dataacademy. ir

ارائه روشی برای تشخیص رکوردهای مشابه در پایگاه داده‌های بزرگ با استفاده از شبکه عصبی

بهراد دهقانی[۱]، محمدرضا حسنی آهنگر[۲]
[۱] دانشجوی کارشناسی ارشد دانشگاه جامع امام حسین (ع) ، dehghani@ihu.ac.ir
[۲] استادیار گروه کامپیوتر دانشگاه جامع امام حسین (ع) ، mrhassani@iust.ac.ir

چکیده - پایگاه داده‌های بزرگ معمولاً شامل رکوردهایی است که به یک موجودیت یکسان اشاره دارند. استفاده از اختصار، سامانه-های متنوع سازمانی، تجمیع منابع داده مختلف دلایلی هستند که منجر به رکوردهای تکراری در پایگاه داده می‌گردند. این مساله بخاطر اینکه تاثیر مخربی بر آمار و گزارشات و نتایج داده کاوی از پایگاه داده‌ها می‌گذارد بسیار مهم است. در این مقاله، راه‌حل جدیدی برای تشخیص رکوردهای مشابه در پایگاه داده‌های بزرگ با استفاده از شبکه عصبی *MLP* ارائه شده است. نتایج ارزیابی نشان می‌دهد در صورتی که شبکه عصبی با داده‌های مناسب آموزش ببیند عملکرد خوبی نسبت به روش‌های قبلی دارد. این روش بصورت ماژولی جهت تشخیص تکرارها، برای سیستم‌های برخط بزرگ مانند موتورهای جستجو قابل استفاده است

کلید واژه- هوش مصنوعی، تشخیص تکرار، شباهت رکوردها، شبکه‌های عصبی

۱-مقدمه

تشخیص شباهت بین توصیف‌های مختلف از شی یکسان، یک مسأله مشترک در فناوری اطلاعات است. پایگاه داده‌های بزرگ، شامل چندین رکورد برای هر محصول، شخص، رخداد یا اشیا دیگر که در اختصار متفاوتند و شامل غلط املایی، اشتباهات چاپی، نمایش‌های غیر معمول و غیر واحد از موجودیت منطقی یکسان هستند. موتورهای جستجو، نرم‌افزار پاک‌کننده داده، رابط‌های کاربری هوشمند، تلاش برای پیشگویی ورودی کاربر، تنها بخشی از ناحیه‌هایی است که این مسأله اتفاق می‌افتد. این مساله مانع از عملکرد درست الگوریتم‌های داده کاوی، جمع‌آوری داده‌های آماری و غیره می‌شود.

راه‌حل‌های مختلفی برای شناسایی ردیف‌های تکراری با سطح مختلف خودکارسازی وجود دارد. در بیشتر موارد، رکوردهای یکسان جهت تایید در اختیار کاربر قرار می‌گیرند. یکی از مهم-ترین مسائل این راه‌حل‌ها، تنظیم پارامترهای جستجو است که اساساً برای کارکرد صحیح الگوریتم هستند و بر روی کارآیی و کیفیت نتایج، تأثیرگذارند، تنظیم سیستم‌ها، ممکن است گیج کننده باشد و نیاز به تلاش زیاد خبره دارد. تلاش برای استفاده از شبکه عصبی برای کالیبره کردن خودکار موتور جستجوی تشخیص تکرار در این مقاله بحث شده است.

در [۱] توضیحی از موتور جستجوی تشخیص تکرار بیان شده است. موتور رکوردی از داده را دریافت و در رکوردهای پایگاه داده برای شبیه/همانند آن را جستجو می‌کند. برای مقایسه عامل مشترک، سطح شباهت بین دو رشته در بین دو رکورد (مجموعه رشته‌هایی با معنای مختلف) ارزیابی می‌شود. موتور جستجو تشخیص تکراری برای تعیین سطح شباهت بین دو رکورد از داده‌ها توسعه داده شده، که شکل (۱) آنرا نشان داده است.

