این آزمایش در شکل نشان داده شده است. جدول 21 و 22 و جدول 6. در حالی که شکل 21 تعدادرشته های به روز شده از مجموعه داده های بعد از اعمال الگوریتم T_Repair را نشان می دهد ، شکل 22 اندازه گیری زمان پاسخ برای استفاده از الگوریتم T_Repairرا نشان می دهد. مشاهده شده است که زمان پاسخ عمدتا بر اساس پایان نامه رابطه و تعداد خطاهای به روز شده است. در جدول 6 درصد از توپ های fixed برای هر مجموعه داده به صورت عددی نشان داده شده است.

6.3 بحث نتایج

نتایج حاصل از آزمایشات انجام شده نشان می دهد که هر دو الگوریتم ICCFD-Miner و MICCFD-Miner برای CCFD_ZartMNR با تکنیک MICCFD-Miner پیش بینی شده اند. فراتر از تکنیک پیشنهادی ICCFD-Miner به دلیل اندازه گیری آسانسور هنگام تولید قوانین قابل اعتماد و حداقل اطمینان دقیق برای تشخیص داده های عدم تناقض داده ها رخ می دهد. علاوه بر این ، تکنیک های پیشنهادی

ICCFD_Miner در الگوهای بسته با ژنراتورهای مرتبط با آنها ، یعنی سوپر مارکت های الگوهای بسته ، به عنوان یک فضای جستجو برای تولید دقیق قوانین حداقل و غیرقابل رعایت دقیق ، مورد علاقه است. از آنجا که تکنیک MICCFD-Miner روی الگوهای مکرر حداکثر تمرکز دارد ، تعداد الگوهای ایجاد شده را در مقایسه با روش CCFD-ZartMNR بدون اطلاع اطلاعات کاهش می دهد. این کاهش باعث می شود که از نظر حافظه مصرفی و زمان پاسخگویی بهتر به دست آمده منجر شود. علاوه بر این ، درجه کیفیت خوب قوانین تولید شده توسط تکنیک های پیش بینی شده با استفاده از واقعیت اطمینان حاصل می شود.

7. نتیجه گیری

و کارهای آینده در این مقاله ، هر دو ICCFD-Miner و MICCFD-Minertechniques برای تولید قوانین دقیق و قابل اعتماد برای حل مشکلات ناسازگاری داده ها در پایگاه داده ارائه داده اند. علاوه بر این ، الگوریتم T_Repair ارائه شده است که عملکرد fi xing تاپل های متناقض را انجام می دهد. هر دو تکنیک پیش بینی شده و الگوریتم برای افزایش کیفیت داده در چارچوب Ageneric گنجانده شده اند. این چارچوبها با استفاده از چندین مجموعه داده از حوزه مراقبت های بهداشتی ارزیابی و تأیید می شوند. انواع آزمایشات برای آزمایش عملکرد تکنیکهای پیشنهادی شامل زمان پاسخ ، فضای ذخیره سازی و قابلیت دستیابی به پایگاه داده انجام می شود. نتایج نشان می دهد که تکنیک های پیشنهادی ، از تکنیک رقیب ، یعنی CCFD-ZartMNR استفاده می کنند. در نهایت ، کار آینده شامل ارائه یک الگوریتم افزایشی برای حفظ و به روزرسانی قوانین در مورد جریانهای داده با استفاده مجدد از اسکن پایگاه داده اصلی ، و همچنین ترمیم داده ها است. علاوه بر این ، عملکرد T_Repairtechnique باید در نظر گرفته شود ، می توان آن را با استفاده از indexing و پیوستن بیرونی برای شناسایی تاپل های بی همتا که نیاز به انجام آن دارند ، افزایش داد.

اسامی دانشجویان :

سمیرا آتشپز

سارا هوفر

آیلین قیصر