مروری بر روش های تشخیص خطا در پایگاه داده

سهیلا عابدینی، محمدرضا ملاحسینی

دانشجوی دکترا کامپیوتر - نرم افزار، گروه کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران

soheylaabedini@gmail.com

عضو هیات علمی، گروه کامپیوتر، واحد میبد، دانشگاه آزاد اسلامی، میبد، ایران

Reza_mollahoseini@yahoo.com

چکیده

کیفیت داده‌ها در امر تصمیم‌گیری سازمان‌ها تاثیرگذار است. به نحوی که تصمیم‌گیری بر اساس داده‌های بی‌کیفیت، سازمان را متحمل هزینه‌های بالایی می‌کند. در این تحقیق پس از بیان اهمیت موضوع از دیدگاه محققین مختلف، روش‌های تشخیص خطا از جمله شبکه‌های عصبی، درخت تصمیم، روش قوانین، هستان‌شناسی و وابستگی تابعی مورد بحث و بررسی قرار گرفته است. نتایج بررسی نشان می‌دهد که هر کدام از روش‌های ذکر شده در قالب روش‌های داده کاوی بر روی پایگاه داده‌های مختلف دارای مزایا و معایب متفاوتی دارند. ولی در کل همه روش‌های ذکر شده قابل پیاده سازی برای این هدف می‌باشند.

کلیدواژه‌ها: تشخیص خطا، پایگاه داده، داده‌کاوی، آشکارسازی خطا، روش‌های فراابتکاری.

1- مقدمه

هوش تجاری از مجموعه‌ای از تکنیک‌ها و ابزار پدید آمده و هدف آن فراهم آوردن مشاغلی با حمایت ضروری از تصمیم‌گیری است. نمونه‌های خدمات هوش تجاری ساده موجود عبارتند از سرویس‌های مختلف جستجو و فیلترینگ و همچنین ارائه گران محتوا و جمع‌آوری کنندگانی که بسته‌های اطلاعاتی نیمه سفارشی را به کاربران خاص ارائه می‌کنند. در یک سطح پیشرفته‌تر مدیریت اطلاعات (IM) می‌تواند به یک مدیر در کنترل و نظارت بر سازمان‌های خاص، تکنولوژی ها یا حوزه‌های تحقیقاتی کمک کرده و به آن‌ها توانایی تحلیل داده‌های ابتدایی را نیز می‌دهد تا بتوانند در سطح رقابت شرکت، بخش یا صنایع به نتیجه‌گیری بپردازد(میکرویانیدیس و تئودولیدیس[1]،2010).

یکی از مهمترین بخش‌های سیستم‌های هوش تجاری، انبار داده‌ی آن‌ها است. انبار داده یک مخزن فیزیکی است که در آن داده‌های رابطه‌ای به طور ویژه‌ای سازمان‌یافته شده‌اند تا بتوانند در سرتاسر سازمان از تصمیم‌گیری پشتیبانی و شامل داده‌های تمیز در یک فرمت استاندارد باشد. از مشکلاتی که در سیستم‌های انبار داده وجود دارد، داده‌های نامتجانس است. داده‌های نامتجانس به دلیل جمع‌آوری اطلاعات از سیستم‌های قدیمی، استانداردهای ناسازگار و قالب‌های نامتجانس به وجود می‌آیند که اغلب باعث بروز مشکلات عدیده‌ای در خصوص یکپارچگی داده و تجزیه و تحلیل‌های پیچیده می‌گردد. انبار داده فرایندی برای دسترسی و استخراج این داده‌های ناهمگن و پالایش (تصفیه) و تبدیل آن‌ها به فرمتی واحد و در نهایت ذخیره‌سازی آن‌ها در یک ساختار واحد که امکان دسترسی آسان و ساده را فراهم می‌کند. حجم داده‌های درون انبار داده بسیار بالا است به ویژه زمانی که سابقه‌ای از داده را نگه‌داری کنیم. ابزارهای تحلیلی نیاز به پویش مقدار بسیار بالایی از داده‌ها را دارند که تاثیر منفی بر محیط‌های عملیاتی دارند. به همین خاطر برای بهبود و بالا بردن کارایی نیاز به جداسازی محیط عملیاتی از محیط انبار داده‌ها می‌باشد (نیکومرام و محمدی، 1391). میزان داده‌های نامتجانسی که امروزه در دسترس سازمان‌ها قرار دارد، مدیریت اطلاعات را تبدیل به کاری بسیار پیچیده و در عین حال حیاتی کرده است؛ چون این داده‌ها ممکن است سرمایه ارزشمندی برای هوش تجاری باشند. امروزه سیستم‌های نرم‌افزاری برای کنار آمدن با پیچیدگی‌هایی که روز به روز هم بیشتر می شوند، از طریق انطباق با تغییراتی که ممکن است در محیط یا بافت عملیاتی‌شان یا در مقتضیات سیستم اتفاق بیفتد، انعطاف‌پذیرتر، قابل اطمینان‌تر، قابل بازیابی، قابل اختصاصی سازی، وابسته‌تر، قابل پیکربندی و خودبهینه ساز و در مصرف انرژی صرفه‌جو شده‌اند (لیموس و همکاران[2]، 2013). رشد روز افزون داده‌ها موجب گردیده تا سازمان‌ها با داده‌های حجیم و منابع ناهمگون و توزیع‌شده روبه‌روگردند. سازمان‌ها جهت تصمیم‌گیری نیازمند استفاده از این داده‌ها و استخراج دانش