آکادمی داده
dataacademy.ir

dataacademy. ir

شکل ۱: مراحل موتور جستجوی تشخیص تکرار[1]

چندین کلید مختلف برای تعیین کلاسترهای ردیف‌های مشابه به کار می‌رود تا بیشترین شباهت بررسی شود و نتایج کلاسترهای مختلف به روش مشابه برای توصیف آن ترکیب می‌شود[۱و۵]. موتور الگوریتم ویرایش فاصله [۲] را به عنوان متدی برای تعیین سطح شباهت بین فیلدهای رشته به‌کار می‌گیرد. که بر اساس فاصله لون اشتاین (Levenshtein) است[۳]، که بصورت کمترین تعداد درج‌ها، حذف‌ها یا جانشینی‌ها لازم برای انتقال یک رشته به دیگری و توسعه نیدلمن (Needleman) و وانچ (Wunsch) که امکان دنباله‌های تکراری از کاراکترهای نامتناسب یا گپ‌ها را در دو رشته هم‌طراز می‌دهد تعریف می شود. الگوریتم مقایسه توسط رینهارد اشنایدر (Reinhard Schneider) برای مقایسه توالی پروتئین‌ها پیاده‌سازی شده است[۶] و برای مقایسه دو رشته اسکی در موتور مورد بحث استفاده شده است. با افزودن چندین ویژگی و تخصیص وزن‌ها و معرفی جداول مشابه (شباهت آوایی، کاراکترها، نزدیکی مکان به هم در صفحه کلید و غیره) گسترش یافته است. موتور با مفسر وابسته به زبان از نمادهای خاص (ć š ä ö ß ž) و...)، اختصارات، آدرس ها، فیلدها و عبارات چندکلمه ای (مانند Delphi automobile system درمقابل automobile systems Delphi) کار می‌کند. موتور به صورت مجموعه‌ای از ماژول‌ها عملیاتی شده و به زبان جاوا نوشته شده است.

در گام بعدی، ضرایب شباهت برای هر فلیدی محاسبه شده است، برای تولید ضریب کلی که سطح شباهت دو رکورد از داده-ها را نشان می دهد ترکیب شده‌اند. واضح است که فیلدها اولویت یکسان ندارند. مقدار یکسانی از اطلاعات را نمی‌آورند. برای نمونه فیلدی که شامل نام یک شخص است حاوی اطلاعات مفیدتری از فیلدی است که تاریخ تولد یا نام شهر است. وقتی نتایج فیلدی جهت محاسبه ضریب تاثیر شباهت نهایی ترکیب شدند، بایستی بر اساس ارزش اطلاعاتی فیلدها وزن‌دهی شوند.

در روش استاندارد برای شباهت رکورد [۷]، رکوردهای پایگاه داده که برای تکرارها جستجو می شوند با مجموعه ای از خصوصیات نمایش داده می شوند. رای برای یک جفت کاندید با y بیان می شود، که y میتواند مجموعه مقادیر {۱و۱-} بگیرد. مقدار ۱، یعنی رکوردهای جفتی به موجودیت یکسان بر می‌گردد و مقدار ۱- ، یعنی به موجودیت‌های متفاوت بر می‌گردد. فرض کنیم x=(x1,x2,xn) برداری از امتیازهای شباهتی را بین خصوصیات رکوردها در جفت کاندید را نشان دهد. احتمال توزیع y از x بصورت زیر تعریف می‌شود:

$$f(x)=\lambda_0+\sum_{i=1}^{1n}\lambda_i x_i \qquad (1)$$

f(x) تابع تشخیص است، *λi* برای *0<i<1* پارامترهای مدل هستند. با این پارامترها و بردار شباهت خصوصیت *x*، رای جفت کاندید مثبت است (تطابق) اگر *f*(x)>0 و در غیر اینصورت منفی (عدم تطابق) می‌باشد. پارامترها معمولا با بیشترین احتمال یا با احتمال شرطی تنظیم هستند[۸].

با استفاده از معادله(۱) تعریف تطابق تکرار بدین شکل منتقل می شود:

$$f(x)=\lambda_0+\sum_{i=1}^{1n}\lambda_i x_i > \lambda'0 \qquad (2)$$

که بصورت ضریب شباهت کلی تفسیر می شود (بر اساس شباهت های بین خصوصیات محاسبه می شود) برای تعریف دو رکورد تکراری باید از حد مشخصی بزرگتر باشد. در راه‌حل کنونی، سطح اولویت‌ها به هریک از خصوصیات که در فرآیند بررسی هستند تخصیص می‌یابد. لذا پارامترهای *λi* در(۲) بصورت (۳) تعریف می شوند.

