Iran J. Insur. Res., 7(1): 27-38, Winter 2018, Serial #23

Iranian Journal of Insurance Research
(IJIR)
Homepage: https://ijir.irc.ac.ir/?lang=en

Insurance Research Center (IRC)

ORIGINAL RESEARCH PAPER

Apply data mining techniques to recruitment agents on the insurance industry

F. Kazemi[1,*], H. Iranmanesh[2]

[1] *Department of Industrial Management - Operations Research, Faculty of Management and Accounting, Qazvin Azad University, Qazvin, Iran*
[2] *Department of Industrial Engineering, School of Industrial Engineering, Qazvin Azad University, Qazvin, Iran*

ARTICLE INFO

Article History
Received: 20 February 2017
Revised: 09 April 2017
Accepted: 03 January 2018

Keywords
Data mining; intelligent decision making systems.

ABSTRACT

Iran's insurance industry has experienced significant growth in the last decade, and a high share of transactions in this industry is allocated to agents; Therefore, the process of choosing a new representative is very important. This research focuses on the characteristics of the data warehouse and data mining methods suitable for choosing an agent in the insurance industry. Regarding the example presented in this research, three data mining methods of discriminant analysis, decision trees, and artificial neural networks are evaluated for predicting service duration, premium sales, and agent continuity index. The results show that work experience, job position, age, marital status, previous job, annual income from the previous job, and sold insurance policies are very important in determining the duration of activity of new agents, sold insurance policies, and renewal of issued insurance policies. The main goal of this article is to design and develop an intelligent decision support system, and in other words, an intelligent representative selection system for insurance companies, so that the managers of this industry can choose quality representatives with its help.

***Corresponding Author:**
Email: *kazemi.fereidoon@gmail.com*
DOI: 10.22056/ijir.2018.01.03

نشریه علمی پژوهشنامه بیمه

سایت نشریه: https://ijir.irc.ac.ir/?lang=fa

مقاله علمی

استفاده از روش‌های داده‌کاوی برای جذب نمایندگان صدور در صنعت بیمه

فریدون کاظمی[۱،*]، حامد ایران‌منش[۲]

[۱]گروه مدیریت صنعتی- گرایش تحقیق در عملیات، دانشکدهٔ مدیریت و حسابداری، دانشگاه آزاد قزوین، قزوین، ایران

[۲]گروه مهندسی صنایع، دانشکدهٔ مهندسی صنایع، دانشگاه آزاد قزوین، قزوین، ایران

اطلاعات مقاله

تاریخ دریافت: ۰۲ اسفند ۱۳۹۵

تاریخ داوری: ۲۰ فروردین ۱۳۹۶

تاریخ پذیرش: ۱۳ دی ۱۳۹۶

کلمات کلیدی

داده‌کاوی

سیستم‌های تصمیم‌گیری هوشمند.

چکیده:

صنعت بیمهٔ ایران رشد چشمگیری را در دههٔ گذشته تجربه کرده است و سهم بالایی از معاملات این صنعت به نمایندگان اختصاص دارد؛ بنابراین، فرایند انتخاب نمایندهٔ جدید اهمیت بسیار بالایی دارد. این پژوهش بر روی مشخصه‌های انبار داده و روش‌های داده‌کاوی مناسب برای انتخاب نماینده در صنعت بیمه تمرکز دارد. در خصوص مثال ارائه‌شده در این پژوهش سه روش داده‌کاوی تحلیل تفکیک‌کننده، درخت های تصمیم، و شبکه‌های عصبی مصنوعی برای پیش‌بینی طول مدت‌زمان خدمت، میزان فروش حق‌بیمه، و شاخص تداوم نمایندگان مورد ارزیابی قرار می‌گیرد. نتایج نشان می‌دهد که تجربهٔ شغلی، موقعیت شغلی، سن، وضعیت تأهل، شغل قبلی، درآمد سالیانه از محل شغل قبلی، و بیمه‌نامه‌های فروخته‌شده اهمیت به‌سزایی در تعیین مدت‌زمان فعالیت نمایندگان جدید، بیمه‌نامه‌های فروخته‌شده و تمدید بیمه‌نامه‌های صادرشده دارند. هدف اصلی این مقاله، طراحی و توسعهٔ یک سیستم پشتیبانی تصمیم‌گیری هوشمند، و به بیان دیگر یک سیستم هوشمند انتخاب نماینده برای شرکت‌های بیمه بوده تا با کمک آن مدیران این صنعت بتوانند نمایندگان کیفی انتخاب کنند.