[1] Mikroyannidis, A., & Theodoulidis
[2] Lemos et al.

از آن‌ها می‌باشند. جمع‌آوری این حجم از داده‌ها موجب ایجاد موضوع دیگری به نام کیفیت داده‌ها[1] برای سازمان‌ها می‌گردد (عطاییان و دانشپور، 1395). زمانی که داده‌ها از منابع و سیستم‌های مختلف به یک سیستم دیگر منتقل می‌شوند، ممکن است دچار خطا گردند و یا اینکه با مشکلاتی از قبیل فرمت و دامنه ناهمگون مواجه گردند. تصمیم‌گیری بر مبنای این داده‌های فاقد کیفیت علاوه بر اینکه به ساختار سازمان آسیب می‌زند، باعث می‌شود هزینه‌های زیادی به سازمان وارد شود. بر اساس مطالعات انجام شده بیش از 30 درصد داده‌ها فاقد کیفیت هستند و این گونه داده‌ها موجب گردیده سالانه 3 تریلیون دلار به دولت آمریکا خسارت وارد کنند (بسکالس و همکاران[2]، 2014). بنابراین کیفیت داده‌ها در منابع داده از اهمیت بالایی برخوردار است. فراهم نمودن داده‌هایی با کیفیت بالا کاری زمان‌بر و هزینه بر است (فان[3]، 2008). اما در برابر هزینه‌های شکست بر اثر داده‌های فاقد کیفیت بسیار ارزشمند است. کیفیت داده‌ها دارای ابعاد متنوعی است که صحت[4] در میان ابعاد دیگر حایز اهمیت بالایی است. مشکلاتی از قبیل ناقص بودن[5]، ناسازگاری[6] و مقادیر از دست رفته[7] باعث نقض این بعد می‌شوند. تصحیح داده‌ها فرایندی است که برای تشخیص داده‌های ناقص، نادرست و ناسازگار و بهبود کیفیت داده‌ها با اصلاح خطاهای شناسایی شده به کار می‌رود (احمد و همکاران، 2011). روند صحیح داده‌ها می‌تواند وقت‌گیر و خسته‌کننده باشد. اما نمی‌توان آن را نادیده گرفت (ویلیامز و همکاران[8]، 2002). با توجه به حجم داده‌ها، راهکارهای تصحیح تعاملی غیرکاربردی است و نیاز به راهکارهای خودکار وجود دارد. داده‌کاوی یکی از تکنیک‌های کلیدی برای تصحیح داده‌ها است (احمد و همکاران، 2011) و تاکنون راهکارهای مختلفی برای تصحیح داده‌ها ارائه شده است. روش های مطرح شده در این حوزه شامل هستان‌شناسی (بروژمن[9]، 2008)، طبقه‌بندها شامل (درخت تصمیم (رحمان و همکاران، 2013)، شبکه عصبی (بریمن[10]، 1996) و قانون بیز (یاکوت و همکاران[11]، 2013))، یادگیری مرکب (بریمن، 1996) و وابستگی تابعی (هی و همکاران[12]، 2014). برای تصحیح داده استفاده نموده‌اند که خود دارای معایب و مزایایی هستند. با توجه به اهمیت موضوع در این تحقیق قصد داریم روش‌های تشخیص خطا را با ذکر تحقیقاتی در این زمینه مورد بررسی قرار دهیم.