(3)

$$\lambda_i=\frac{weight_i}{\sum_{i=1}^{n} weight_i} \quad \text{که} \quad weight_i=max(pririty)-priority_i$$

تخصیص مستقیم اولویت‌ها یک کار خبرگی است. فردی که مقداردهی این ضرایب را انجام می‌دهد علاوه بر آشنایی با محتوای کلی داده باید به جزئیات ساختار پایگاه داده نیز بخوبی آشنا باشد. لذا راه حل آماده شده نیاز به دانش خبره در ناحیه بحث شده دارد. بنا به این دلایل راه حل جدید پیشنهاد و ارائه شده است که از شبکه عصبی برای تولید خودکار تابع تشخیص استفاده می‌کند.

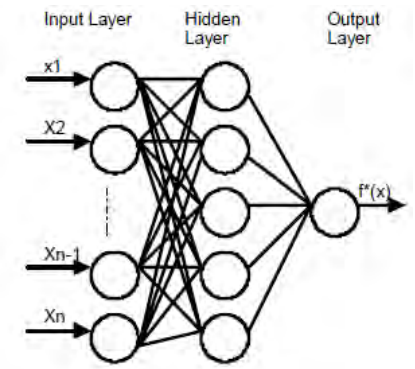
آکادمی داده
dataacademy.ir

dataacademy. ir

در بخش دوم راه‌حل پیشنهادی با جزئیات بیشتر بحث شده است. در بخش سوم، با استفاده از شبکه عصبی MLP پیاده-سازی انجام شده و نتیجه و مزایای به کارگیری بیان شده است. بخش چهارم نتایج و طرح‌هایی برای بهبود و توسعه مدل در آینده را شامل می‌شود.

۲- شبکه عصبی MLP

تصمیم گیری بین راه‌حل های مرسوم و عصبی همیشه بصورت کامل شفاف نیست. برای هر کدام از راه حل‌های مرسوم و عصبی مسائلی وجود دارد که راه حل‌های مناسبی ایجاد می نمایند. انتخاب آن به منابع در دسترس و اهداف نهایی پروژه بستگی دارد. در هنگام تصمیم‌گیری برای مساله‌ای که با روش محاسبات عصبی حل خواهد شد سه شرط اصلی باید اعمال شود.



شکل ۲: ورودی و خروجی یک الگوی آموزشی در شبکه عصبی MLP

۱. راه حل مساله صریحاً توسط یک الگوریتم، یکسری معادلات (برای نمونه، ارائه مدل فیزیکی) یا مجموعه‌ای از قواعد نمی‌تواند حل شود یا بسیار مشکل است.

۲. شواهدی مبنی بر یک نگاشت ورودی-خروجی بین مجموعه ای از متغیرهای ورودی X و داده های خروجی متناظر y وجود دارد بطوریکه: y=f(x). اگرچه فرم f() ناشناخته است.در این راه حل مشخص کردن f() یک کار خبرگی است.

۳. تعداد زیادی داده باید در دسترس باشد مثلاً نمونه‌های زیادی برای آموزش شبکه.

۲-۱ ساختار شبکه عصبی MLP و آموزش آن

تعدادی ورودی‌ها در شبکه چندلایه توسط تعداد ویژگی‌ها یا پارامترهای ورودی در دسترس برای مساله مطرح شده مشخص می شود. بنابراین شبکه عصبی محقق شده، ۱۴ ورودی و یک خروجی دارد.(مطابق با اندازه داده های ورودی و خروجی به ترتیب در یک الگوی آموزش).

شبکه عصبی یک لایه مخفی دارد. تابع فعال ساز آن یکی از توابع فعال ساز نمونه است[۹]، تابع sigmoid باینری که دامنه (0,1) دارد.

مقداردهی اولیه وزن ها شامل مجموعه مقادیر تصادفی کوچک (-0.1,0.1) است. علاوه بر وزن ها، مقادیر بایاس نرون ها در محاسبه خروجی شبکه لحاظ شده است. نرخ یادگیری برای هر الگو ۰.۳ است. انتخاب نرخ یادگیری تاثیر مهمی در کارآیی شبکه دارد.