***نویسنده مسئول:**

ایمیل: *kazemi.fereidoon@gmail.com*

DOI: 10.22056/ijir.2018.01.03

مقدمه

امروزه، داده‌ها یک دارایی استراتژیک و بسیار مهم برای افراد و شرکتها محسوب می‌شوند. این امر نتیجهٔ فناوریهای جدید است که «مخزن داده‌ها»[۱] یکی از این موارد است. شرکتها در حوزه‌های مختلف مانند بانکداری، بیمه، خرده‌فروشی، و مراقبتهای پزشکی با مهار داده‌های انبوه عملیاتی سعی در درک بهتر و پیشرفت کسب‌وکار خود دارند (Brockett et al., 1997; Delmater and Hancock, 2001).

داده‌ها از واحدهای مختلف یک شرکت جمع‌آوری و در یک انبار مرکزی به نام مخزن داده‌ها ذخیره می‌شود. تحلیلگران با استفاده از مخزن داده‌ها اطلاعات بازار تجارت را استخراج می‌کنند تا قادر به تصمیم‌گیری بهتر باشند (Cho, Wüthrich Cho and Wüthrich, 2002 and Zhang, 1999;). به این فرایند پشتیبانی تصمیم‌گیری تعاملی یا در اصطلاح OLAP[۲] می‌گویند. می‌توان به جای OLAP از واژهٔ پردازش سریع اطلاعات چندبُعدی و یا به عبارت بهتر از «فن‌آوری تحلیل داده‌ها» نیز استفاده کرد. OLAP با تمرکز بر روی حجم زیادی از داده‌ها، فرایند تصمیم‌گیری را تسهیل می‌کند.

به صورت سنتی، گزارشهای معمول پس از ساعت کاری جمع‌آوری و چاپ می‌شوند. در این فرایند مسئولیت بررسی روزانهٔ فعالیتها و معاملات وجود ندارد. علاوه‌براین، ساختار یک بانک دادهٔ سنتی به صورت گسترده برای عملیاتهای روزانه مناسب است. در رویه‌های گزارش‌دهی و استعلامهای ویژه و پیچیده از مدیریت ارشد، مانند جدولهای پیشایندی، معمولاً نیاز به بازیابی و ادغام سوابق تاریخی معاملات و تراکنشها وجود دارد. این فرایند با عملیات روزانهٔ عادی، به‌ویژه با کسب‌وکارهایی مانند بانک، شرکتهای بیمه، و خطوط هوایی که شدیداً عملیاتی هستند، تداخل دارد؛ بنابراین، اخذ آنلاین استعلامات ویژه نیز شدنی نیست. در شیوهٔ سنتی وجود تأخیر تا پس از ساعت کاری اجتناب‌ناپذیر است؛ که این امر مانعی بزرگ بر سر راه مدیران برای اتخاذ تصمیمهای آنی بوده و از این رو عملکرد اجرایی شرکت را کاهش می‌دهد. به منظور تسهیل این مشکلات، برخی از آمارهای معمول در ساختار و معماری مخزن داده‌ها مانند خلاصهٔ فروش در مناطق مختلف، دسته‌بندیهای محصولات مختلف و نقاط تقاضای مختلف، هنگام ثبت هر تراکنش به‌روزرسانی می‌شود. این آمارهای معمول نشان‌دهندهٔ دیدگاه چندبُعدی از مخزن داده‌ها بوده و اجازه می‌دهد تحلیل آنلاین با کمترین وقفه برای عملیاتهای معاملاتی روزانه انجام شود. این فرایند که نخستین گام به سوی OLAP است را می‌توان حرکتی جدید در مدیریت اطلاعات دانست.

تحلیل ساختار چندبُعدی مخزن داده‌ها چالشهای جدیدی را ایجاد می‌کند. علاوه بر ارائهٔ خلاصه‌های آماری از جنبه‌های مختلف داده، روشهای داده‌کاوی نیز بایستی در رابطهٔ با OLAP اعمال شود. درحقیقت این ترکیب، راه‌حلی یکپارچه برای کسب‌وکار ایجاد می‌کند. با توجه به تحولات آتی در این حوزه، انتظار می‌رود که داده‌کاوی اهمیت بسیار بیشتری به خود بگیرد. با این حال، روشهای داده‌کاوی موجود، در حال حاضر، ارتباط چندانی با ابزارهای OLAP ندارند. علاوه‌براین، بیشتر روشهای داده‌کاوی مانند C4.5، CN2 و ITRules، نرم‌افزارهایی مستقل با تعداد زیادی پارامتر کنترلی هستند؛ بنابراین، کار با این ابزار برای کاربران معمولی ساده نبوده و نیاز به کارشناس و متخصص دارد. از همهٔ موارد فوق مهم‌تر اینکه، این روشها عموماً برای بانکهای اطلاعات سنتی به کار برده می‌شوند. یکی از چالشهای پیش روی این پژوهش، چگونگی تطبیق روشهای داده‌کاوی موجود با ساختار چندبُعدی مخزن داده‌هاست. چالش دیگر این است که چگونه خطوط ارتباطی کارآمد برای مدیران ارشد اجرایی ایجاد شود تا بتوانند از چنین روشهای پیچیده‌ای بهره جویند.