2- مروری بر روش‌های داده‌کاوی تشخیص خطا در پایگاه داده

[1] Data Quality
[2] Beskales et al.
[3] Fan
[4] Accuracy
[5] Incomplet
[6] Inconsistent
[7] Missing value
[8] Williams et al.
[9] Brüggemann
[10] Breiman
[11] Yakout et al.
[12] He et al.

2-1-هستان شناسی

هستان شناسی درک مشترکی از مفاهیم موجود در یک حوزه به همراه روابط بین آن‌ها را ارایه می‌نماید که برای حل مشکل چند معنائی موجود در اکثر کاربردها مفید است. .هستان‌شناسی عبارت است از یک تعریف رسمی از یک دامنه خاص با تعریف نهادها یا هویت‌های هستان شناسی که توصیف کننده آن دامنه مثلا مفاهیم یا طبقات و همچنین نمونه‌ها و روابط آنها هستند. هستان شناسی اطلاعات را به شکل قابل پردازش توسط ماشین شرح داده و لذا توسط عوامل نرم‌افزاری امکان اصلاح موثر آن را فراهم می‌آورد. این اطلاعات معمولا با استفاده از زبان مبتنی بر XML مانند RDFS و OWL نمایش داده می‌شود. تطبیق هستان‌شناسی به طور کلی از سه مرحله اندازه‌گیری شباهت تشکیل شده است : اندازه‌گیری شباهت بین مفاهیم ، اندازه‌گیری شباهت بین خواص و منطق اندازه-گیری شباهت استنتاج (کاک و یانگ[1]، 2010). هستان‌شناسی می‌تواند به لحاظ معنایی یک پایگاه دانش غنی در سیستم‌هایی می‌باشد که در مدیریت اطلاعات تخصص دارند.

هستان شناسی یک پایگاه دانش غنی به لحاظ معنایی را جهت تفسیر موضوعات فاقد ساختار برای سیستم های مدیریت اطلاعات فراهم می‌آورد. برمبنای معانی کدگذاری شده در هستان شناسی، اطلاعات را می‌توان از متونی با زبان طبیعی استخراج کرد و در سطح پیشرفته تر پردازش، دانشی را می توان کشف کرد که به هوش تجاری کمک می‌کند. درهرحال نحوه مدیریت کردن هستان‌شناسی در سیستم های مدیریت اطلاعات پیچیده نیست و عوامل مهم را نادیده می‌گیرد. از لایه بندی هستان‌شناسی یا یکپارچه سازی آن به ندرت استفاده شده و جنبه پویای هستان شناسی که به مکانیزم های ارزیابی مناسب نیاز دارد، اغلب نادیده گرفته می شود. به طور کلی پتانسیل هستان شناسی در مدیریت اطلاعات و هوش تجاری هنوز هم کاملا شناسایی نشده و در عمل مورد استفاده قرار نگرفته است((میکرویانیدیس و تئودولیدیس، 2010).

درپژوهش(مارتین و همکاران[2]، 2011) کاربرد یک ساختار برای هوش تجاری با استفاده از طبقه‌بندی آنتولوژی پرداختند. از این مدل آنتولوژی ارتباط بین داده‌ها با استفاده از درخت تصمیم‌گیری به دست آمد که منجر به توسعه یک مدل هوش تجاری گردید.

در پژوهش امینی و همکاران (1390)، روش ارائه شده از هستان‌شناسی برای شناسایی مقادیر معتبر غیر معتبر پس از شناسایی به همراه تمامی مقادیر معتبر که در هستان‌شناسی موجود است به کاربر نشان داده شده و از او خواسته می‌شود که برای رکورد مورد نظر از میان مقادیر معتبر مقداری را انتخاب کند. مقادیر انتخاب شده به وسیله کاربر به یک حد آستانه از پیش‌تعیین شده رسیده است. از آنها قوانین استخراج می‌شود. از این پس، هنگامی که مقادیر تصحیح‌شده توسط کاربر به یک حد آستانه از پیش‌تعیین شده رسید، از آن‌ها قوانین استخراج می‌شود. از این پس، هنگامی که مقادیر غیر معتبر به وسیله هستان‌شناسی تشخیص داده شد به جای تعامل با کاربر برای تصحیح، از قوانین

[1] Kwak & Yong
[2] Martin et al.

استخراج شده برای اصلاح اتفاده می‌شود. این روش به دلیل نیاز به کاربر در فاز اولیه در حجم داده‌ای بالا مناسب نمی‌باشد.