دو عامل اصلی مشخص هستند؛ جمع آوری، آماده سازی و تحلیل داده‌های آموزشی و طراحی، آموزش و تست شبکه عصبی، دو آزمایش برای تصمیم‌گیری کار تحقیق انجام شده است. آزمایش ها ابتدا برای مدل سیستم خبره و سپس برای مدل شبکه عصبی انجام شده‌اند. هر دو آزمایش کار یکسانی (تعیین درصد شباهت رکورد بر اساس شباهت فیلدهایش) روی مجموعه داده یکسان، و متغیرهای وابسته یکسان بجز اولویت فیلدی که در مدل سیستم خبره معین شده ولی در شبکه عصبی قابل شناخت نبود را استفاده کرده اند. داده هایی که در روش سیستم خبره بدرستی بدست آمده بودند در کلاس های ورودی صحیحی برای مجموعه آموزشی شبکه عصبی (بردارهای ورودی) خوشه بندی شدند و نتایج بعنوان بردارهای خروجی صحیح استفاده شدند. آن داده ها برای آموزش و هم برای تست الگوها جدا شدند.

ادبیات تحقیقی محاسبات عصبی غنی از مقالاتی است که ساختارهای شبکه مختلفی را ارائه کرده اند اما اغلب آنها از ساختارهای پرسپترون چندلایه و تابع پایه شعاعی استفاده کردند[۱۰]. از این‌رو شبکه عصبی جهت نتایج عملی با پرسپترون چندلایه با یک لایه مخفی و یادگیری با ناظر پس انتشار بصورت استاندارد است. توپولوژی شبکه به تغذیه رو به جلو محدود شده است: مانند حلقه باز که عموما ارتباطاتش از لایه ورودی به لایه مخفی و از لایه مخفی به لایه خروجی است.

آکادمی داده
dataacademy.ir

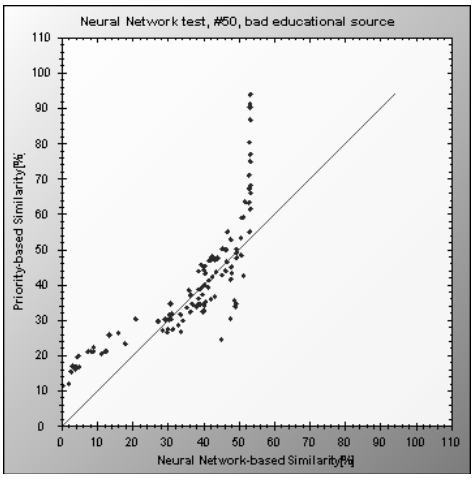
dataacademy. ir

لایه مخفی رکورد کردن(یا فراهم کردن نمایش برای) ورودی ها را یاد می گیرد[۱۱].

برنامه‌هایی که از روش شبکه عصبی استفاده می‌کنند شامل نگاشتی از یک مجموعه ورودی به مجموعه مشخصی از خروجی-های هدف هستند. هدف بدست آوردن شبکه ی است که بتواند پاسخ های قابل قبولی علاوه بر موارد آموزش دیده به ورودی هایی که متشابه اند بدهد. آموزش شامل تغذیه روبه جلو از الگوی ورودی، محاسبه خطا و تنظیم وزن ها و بایاس است. در برخی موراد کند است، اما یک شبکه آموزش دیده خروجی هایش را سریعتر تولید می کند.

۳- پیاده‌سازی شبکه عصبی MLP

تخمین قابل قبول از اینکه چه مقدار داده برای آموزش درست شبکه عصبی لازم است مهم می‌باشد. اگر مقدار بسیار کمی از داده جمع‌آوری شده است، محدوده کاملی از نسبتی که شبکه عصبی باید آموزش ببیند را ممکن است پوشش ندهد. آزمایش به داده‌های کافی نیاز دارد تا به فرمی از نگاشت اشاره کند که بدرستی تمامی محدوده از فضای ورودی از الگوها را مشخص نماید. اگر داده های آموزشی برای ناحیه‌ای از فضای ورودی از برخی داده های تستی وجود نداشته باشد، تولیده کننده مناسبی برای این الگوها نمی‌تواند باشد. در چنین حالتی شبکه عصبی پاسخ می‌دهد ولی دقیق نخواهد بود. آزمایش های مختلفی با پارامترهای متفاوت در محیط شبیه‌سازی مطلب برای رسیدن به پاسخ دقیق انجام شده است که شکل (۳) آزمایشی با الگوهای آموزشی نامناسب را نشان می‌دهد.