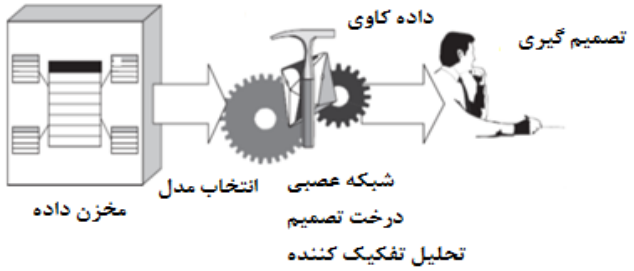
هدف اصلی این مقاله، طراحی و توسعهٔ یک سیستم پشتیبانی تصمیم‌گیری هوشمند، یا یک سیستم هوشمند انتخاب نماینده برای شرکتهای بیمه[۳]، با استفاده از روشهای داده‌کاوی برای یک مخزن دادهٔ چندبُعدی است. در این پژوهش از این سیستم برای مفهوم انتخاب نماینده در صنعت بیمه استفاده می‌شود. در شکل ۱ چارچوب سیستم تصمیم‌گیری هوشمند نمایش داده شده است.

---

۱. Data Warehousing

2. Online Analytical Processing

۳. Intelligent Agent Selection Assistant for Insurance (IASAI)



شکل ۱: چارچوبی برای سیستم پشتیبانی تصمیم‌گیری هوشمند

بیمه نقش مهمی در جامعه و بازار تجارت بازی می‌کند و گسترهٔ وسیعی از موضوعات را در شاخه‌های مختلف مانند اموال، اشخاص، و مسئولیت در بر می‌گیرد. صنعت بیمه ایران در دههٔ گذشته رشد چشمگیری را تجربه کرده است. تعداد کارکنان یک شرکت بیمه یا یک شرکت مالی مرتبط به بیمه در ایران، بین چند صد نفر تا چند هزار نفر متغیر است. در حالت کلی، یکی از موضوعات مهم در این صنعت، حجم بالای معاملات برای نمایندگان بیمه است. از این رو، به رویهٔ انتخاب نمایندگان جدید به‌عنوان یک فرایند استخدامی دائم و منظم نگاه می‌شود.

این پژوهش بر روی ویژگیهای مخزن داده‌ها و روشهای مناسب داده‌کاوی برای انتخاب نماینده در صنعت بیمه تمرکز دارد. در اینجا، سه روش معمول داده‌کاوی (تحلیل تفکیک‌کننده[۱]، درختهای تصمیم[۲]، و شبکه‌های عصبی مصنوعی[۳]) مدنظر قرار گرفته؛ و بررسی می‌شود که چگونه می‌توان این سه ابزار را می‌توان در OLAP برای یک مخزن داده‌ها چندبُعدی گنجاند.

سیستم‌های پردازش تحلیلی آنلاین

سیستمهای اطلاعاتی تحلیلی در مقایسهٔ با سیستمهای عملیاتی، سیستمهای هستند که امکان تحلیل داده‌های انبوه حاصل از سیستمهای عملیاتی را برای تمامی سطوح کاربران فراهم می‌کنند. این در حالی است که سیستمهای عملیاتی در سازمانهای بزرگی مانند شرکتهای بیمه به صورت روزانه پردازشهای اطلاعاتی فراوانی را به انجام رسانده و به تولید اطلاعات گوناگون می‌پردازند. بانکهای اطلاعاتی این سازمانها با داده‌های فراوان حاصل از تراکنشهای مالی، اداری، حسابداری و... روبه‌رو می‌شوند. اطلاعات پایهٔ سیستمها همانند اطلاعات کاربران و سطوح دسترسی آنها معمولاً با تغییرات روزانه مواجه نیستند اما اطلاعات عملیاتی نظیر عملیات تجاری، خریدوفروش محصولات و ...می‌توانند حتی به طور لحظه‌ای تغییر کنند.