لیهونگ و همکاران (2010)، یک روش هستان شناسی را در فرآیند ETL از انباره داده ارائه نمودند. بطور کلی آنها در پژوهش خود یک دامنه هستان شناسی را تعریف کردند تا بتواند منابع داده را ردیابی نموده و جستجو کند، قوانینی را برای انتقال داده ها تعریف نماید و ار بین بردن عدم هماهنگی را انجام دهد. در مقاله ارائه شده، دامنه هستان شناسی با ابر داده های انباره داده ادغام شده است. رکوردهای انباره داده به کلاسهایی از آنتولوژی تحت عنوان OWL[1] نگاشت شده است.

جوئل و همکاران (2011)، یک هستان شناسی را مبتنی بر فرآیند ETL برای ساخت یک هستان شناسی انباره داده ارائه نمودند. هدف اصلی آنها در پژوهش خود تولید نمودن انباره داده ای بوده است که بتواند هستان شناسی ای فرآیند ETL را افزایش دهند.

بلاترچ و همکاران[2] (2012)، با استفاده از هستان شناسی و نیازمندیها اقدام به بهینه سازی و تهیه ساختاری برای انباره داده ها نمودند. در این پژوهش یک مدیریت منابع انسانی جدید برای کاربردهای پایگاه داده ای مطرح شد. بجای جایگزاری DBA، هدف طراحی یک سیستم مفهومی مبتنی بر DBA بوده است. بلاترچ و همکارانش در پژوهش خود ابتدا درخواست‌ها و نیازمندی‌های کاربران را مورد تجزیه و تحلیل قرار داده، سپس درخواست ها را در قالب کوئری های SQL Server ارائه می نماید. در نهایت نیز شاخصه هایی را مبتنی بر نیاز های نرم افزار انتخاب نموده و در معماری پیشنهاد شده مورد ارزیابی و مقایسه قرار می‌دهد.

گولیک[3] (2013)، فرآیند انتقال فایل‌های OWL مربوط به هستان‌شناسی را اعمال نمودند. در این مقاله یک روش که بتواند ساختار OWL را در نمای انبار داده‌ها انتقال دهد مورد توجه قرار گرفته است. پس از اینکه قوانین و شرایط مربوط به هستان‌شناسی طراحی شد، یک روش انتقال OWL در نمای ستاره‌ای تولید می‌گردد.

وانگ توم هام و همکاران[4] (2015)، از هستان شناسی و انباره های داده مبتنی بر اعتماد در هوش تجاری استفاده نمودند. بطور خلاصه، آنها روشی امن را طراحی نمودند می توان در بسیار از کاربردها از آن استفاده نمود. در تحقیق دیگر در سال 2016، بوچرا و همکاران[5]، یک معماری مبتنی بر هستان‌شناسی را برای انباره‌های داده به منظور خودکار کردن فرآیند موجود در این انباره ها ارائه نمودند. آنها در تحقیق خود جهت خودکار سازی فرآیند موجود در انباره‌های داده دو مرحله را دنبال کردند، ابتدا کلیه نیازمندی‌های کاربران را در قالب هستان‌شناسی تحلیل و طراحی نموده و سپس نیازمندی‌های تحلیل شده را در قالب یک پیاده‌سازی جدید ارائه نموده تا فرآیندها به صورت خودکار اجرا شوند (بوچرا و همکاران، 2016).

[1] Web Ontology Language (OWL)
[2] Bellatreche et al.
[3] Gulić
[4] Wongthongtham et al/
[5] Bouchra

2-2- شبکه‌های بیزین

در مقایسه با شبکه‌های عصبی، این گروه از دسته‌بندی‌ها، میزان عضویت یک نمونه به هر کلاس را با یک احتمال نشان می‌دهد، همچنین از مفاهیم اماری مانند میانگین، انحراف معیار و یا از هیستوگرام مقادیر ویژگی، برای تولید قانون استفاده می‌نمایند (جین و آبراهام[1]، 2014). یکی از مهم‌ترین روش‌های دسته‌بندی آماری شبکه‌های NB و شبکه‌های بیزی هستند. شبکه بیزی یک مدل گرافیکی است که رابطه احتمالی بین یک مجموعه از متغیرها را بیان می‌کند. ساختار سک شبکه بیزی S، یک گراف جهت‌دار بدون دور است که گره‌ها در آن متغیرهای تصادفی را نشان می‌دهد و یال‌های آن شبکه، یک ارتباط یک به یک بین متغیرها می‌باشد (کوتسیانتیس و همکاران[2]، 2007).