شکل ۳:آزمایش با الگوهای آموزشی نامناسب

در تست نشان داده شده در شکل (۳)، شبکه عصبی با شباهت هایی بین ۳۰٪-۶۰٪ آموزش دیده شده است. اختلاف بین شباهت محاسبه شده با استفاده از روش اولویت محور موجود و دیگری که با استفاده از شبکه عصبی محاسبه شده بسیار زیاد بود. خصوصاً در ناحیه‌هایی خارج از این محدوده (اگر هیچ اختلافی بین بین دو روش وجود ندارد نقاط باید روی خط کشیده شده باشند). لذا برای یقین از اینکه شبکه عصبی بخوبی آموزش دیده تا محدوده کاملی از مقادیر را تشخصی و پاسخ دهد، ۲۷۰.۰۰۰ الگوی آموزشی بالقوه (بردارهای ورودی-خروجی) در چندین کلاس که با استفاده از روش اولویت محور جمع شده بودند پخش شدند. هر کلاسی شامل نمونه هایی است که ورودی در محدوده ۱۰ درصد و مطابق به ورودی هایش دارد. برای نمونه شباهت رکوردی از ۰-۹٪ از کلاس یک، ۱۰-۱۹٪ از کلاس دو و غیره. دو جفت متعادل از مجموعه ها یکپارچه شده اند و در همه کلاس ها بصورت یکسانی نمایش داده شده‌اند. اولین جفت شامل ۵۰ الگوی یادگیری و ۲۵۰ تستی می باشد. دومین ۱۰۰ یادگیری و ۲۵۰ الگوی تستی. برای شروع ساخت شبکه عصبی ۱۴ متغیر از الگوی یادگیری بعنوان ورودی مشخص شدند. تعداد ۱۴ بخاطر اینکه هر رکورد شامل ۱۴ فیلد است(نام، نام خانوادگی، شهر و غیره). در نمونه زیر مشخص هستند.

1 47.50 2 0.00 3 31.67 5 100.00 6 100.00 7 -3.00 8 25.00 9 0.00 11 0.00 12 0.00 13 0.00 14 0.00

در توالی مقادیر داده صحیح و اعشاری که ترکیب زوج های اعدادند یکی در میان هستند. مقدار صحیح مربوط به شماره بعدی از فیلد داده ی است که در بردار آموزشی شرکت دارد (بدین معنی است که مقدار صحیح شماره نرون ورودی پذیرنده را دارد). مقدار اعشار، شباهت رشته ی برای آن شماره فیلد را دارد. اگر فیلدی وجود دارد که که در ارزیابی شرکت ندارد. زوجش خالی است. یعنی آن داده برای این فیلد خاص در پایگاه داده وجود ندارد. لذا مکانش در فرم بردار ورودی صفر است. نمونه بالا به ورودی زیر از شبکه عصبی تبدیل شده است.

47.50 0.00 31.67 0.00 100.00 100.00 -3.00 25.00 0.00 0.00 0.00 0.00 0.00 0.00

هر یک از این مقادیر ورودی برای یک نرون از لایه ورودی شبکه است. بدلیل نیاز به کم کردن تاثیر شدت، در طول مرحله پیش پردازش مقادیر ورودی نرمالایز شده اند.$X_i \in (-1,1)$. فرم نهایی از الگوی ورودی بدین شکل است:

آکادمی داده
dataacademy.ir

dataacademy. ir

0.475 0.00 0.3167 0.00 1.0 1.0 -0.03 0.25 0.00 0.00 0.00 0.00 0.00 0.00

خروجی متناظر از شبکه عصبی شامل فقط یک مقدار است- مقدار تخمینی از شباهت جفت رکوردها، تابع تشخیص $f(x)$ در (۲) تعریف شده است.