تحلیل و پردازش درست و دقیق اطلاعات عملیاتی می‌تواند در تولید نتایج آماری در جهت تصمیم‌گیریهای کلان مدیریتی مؤثر بوده و به مدیران کمک کند تا تصمیمات بهینه‌ای برای موفقیت سازمان خود بگیرند. برای تحلیل و پردازش این اطلاعات و تسهیل و سرعت‌بخشیدن به عملیات گزارش‌گیری و پرس‌وجوهای متنوع به جای تحلیل مستقیم داده‌ها از درون سیستمهای عملیاتی، از سیستمها و پایگاه‌داده‌های تحلیلی استفاده می‌شود که خارج از حوزهٔ سیستمهای عملیاتی قرار داشته و سرعت بسیار بالایی دارند. پایگاه‌داده‌های تحلیلی نسخه‌های متنوعی از داده‌های تراکنشی را به صورت اختصاصی برای پرس‌وجوها و گزارش‌گیری، سازمان‌دهی می‌کنند. به این ترتیب کاربرانی مانند مدیران سازمان که خارج از سیستمهای عملیاتی قرار دارند می‌توانند گزارشها و پرس‌وجوهای مورد نظر خود را تهیه کنند. پایگاه‌داده‌های تحلیلیOLAP از منابع داده‌ای متفاوت یک سازمان و یا حتی چندین سازمان و ارگان وابسته به هم تهیه می‌شود. این پایگاه‌داده بستر مناسبی را فراهم می‌آورد که داده‌های بایگانی‌شده در سیستمهای عملیاتی و مستقل از هم سازمان، به صورت مجتمع، خلاصه شده، و یکپارچه و سازمان‌یافته درآمده و برای استخراج مناسب اطلاعات در دسترس مدیران باشند (Blanco et al., 2015).

سیستمهای OLAP برای ارائهٔ پاسخهای سریع به سؤالات و جستجوهای تحلیلی روی داده‌های «چندبُعدی» طراحی شده‌اند. به طور معمول اگر بخواهیم مشابه همین پرس‌وجوهای تحلیلی را روی سیستمهای اطلاعاتی عادیOLTP[۱] اجرا کنیم ممکن است نتایج در زمانی طولانی و

---

۱. Discriminant Analysis (DA)

۲. Decision Trees (DTs)

۳. Artificial Neural Networks (ANNs)

غیرکاربردی بازگردانده شود در حالی که استفاده از OLAP تضمین می‌کند که اطلاعات و گزارشهای تحلیلی با زمان پاسخ مناسبی به کاربر تحویل داده شود (Houaria et al., 2016). کاربردهای معمول OLAP عبارت‌اند از: گزارشهای تجاری فروش، بازاریابی، گزارشهای مالی، و مواردی از این دست. این سیستمها داده‌های خود را به‌نحوی خاص نگهداری می‌کنند که از نظر سرعت در برخورد با داده‌های چندبُعدی بهتر از سیستمهای OLTP عمل می‌کنند و از این رو به آنها بانکهای اطلاعاتی سلسله‌مراتبی[۲] هم گفته می‌شود.

تعاریف عمدهٔ فناوری OLAP

مخزن داده: انبار داده‌ها یک پایگاه اطلاعاتی بزرگ است که به جمع‌آوری، یکپارچه‌سازی و ذخیرهٔ اطلاعات متنوع یک سازمان، با هدف تولید گزارشهای چندجانبه و دقیق می‌پردازد (Bellatrechea et al., 2015).

مرکز داده‌ها (DM[۳]): انبار داده‌ها حجم عظیمی از اطلاعات را در واحدهای منطقی کوچکتری به نام مرکز داده‌ها نگهداری می‌کند. مرکز داده‌ها نمونه‌های کوچکی از انبار داده‌ها بوده و همانند آنها حاوی نسخه‌هایی ثابت از داده‌هایی هستند که در موارد خاص استفاده می‌شوند. مرکز داده‌ها می‌تواند وابسته یا مستقل از هم باشند. هر مرکز داده، داده‌ها و ابعاد[۴] خاص خود را دارد که می‌تواند با بقیه به اشتراک بگذارد.

داده‌کاوی[۵]: ابزارهای داده‌کاوی به دنبال طرحها و گروه‌بندیهایی در داده‌ها می‌شود که ممکن است از دید ما پنهان مانده باشد. در داده‌کاوی این ابزار است که استفاده‌کننده را هدایت می‌کند. ابزار فرض می‌کند که شما خود نیز دقیقاً نمی‌دانید که چه می‌خواهید. اولین گام داده‌کاوی هدف‌دار، انتخاب مجموعهٔ داده‌ها برای تحلیل است. داده‌ها می‌تواند از انبار داده‌ها و یا بانکهای اطلاعاتی عملیاتی استخراج شود. داده‌ها پس از جمع‌آوری و حذف موارد تکراری در قالبهای یکسان تجمیع و پاکسازی می‌شوند. سپس با استفاده از منابع مناسب، اطلاعات ناقص اصلاح و کدگذاری شده و با ساختار جدیدی آماده می‌شوند. به این ترتیب داده‌ها برای داده‌کاوی آماده است.