شبکه‌های NB، شبکه‌های بیزی خیلی ساده‌ای هستند که از گراف‌های بدون دور جهت‌دار، با تنها یک والد و چندین فرزند تشکیل شده‌اند و نودهای فرزندان را مستقل در نظر می‌گیرد. این الگوریتم احتمال شرطی هر ویژگی داده شده را به توجه به دسته مربوطه‌اش یاد می‌گیرد. سپس عمل دسته‌بندی با به کار بردن قوانین بیز برای محاسبه مقدار احتمالی دسته نتیجه نموده داده شده، با دقت بالایی انجام می‌شود. مهم‌ترین مزیت دسته‌بندی کننده NB، زمان کمتر محاسبه برای آموزش دیدن است (جان و لنگلی[3]، 1996).

این کلاسیفایر همچنین توانایی غلبه بر پدیده شاخه گمشده را دارد که به این دلیل است که از هر دو احتمال پیشین و شرطی برای محاسبه احتمال هر کلاس بندی ممکن استفاده می کند. از آنجاییکه NBC مبتنی بر رابطه 1 است، ممکن است برخی از مقادیر $P(x(k)|C(i))$ صفر شوند چرا که امکان دارد $x(k)$ مدنظر در $C(i)$ مجموعه‌های آموزشی موجود نباشد.

$$P(X|C(i)) = \prod_{k=1}^{n} P(x(k)|C(i)) \qquad (1)$$

این موضوع ممکن است سبب صفر شدن $P(X|C(i))$ و این یک ضعف جدی برای آن کلاس می باشد. از این رو بجای صفر از یک عدد کوچک استفاده می شود. به لحاظ تئوری کلاسیفایرهای بیز کمترین نرخ خطا را نسبت به تمامی دیگر کلاسیفایرها دارند. هرچند در کاربردهای عملی فرض های آماری موجب کاهش دقت کلاسیفایر می شود. دیگر ضعف کلاسیفایر های بیزی این است که آن‌ها تنها برای مشخصه های گسسته قابل به کار بستن هستند و این موضوع، کاربردهای این کلاسیفایر را محدود می کند (برامر[4]، 2007). علاوه بر آن یک معادله پیچیده در این کلاسیفایر لازم است تا احتمال کلاس‌بندی تخمین زده شود و تمام مشخصه‌ها باید متقابلا مستقل و نابسته باشند.

[1] Jain & Abraham
[2] Kotsiantis et al.
[3] John, G. H., & Langley
[4] Bramer

2-3- روش قوانین

الگوریتم ارائه شده در (تانگ،2010) با استفاده از قوانین تعریف شده به تصحیح یک منبع ناسازگار می‌پردازد. این قوانین از سه بخش اصلی تشکیل می‌شوند که بخش اول و دوم سمت چپ قانون و بخش سوم سمت راست قانون را تشکیل می‌دهند. بخش اول الگوهای شاهد نامیده می‌شود که بیانگر آن دسته از ویژگی‌هایی که با یکدیگر در ارتباط هستد،می‌باشد. بخش دو شامل الکوهای منفی است که بیانگر مقادیر اشتبا برای ویژگی‌ها است. بخش آخر مقادیر واقعی ویژگی‌هایی است که مقدار صحیح را به ازای مقدار غلط بیان می‌کند. در ابتدا هر یک از ویژگی‌های ارائه شده در بخش الگوهای شاهد به تنهایی به عنوان یک کلید انتخاب شده و سپس هر یک با مقادیر خود در یک لیست با نام قانون ذخیره می‌شود. در هر رکورد به جستجوی کلیدهای ذخیره‌شده در لیست پرداخته شده و در صورتی که بخش دوم ر رکورد موجود بود، آن رکورد باید با بخش سوم تصحیح گردد و این روند باید برای تمامی رکوردها تکرار شود. هدف از تعیین کلید در این الگوریتم صرفه‌جویی در هزینه مقایسات است. زیرا با این کار لازم نیست تمام قسمت‌های یک قانون برایبک کلید مورد بررسی قرار گیرد. در حجم داده‌های بالا تعداد قوانین افزایش خواهد یافت و همچنین هر قانونی حاوی چندین کلید است. بررسی این حجم کلید برای داده‌هایی با حجم بالا بسیار زمان‌گیر است.