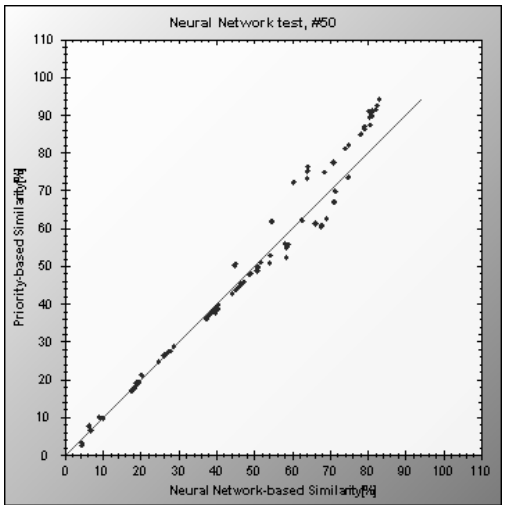
0.475 0.00 0.3167 0.00 0.1 0.1 -0.03 0.25 0.00 0.00 0.00 0.00 0.00 0.00 – بردار ورودی

0.2447 – بردار خروجی

داده های تست نیز به همین شکل است و بطور یکنواخت از کلاس هایی که بالا بحث شد به همان طریق داده های آموزشی انتخاب شدند. این دو مجموعه داده تصادفی هستند و دیگری را نمی پوشانند. هر بردار ورودی از نمونه های تست به شبکه داده شده اند که بردار خروجی را تولید می کند(یک مقدار). سپس نتایج دو روش کنار هم قرار گرفتند تا خلاصه شوند و بصورت گرافیکی در فرم قابل قبولی نمایش داده شدند.

چندین قاعده مطرح برای تعداد نرون ها در لایه مخفی وجود دارد. اگر تعداد کمی نرون باشد underfitting را در پی خواهد داشت و اگر تعداد زیادی نرون باشد overfitting و افزایش زمان برای آموزش شبکه را به همراه خواهد داشت. در مورد ما تعداد مجموعه‌ای است که از روش سعی و خطا استفاده می‌کند. در نقطه شروع تعداد نرون های مخفی از دو برابر اندازه لایه ورودی کمتر انتخاب شدند. چندین تست انجام شد تا اینکه ترکیب مناسبی از تعداد الگوهای آموزشی و دیگر پارامترها پیدا شد.

تست با ۵۰ الگوی آموزشی، که بطور یکنواخت بین ۰٪ . . ۱۰۰٪ توزیع بودند، نمایش خوبی در قسمت پایین(بالای ۵۰٪ شباهت) داشت، اما اختلاف بیشتری در قسمت بالایی از نمودار (شکل ۴) داشت. تست با شرایط جدول (۱) انجام شده است.



شکل ۴: تست با ۵۰ الگوی آموزشی

تست با ۱۰۰ الگوی آموزشی ، با توزیع یکنواخت بین ۰٪ .. ۱۰۰٪ ، توزیع بسیار خوبی در مقایسه با موارد اولویت محور در بین همه مقادیر تست نشان داد(شکل ۵).

تست با شرایط جدول (۲) انجام شده است.

جدول ۱: تست با ۵۰ الگوی آموزشی

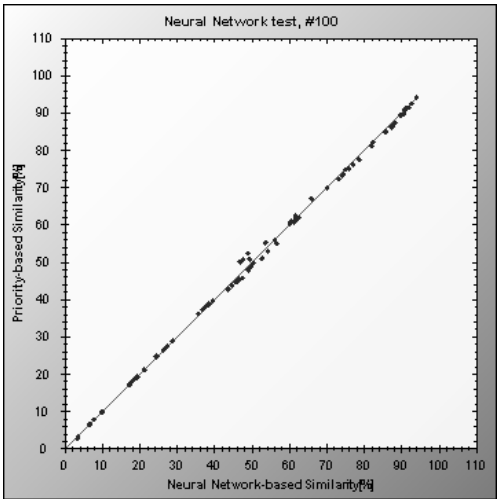
تعداد الگوهای آموزشی	۵۰
تعداد الگوهای تست	۲۵۰
تعداد نرون های ورودی	۱۴
تعداد لایه مخفی	۱
تعداد نرون ها در لایه مخفی	۲۰
تعداد نرون خروجی	۱
تعداد الگوهای آموزشی	۵۰
نرخ یادگیری	۰.۳
ضریب ممنتم	۰.۰
خطای بدست آمده	۰.۰۰۴۶۹۵۵۱۶۰۶۵۵۹۷۶۶