نحوهٔ عملکرد سیستم‌های OLAP

سیستم OLAP به صورت مرتب از داده‌های منابع اطلاعاتی مختلف نسخه‌های خلاصه‌شده برداشته و آنها را در مکعبهای داده‌ای مرتب می‌کند. پرس‌وجوهای کاربران می‌تواند روی این مکعب اجرا شود. روشهای مختلف طراحی انبار داده‌ها امکان پردازشهای بهینه را بر روی مقادیر زیادی از داده‌ها فراهم می‌آورند. پرس‌وجوهای پیچیده روی سیستمهای OLAP به زمانی حدود تنها ۰/۱ درصد از زمان اجرای جستجوهای مشابه روی سیستمهای OLTP احتیاج دارند.

انواع ویژه‌ای از الگوهای پایگاه‌داده‌ها به نام ستاره‌ای[۶] یا دانه برفی[۷] برای طراحی انبار داده چندبُعدی وجود دارد. در این حالت، پایگاه‌داده‌ها از یک جدول مرکزی و جدولهای چندبُعدی تشکیل شده است روابط بین آنها کاملاً مشخص است. برای دستیابی به سرعت بالا و زمان کوتاه، سیستمهای OLAP جدولهای اطلاعاتی خود را در آرایشهای ستاره‌ای یا دانه برفی مرتب می‌کنند. ساختار OLAP مثل یک مکعب روبیک است که می‌توانید آن را در جهتهای مختلف بچرخانید تا بتوانید تحلیلهایی از دیدگاه‌های مختلف را بررسی کنید. نحوهٔ عملکرد این سیستمها به این صورت است که معیارهای اساسی تحلیل به‌عنوان ابعاد مختلف یک مکعب در نظر گرفته شده و این مکعب در انبار داده‌ها شکل می‌گیرد. این ابعاد می‌توانند در سطوح مختلف و به صورت سلسله‌مراتبی نیز وجود داشته باشند.

انواع مختلف سیستم‌های OLAP

در حال حاضر انواع مختلف OLAP وجود دارد (Dehne et al., 2015):

---

۱. Normal Online Analytical Processing

۲. Hierarchical

۳. Data Mart

۴. Dimension

۵. Data Mining

۶. Star Schema in Data Warehouse

۷. Snowflake Schema in Data Warehouse

سیستمهای OLAP چندبُعدی[۱]،

فریدون کاظمی و حامد ایران‌منش

سیستمهای OLAP رابطه‌ای[۲]،
سیستمهای OLAP ترکیبی[۳].

روش شناسی پژوهش

چارچوب و روش‌شناسی

یک مخزن داده را می‌توان به‌عنوان یک انبار آنلاین از داده‌های شرکت معرفی کرد، که به منظور پشتیبانی از فرایند تصمیم‌گیری مورد استفاده قرار می‌گیرد (Inmon, 2002). OLAP روشی است برای تحلیل داده‌هایی که در مخزن داده‌های چندبُعدی ذخیره می‌شوند (Kimball, 1996). این روش نخستین بار به‌وسیلهٔ کاد[۴] و همکاران (۱۹۹۳) با هدف ساده‌سازی سیستمهای پشتیبانی تصمیم معرفی شد. همچنین این روش، سنگ بنای اولیه برای سیستمهای اطلاعات اجرایی مدرن نیز است.

در رابطهٔ با روش OLAP چالشهای جدیدی پیش روی شرکتها قرار می‌گیرد؛ برای مثال چگونه می‌توان با ترکیب مفاهیم داده‌کاوی، روشهای OLAP را ارتقا داد؛ و چگونه می‌توان با ایجاد ارتباط مناسب بین این روشها از فرایند تصمیم‌گیری پشتیبانی کرد. ساختار سیستمهای پشتیبانی تصمیم بایستی به اندازهٔ کافی انعطاف‌پذیر بوده که بتوان از آنها برای روشهای OLAP که برای مخزن داده‌های در حال رشد به کار برده می‌شوند، استفاده کرد. یکی از اهداف این پژوهش پُرکردن شکاف تحقیقاتی موجود با استفاده از ادغام روشهای داده‌کاوی در تکنولوژی OLAP است. شکل ۱، چارچوب IASAI (سیستم پیشنهادی برای پشتیبانی تصمیم‌گیری هوشمند) را برای انتخاب نمایندگان جدید نشان می‌دهد. برای کاهش وقفه در عملیاتهای معاملاتی روزانه، DMهایی از مخزن داده‌ها استخراج شده و در OLAP مورد استفاده قرار می‌گیرد. درحقیقت، هر مخزن داده می‌تواند از چند ساختار کوچکتر به نام مرکز داده تشکیل شود. هر مرکز داده نیز متشکل از داده‌های مرتبط به هم از یک مخزن داده هستند.