تانگ (2014)، روش جهت تصحیح ناسازگاری‌ها با استفاده از قوانین معرفی شده است. در این روش ابتدا پایگاه داده مورد بررسی انتخاب می‌شود. قوانین موجود در آن منبع با استفاده از الگوریتم‌های موجود استخراج می‌گردد و پس از آن الگوریتم اطمینان[1] هر قانون را محاسبه می‌کند.

2- 4- وابستگی تابعی

روش‌های زیادی از وابستگی تابعی برای تشخیص خطا استفاده کرده‌اند اما به ازای ناسازگاری‌های شناخته شده، تصحیح‌های متنوعی را می‌توان ارائه نمود که باید از میان آنها یک تصحیح را به عنوان تصحیح نهایی ارئه کرد. برای انتخاب تصحیح بهینه دو پارامتر هزینه و تنوع مطرح است. الگوریتم ارائه شده توسط هی و همکاران (2014) برای نخستین بار هر دو پارامتر را به عنوان هدف خود قرار داده است. برای محاسبه میزان تنوع از تابع فاصله استفاده می‌شود که میزان عدم شباعت رکوردها را محاسبه می‌کند. برای محاسبه هزینه نیز میزان تغییرات در منبع اولیه با منبع در دسترس را در نظر می‌گیرد. در حجم بالای داادهها محاسبه توابه بسیار زمان‌گیر است.

2- 5- شبکه عصبی

شبکه‌های مصنوعی عصبی در شناسایی خطا و ایزوله کردن آن (FDI) سیستم‌ها چه بصورت سیستم‌های غیرخطی دینامیک و چه دسته‌بندی کننده‌های الگو برای تولید نشانه‌ها و ارزیابی آن‌ها بکار می‌روند (فرانک[2]، 1994). این موارد برای پروسه‌های پیچیده‌ای که از آنها مدلی در دسترس

[1] - Confidence
[2] Frank

نیست یا به سختی بدست می‌آید بطور خاص استفاده می‌شوند. مدل ANN نیازی به مدل دقیق است از سیستم ندارد ولی نیازمند اطلاعات برای آموزش شبکه است. یک مشکل اصلی کنار آمدن با دینامیک‌ها می‌باشد هنگامی که نشانه‌ها تولید می‌شوند (آیزرمن[1]، 1997).

در FDI هدف، بهترین‌تخمین به معنای دست یافتن به رفتار دینامیک پروسه بوسیله یادگیری گرایش مقادیر نهایی و فیلتر کردن نویز است. به عبارت دیگر شبکه‌های عصبی باید ساختار مینیمال داشته باشند تا امکان محاسبات سریع در موارد تولید نشانه‌ها به شکل خطای پیش بین را بدهد.

یکی از الگوریتم‌های پرکاربرد شبکه عصبی برای تشخیص خطا، الگوریتم پس انتشار (BP) است. شبکه عصبی معمولا در یک روش نظارتی با الگوریتم عمومی آموزش می‌بینند،که با عنوان الگوریتم پس انتشار شناخته می‌شود. اطلاعات مربوط به خطا فیلتر شده و به سیستم بازمی‌گردد تا ارتباط بین لایه‌ها را تنظیم کنو، در نتیجه عملکرد بهبود می‌یابد. فرایند خطای BP ترکیبی از دو لایه مختلف شبکه به نام‌های مسیر مستقیم (پیش‌خور) و مسیر بازگشتی (رو به عقب) است. در مسیر رو به جلو یک الگوی فعال برای حساسیت گره‌های شبکه به لایه انتشار می‌یابد. در نهایت، مجموعه‌ای از خروجی‌ها به عنوان پاسخ حقیقی شبکه تولید می‌شود. پاسخ حقیقی شبکه از مقدار پاسخ مطلوب کم می‌شود تا سیگنال خطا تولید شود. این سیگنال خطا سپس به صورت بازگشتی در طول شبکه و در خلاف جهت ارتباطات وزنی پخش می‌شود. وزن‌ها به صورت تنظیم شده‌اند تا پاسخ حقیقی شبکه را به پاسخ مطلوب نزدیکتر کنند (یوسف[2]، 2003).

مطالعات زیادی از شبکه عصبی برای تشخیص خطا استفاده کرده‌اند. به عنوان مثال مرادی و همکاران (1389) در پژوهشی با عنوان تشخیص جریان هجومی از جریان خطا در ترانسفورماتورهای قدرت با استفاده از الگوریتم جستجوی گرانشی، با استفاده از شبکه عصبی داده‌های مربوط را با استفاده از شبکه عصبی پردازش کردند و سپس با استفاده از الگوریتم‌های جستجوی گرانشی آموزش خود را بهبود دادند.