جدول ۲: تست با ۱۰۰ الگوی آزمایشی

تعداد الگوهای آموزشی	۱۰۰
تعداد الگوهای تست	۲۵۰
تعداد نرون های ورودی	۱۴
تعداد لایه مخفی	۱
تعداد نرون ها در لایه مخفی	۱۰
تعداد نرون خروجی	۱
تعداد الگوهای آموزشی	۵۰
نرخ یادگیری	۰.۳
ضریب ممنتم	۰.۰
خطای بدست آمده	۰.۰۰۴۳۲۶۱۲۲۴۶۵۸۱۰۳

آکادمی داده
dataacademy.ir

dataacademy. ir



شکل ۵: تست با ۱۰۰ الگوی آموزشی

۴- نتیجه گیری و کارهای آتی

تشخیص و شناسایی رکوردهای تکراری در پایگاه داده‌های بزرگ بسیار اهمیت دارد. در این مقاله راه‌حل جدیدی با استفاده از شبکه عصبی MLP برای تشخیص رکوردهای تکراری استفاده ارائه شد. نتایج ارزیابی نشان می‌دهد در صورتی که شبکه عصبی با داده‌های مناسب آموزش ببیند عملکرد خوبی نسبت به روش-های قبلی دارد. تنظیم پارامترهای مختلف (تعداد نرون‌های پنهان، تعداد اپک های آموزش، تعداد و توزیع الگوهای داده آموزشی و غیره) چالش اصلی است. برخی از این پارامترها بر اصل سعی و خطا انتخاب می‌شوند. شبکه ایجاد شده همیشه راه حل بهینه نمی دهد، اما با بهبود تنظیمات درست میتواند پاسخ های بهتری تولید نماید.

یکی از جهت گیری‌ها برای بهبود آتی راه حل بحث شده، آزمایش‌های بیشتر با شبکه عصبی است. شبکه عصبی می‌تواند از شیوه های موثرتری برای مقداردهی اولیه وزنها، بهبود در لایه پنهان یا توابع فعالساز متناسب با داده‌ها، تکنیک‌های مختلف برای پویایی در تعیین مقادیر نرخ یادگیری و ضریب ممنتم در طول آموزش استفاده نماید.

مراجع

[1] P. Paskalev, A. Antonov, "Intelligent application for duplication detection". In proceedings of the *International Conference CompSysTech 2006*, IIIA.27.1-IIIA.27.8, June. 2006.

[2] Gusfield "*Algorithms on strings, trees and sequences*". Cambridge Univ. Press, NY, 1997.

[3] Bilenko M., Mooney R. "Adaptive duplicate detection using learnable string similarity measures" In Proceedings of the *Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*(KDD-2003), Washington DC, pp.39-48, August, 2003

[4] Needleman S. B. and Wunsch C. D. "A general method applicable to the search for similarities in the amino acid sequences of two proteins." *Journal of Molecular Biology*, 48:443–453, 1970

[5] Hernandez M. A. and Stolfo S. J. "The merge/purge problem for large databases." In Proceedings of the *1995 ACM SIGMOD*, pages 127–138, San Jose, CA, May 1995.

[6] Sander C. and Schneider R., "Database of homology-derived protein structures and the structural meaning of sequence alignment,"*Proteins*, vol. 9, no. 1, pp. 56--58, 1991.

[7] I. Fellegi and A. Sunter. "A theory for record linkage*"*. *Journal of the American Statistical Association*, 64:1183-1210, 1969

[8] Parag and Pedro Domingos. "Multi-relational record linkage". KDD-2004 *Workshop on Multi-Relational Data Mining* (pp. 31-48), 2004

[9] Laurene Fausett, "*Fundamentals of neural networks. architectures,algorithms and applications*". Prentice Hall, 1994, ISBN: 0133341860

[10] M. Young, "*The Technical Writer's Handbook*". Mill Valley, CA: University Science, 1989.

[11] Lionel Tarassenko, "*A Guide to Neural Computing Applications*". Butterworth-Heinemann, 1998, ISBN: 0340705892

[12] Barbara D. Klein and Donald F. Rossin "Data errors in neural network and linear regression models: An experimental comparison" Data Quality, vol5, n1, 1999– www.dataquality.com/999KR.htm

[13] Leslie Smith, "*An Introduction to Neural Networks*", http://www.cs.stir.ac.uk/~lss/NNIntro/InvSlides.html

آکادمی داده
dataacademy.ir