مرکز داده بخشی از مخزن داده است که شامل تمام اطلاعات مربوط به یک کاربرد خاص می‌شود. برای مثال، به بانک اطلاعاتی که تنها شامل اطلاعات یک دپارتمان خاص است مرکز داده گفته می‌شود، در حالی که مخزن داده‌ها منبع همهٔ داده‌ها مربوط به همهٔ دپارتمانهاست (Inmon, 2002). توالی به‌روزرسانی برای یک مرکز داده به کاربرد آن بستگی خواهد داشت که این کاربرد توسط مدیریت ارشد تعیین می‌شود. به صورت سنتی، شرکتهای بیمهٔ نمایندگانی را استخدام می‌کنند که به محض اینکه در فروش بیمه‌نامه با مشکل مواجه می‌شوند، استعفا می‌دهند. به علت وجود هزینه‌های بالای آموزش برای نمایندگان و دیگر هزینه‌های سربار اداری مانند چاپ کارتهای ویزیت، استعفای نمایندگان به هیچ عنوان مقرون به صرفه نیست. علاوه‌براین، به علت عدم پیگیری، بیمه‌نامه‌های فروخته‌شده توسط این دست نمایندگان شانسی کمی برای تمدیدشدن دارند.

در این پژوهش، سه کاربرد داده‌کاوی مدنظر است. ابتدا، نیاز به پیش‌بینی مدت اشتغال نمایندگان جدید وجود دارد. دوم، نیاز به پیش‌بینی حق‌بیمهٔ ورودی به شرکت توسط این نمایندگان است؛ و سوم، پیش‌بینی تداوم و تمدید حق‌بیمه‌هایی که توسط نمایندگان جدید صادر می‌شود. این سه کاربرد در روند استخدام نمایندهٔ جدید بسیار مفید هستند.

پیش از بررسی چگونگی اجرای این کارکردها، می‌توان توضیح کاملی در خصوص روشهای تحلیل تفکیک‌کننده (DA)، شبکه‌های عصبی مصنوعی (ANN) و درختهای تصمیم (DT) را در مقالات کاس[۵] (۱۹۸۰)، برایمن[۶] و همکاران (۱۹۸۴)، ریملهارت[۷] و همکاران (۱۹۸۶)، کویین‌لن[۸] (۱۹۸۷)، وربس[۱] (۱۹۹۴)، ریپلی[۲] (۱۹۹۴)، دودا و همکاران[۳] (۲۰۰۱) و کونن[۴] (۲۰۰۱) می‌توان جستجو کرد.

---

۱. Multi dimensional OLAP (MOLAP)
۲. Relational OLAP (ROLAP)
۳. Hybrid OLAP
۴. Codd
4. Kass
5. Breiman
6. Rumelhart
7. Quinlan

گردآوری داده‌ها

در این پژوهش از داده‌های مربوط به نمایندگان شرکت بیمهٔ ملت استفاده شده است. برای انتخاب نمایندهٔ بیمه در این شرکت از سیستم پشتیبانی تصمیم‌گیری توسعه‌یافته استفاده می‌شود. مجموعه‌دادهٔ مورد استفاده برای این پژوهش که از مخزن داده‌های شرکت استخراج شده است، شامل اطلاعات ثبت‌شده از بیش از ۱۴۵۰ نمایندهٔ حقیقی و حقوقی در مدت ۱۰ سال ۲۰۰۵- ۲۰۱۵ است.

ویژگیهایی که در این پژوهش مورد نظر هستند عبارت‌اند از: جنسیت، تاریخ تولد، ملیت، میزان تحصیلات، تعداد افراد تحت تکفل، موقعیت شغلی، تجربهٔ کاری، تجربهٔ مدیریتی، مجموع حق‌بیمهٔ فروخته‌شده، شایستگی برای فروش محصولات خاص، تاریخ شروع فعالیت، تاریخ خاتمهٔ فعالیت، زمینهٔ شغلی سابق و درآمد سالیانهٔ قبلی. در اینجا، داده‌های تجمعی مانند مجموع حق‌بیمهٔ فروخته‌شده، به منظور تسریع روند پرس‌وجو در مخزن داده‌ها ذخیره می‌شوند. در بخش بعد، روشهای داده‌کاوی بر روی مجموعه‌ای از داده‌های بیمه‌ای اعمال می‌شوند.