2- 6- درخت تصمیم

یک درخت تصمیم معمولا تشکیل شده است از ریشه[3]، شاخه‌ها[4]، گره‌ها[5] و برگ‌ها[6]. درخت تصمیم هم به صورت مشابه از گره‌ها که با دایره نشان داده می‌شوند و شاخه‌ها که با پاره خط نشان داده می‌شوند، تشکیل شده اند. درخت تصمیم[7] را به منظور شادگی در رسم معمولا از چپ به راست یا از بالا به پایین رسم می‌کنند به طوری که ریشه در بالا قرار بگیرد. گره اول را ریشه می‌گویند. نقیصه

[1] Isermann et al
[2] Youssef
[3] Root
[4] Beach
[5] Node
[6] Leaf
[7] Decision Trees

درخت تصمیم در واقع این است که در موردی که تمام خصوصیت‌ها کمی باشند، درخت تصمیم تخمین‌های غیر دقیقی را از جواب نهایی به دست می‌دهند.

در پژوهش تنگ (2004) هدف شناسایی ویژگی‌های مشکوک به خطا و کلاس آن ویژگی‌ها و اصلاح آن به وسیله الگوریتم polishing که دارای دو فاز پیش‌بینی و تنظیم مقادیر می‌باشد است. در فاز پیش‌بینی از میان الگوریتم‌های طبقه‌بنید، یک روش انتخاب شده و به ازای هر ویژگی 10 دسته بر روی داده‌ها ایجاد می‌شود. در این مرحله مقدار ویژگی مورد نظر برای هر رکورد به ازای هر 10 طبقه‌بند پیش‌بینی می‌گردد. در صورتی که مقدار اصلی آن رکورد با مقدار پیش‌بینی شده تناقض داشته باشد، مقادیر حاصل از پیش‌بینی برای آن رکورد جهت تصحیح در فاز بعدی ذخیره می‌شود.

3- جمع‌بندی

صحت داده‌هایی که در اختیار سازمان‌ها قرار می‌گیرد. از ابعاد مختلف حائز اهمیت فراوان است. به گونه‌ای که داده‌های ناصحیح ممکن است سازمان را متحمل هزینه‌های جبران ناپذیری کند. روش‌های داده‌کاوی از جمله روش‌های کارا به منظور شناسایی و تصحیح خطا به عنوان ابزار قدرتمندی مورد استفاده قرار می‌گیرد. از جمله این روش‌ها می‌توان به شبکه‌های عصبی، درخت تصمیم، روش قوانین، هستان‌شناسی و وابستگی تابعی اشاره کرد. نتایج بررسی در این پژوهش نشان داد که روش‌های فوق قابلیت استفاده در شناسایی و تصحیح خطا را دارند.

4- منابع

امینی، ملیحه. نقیب‌زاده محمود. محتشمی سید هاشم.(1390). استنتاج توزیع شده بر روی آنتولوژی‌ها و قوانین در منطقorder-sorted. سیستم‌های هوشمند در مهندسی برق، سال دوم، شماره سوم، پاییز.

عطاییان، م. و دانشپور، ن. (1395) تشخیص خودکار خطا در پایگاه داده، مبتنی بر خوشه‌بندی و نزدیک‌ترین همسایگی. نشریه مهندسی برق و کامپیوتر ایران، 14 (4) 349-356.

مرادی، ع.؛ عبادیان، م. و دریاباری، م. (1389) تشخیص جریان هجومی از جریان خطا در ترانسفورماتورهای درت با استفاده از الگوریتم جست‌وجوی گرانشی، مجله علمی-پژوهشی سیستم‌های هوشمند در مهندسی برق، 1 (1): 43-58.

Bellatreche, L., Khouri, S., Boukhari, I., & Bouchakri, R. (2012, May). Using ontologies and requirements for constructing and optimizing data warehouses. In *MIPRO, 2012 Proceedings of the 35th International Convention* (pp. 1568-1573). IEEE.

Beskales, G., Ilyas, I. F., Golab, L., & Galiullin, A. (2014). Sampling from repairs of conditional functional dependency violations. *The VLDB Journal*, *23*(1), 103-128.

Bouchra, A., Wakrime, A. A., Sekkaki, A., & Larbi, K. (2016, May). Automating Data Warehouse Design Using Ontology. In *International Conference on Electrical and Information Technologies*.