آزمایش

این آزمایش با استفاده از سه روش داده‌کاوی تحلیل تفکیک‌کننده (DA)، شبکه‌های عصبی مصنوعی (ANN) و درختهای تصمیم (DT) برای پیش‌بینی طول مدت‌زمان خدمت، میزان فروش حق‌بیمه و شاخص تداوم نمایندگان بیمه اجرا می‌شود. انتظار می‌رود این سه روش به مدیران برای انتخاب نمایندگان بیمه با بهره‌وری بالا کمک کند. در طول این دورهٔ ۱۰ ساله، ۱۴۵۸ رکورد از نمایندگان ثبت و برای اهداف آموزشی این پژوهش مورد ارزیابی قرار گرفته‌اند. این در حالی است که ۵۰۰ رکورد نیز به صورت تصادفی، برای اعتبارسنجی عملکرد انتخاب شده است. در این پژوهش، از مدل پیش‌خور با تک لایهٔ مخفی[۵] برای شبکه‌های عصبی استفاده می‌شود؛ درخت تصمیم با روش C4.5 ساخته می‌شود؛ و برای تحلیل تفکیک‌کننده نیز از تابع تفکیک خطی فیشر[۶] استفاده می‌شود. در روشهای درخت تصمیم و شبکه‌های عصبی مصنوعی، ویژگیهای اسمی مانند جنسیت، به صورت متغیرهای ساختگی مانند جنسیت ۱ (مرد) و جنسیت ۲ (زن) بیان می‌شوند.

پیش‌بینی طول مدت‌زمان خدمت

طول مدت‌زمان خدمت را می‌توان در دو رده قرار داد: کمتر از یک‌سال (ردهٔ ۱) و مساوی یا بیشتر از یک سال (ردهٔ ۲). به‌وضوح مشخص است که شرکتهای بیمه تمایل به استخدام نمایندگانی دارند که بیشتر از یک‌سال کار خواهند کرد. دقت پیش‌بینی‌های روش تحلیل تفکیک‌کننده برای طول مدت‌زمان خدمت و برای داده‌های آموزشی و آزمون به‌ترتیب برابر با ٪۶۰/۰۷ و ٪۵۷/۲۰ است؛ دقت پیش‌بینی‌های روش شبکه‌های عصبی نیز به‌ترتیب برابر با ٪۶۹/۲۸ و ٪۶۱/۴ است؛ و دقت پیش‌بینی‌های درخت تصمیم نیز به‌ترتیب برابر با ٪۷۵/۳۰ و ٪۵۹/۴۰ است. در جدول ۱، ماتریس رده‌بندی برای داده‌های آزمون همان‌گونه که توسط این سه روش پیش‌بینی شده‌اند، ارائه شده است.

جدول ۱: ماتریس آشفتگی برای پیش‌بینی طول مدت‌زمان خدمت (به درصد)

روش داده‌کاوی	مدت‌زمان خدمت	پیش‌بینی		کلی
		کمتر از یک سال	بیشتر از یک سال	
شبکهٔ عصبی مصنوعی				
واقعی	کمتر از یک سال	۳۷/۶	۱۶/۸	۵۴/۴
	بیشتر از یک سال	۲۱/۸	۲۳/۸	۴۵/۶
کلی		۵۹/۴	۴۰/۶	۱۰۰/۰
تحلیل تفکیک‌کننده				
واقعی	کمتر از یک سال	۲۸/۸	۲۵/۶	۵۴/۴
	بیشتر از یک سال	۱۷/۲	۲۸/۴	۴۵/۶
کلی		۴۶/۰	۵۴/۰	۱۰۰/۰

[8]. Werbos
[9]. Ripley
[10]. Duda
[11]. Kohnen
[۵]. A Single Hidden Layer Feed-Forward Model
[۶]. Fisher's Linear Discriminant Function

درخت تصمیم				
واقعی	کمتر از یک سال	۴۲/۶	۱۱/۸	۵۴/۴
	بیشتر از یک سال	۲۸/۸	۱۶/۸	۴۵/۶
کلی		۷۱/۴	۲۸/۶	۱۰۰/۰