Bramer, M. (2007). *Principles of data mining* (Vol. 180). London.: Springer.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Brüggemann, S. (2008). Rule mining for automatic ontology based data cleaning. *Progress in WWW Research and Development*, 522-527.

Chávez, J. V., & Li, X. (2011, October). Ontology based ETL process for creation of ontological data warehouse. In *Electrical Engineering Computing Science and Automatic Control (CCE), 2011 8th International Conference on* (pp. 1-6). IEEE.

De Lemos, R., Giese, H., Müller, H. A., Shaw, M., Andersson, J., Litoiu, M., ... & Weyns, D. (2013). Software engineering for self-adaptive systems: A second research roadmap. In *Software Engineering for Self-Adaptive Systems II* (pp. 1-32). Springer, Berlin, Heidelberg.

Fan, W. (2008, June). Dependencies revisited for improving data quality. In *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 159-170). ACM.

Frank, P. M. (1994). Enhancement of robustness in observer-based fault detection. *International Journal of control*, *59*(4), 955-981.

Gulić, M. (2013, May). Transformation of OWL ontology sources into data warehouse. In *Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on* (pp. 1143-1148). IEEE.

He, C., Tan, Z., Chen, Q., Sha, C., Wang, Z., & Wang, W. (2014, April). Repair diversification for functional dependency violations. In *International Conference on Database Systems for Advanced Applications* (pp. 468-482). Springer, Cham.

Isermann, R., Ernst, S., & Nelles, O. (1997). Identification with dynamic neural networks-architectures, comparisons, applications. *IFAC Proceedings Volumes*, *30*(11), 947-972.

J. Hipp, U. Guntzer, and U. Grimmer, "Data quality mining-making a virute of necessity," in Proc. 6th Int. SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD'01, pp. 52-57, Santa Barbara, California, USA, May, 2001.
Jain, R., & Abraham, A. (2004). A comparative study of fuzzy classification methods on breast cancer data. *Australasian Physical & Engineering Science in Medicine*, *27*(4), 213-218.

Jiang, L., Cai, H., & Xu, B. (2010, November). A domain ontology approach in the ETL process of data warehousing. In *e-Business Engineering (ICEBE), 2010 IEEE 7th International Conference on* (pp. 30-35). IEEE.

John, G. H., & Langley, P. (1995, August). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp. 338-345). Morgan Kaufmann Publishers Inc.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

Kwak, J. A., & Yong, H. S. (2010). Ontology Matching based on hypernym, hyponym, holonym, and meronym sets in word net. *International Journal of Web & Semantic Technology*, *1*(2), 1-14.

Martin, A., Maladhy, D., & Venkatesan, V. P. (2011). A framework for business intelligence application using ontological classification. *arXiv preprint arXiv:1109.1088*.

Mikroyannidis, A., & Theodoulidis, B. (2010). Ontology management and evolution for business intelligence. *International Journal of Information Management*, *30*(6), 559-566.

N. Tang, "Big data cleaning," in Proc. 16th Int. Conf.in Web Technologies and Applications, pp. 13-24, Changsha, China, 5-7 Sept. 2014

Rahman, Md Geaur, and Md Zahidul Islam. "Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques." *Knowledge-Based Systems* 53 (2013): 51-65.

Teng, C. M. (2004). Polishing blemishes: Issues in data correction. *IEEE Intelligent Systems*, *19*(2), 34-39.

W. Ahmed Malik and A. Unwin, "Automated error detection using association rules," Intelligent Data Analysis, vol. 15, no. 5, pp. 749- 761, Sept. 2011.

Williams, P. H., Margules, C. R., & Hilbert, D. W. (2002). Data requirements and data sources for biodiversity priority area selection. *Journal of biosciences*, *27*(4), 327-338.

Wongthongtham, P., & Abu-Salih, B. (2015, July). Ontology and trust based data warehouse in new generation of business intelligence: State-of-the-art, challenges, and opportunities. In *Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on* (pp. 476-483). IEEE.

Yakout, M., Berti-Équille, L., & Elmagarmid, A. K. (2013, June). Don't be SCAREd: use SCalable Automatic REpairing with maximal likelihood and bounded changes. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 553-564). ACM.

Youssef, O. A. (2003). A wavelet-based technique for discrimination between faults and magnetizing inrush currents in transformers. *IEEE Transactions on Power Delivery*, *18*(1), 170-176.