با توجه به جدول، مشخص است که شبکه‌های عصبی مصنوعی بالاترین دقت، ٪۶۱/۴= ۲۳/۸ + ۳۷/۶، را در مقایسهٔ با دو روش دیگر دارند. برای تفسیر این درصدها (برای مثال در روش شبکه‌های عصبی) اگر پیش‌بینی شود که یک متقاضی نمایندگی، در ردهٔ ۱ قرار دارد، آنگاه ٪۶۳/۳ = (۲۱/۸ + ۳۷/۶) ÷ ۳۷/۶ احتمال وجود دارد که متقاضی واقعاً در آن رده وجود داشته باشد. اگر پیش‌بینی شود که این متقاضی در ردهٔ ۲ قرار دارد، آنگاه ٪۵۸/۶ احتمال وجود دارد که متقاضی واقعاً در ردهٔ ۲ وجود داشته باشد. با مقایسهٔ نتایج در جدول ۱ مشاهده می‌شود که مدل شبکه‌های عصبی، بالاترین دقت مشروط را برای هر دو ردهٔ ۱ و ۲ در پیش‌بینی طول مدت‌زمان خدمت نماینده داراست.

موقعیت شغلی و بیمه‌نامهٔ فروخته‌شده، دو مورد از مهم‌ترین پارامترها برای تعیین طول مدت‌زمان خدمت هستند. ضرایب استاندارد در روش تحلیل تفکیک‌کننده برای پیش‌بینی طول مدت‌زمان خدمت در جدول ۲ نشان داده شده است. این ضرایب، اهمیت نسبی متغیرهای مستقل را اندازه‌گیری می‌کنند. ضرایب موقعیت شغلی و وضعیت تأهل بالاترین اهمیت را دارند. این امر نشان می‌دهد که این دو مورد اساسی‌ترین عاملها در پیش‌بینی طول مدت‌زمان خدمت نماینده هستند. درنتیجه، بررسی هم‌زمان هر سه روش داده‌کاوی نشان می‌دهد که مؤثرترین متغیرها برای تعیین مدت‌زمان احتمالی خدمت نمایندهٔ جدیدالورود عبارت‌اند از: ۱. موقعیت شغلی، ۲. بیمه‌نامهٔ فروخته‌شده و ۳. وضعیت تأهل.

جدول ۲: ضرایب تابع تفکیک متعارف استاندارد (طول مدت‌زمان خدمت)

متغیر	ضریب استاندارد
موقعیت	۰/۷۵۴
جنسیت	۰/۳۳۵
سن	۰/۰۴۰
وضعیت تأهل	
مجرد	۰/۶۷۱
متأهل	۰/۳۱۲
مطلقه	۰/۳۳۷
تعداد افراد تحت تکلف	۰/۱۲۷
سطح علمی	-۰/۰۹۸
ماهیت شغل قبلی	
دفتری/فنی	-۰/۱۷۹
بخش فروش/بخش خدمات	۰/۰۰۳
تخصصی	-۰/۰۴۳
مدیریتی	-۰/۱۰۵
وضعیت استخدام	-۰/۰۵۴
درآمد ماهیانهٔ گذشته	-۰/۴۶۶
سابقهٔ کاری	-۰/۱۰۵
سابقهٔ مدیریتی	-۰/۱۸۲
پورتفوی بیمه‌ای آورده شده	-۰/۳۶۸

پیش‌بینی فروش حق‌بیمه

برای حق‌بیمهٔ فروخته‌شده، یا مجموع کل بیمه‌نامه فروخته شده توسط یک عامل بیمه، دو رده در نظر گرفته می‌شود: کمتر از ۲۲۵۴۳۲۲ (ردهٔ ۱) و مساوی یا بیشتر از ۲۲۵۴۳۲۲ (ردهٔ ۲)؛ بنابراین، نمایندگانی که در ردهٔ ۲ قرار می‌گیرند عملکرد بهتری دارند. تحلیل تفکیک‌کننده دارای دقت پیش‌بینی ٪۵۷/۴۴ و ٪۶۴/۶۰ (٪۱۸/۶ + ۴۶ مطابق جدول ۳) به‌ترتیب برای داده‌های آموزشی و داده‌های آزمون است. دقت پیش‌بینی برای شبکه‌های عصبی مصنوعی ٪۶۷/۲۵ و ٪۵۳/۲۰ (٪۳۱/۶ + ٪۲۱/۶ مطابق جدول ۳) و برای درخت تصمیم ٪۷۷/۱۰ و ٪۵۳

(٪۳۲/۴ +٪۲۰/۶ مطابق جدول ۳) است. درخت تصمیم، بالاترین دقت کلی را برای مجموعه داده‌های آزمون، بالاترین دقت مشروط برای پیش‌بینی ردهٔ ۱ (٪۵۷/۸) و بالاترین دقت مشروط برای پیش‌بینی ردهٔ ۲ (٪۷۲) را تولید می‌کند.

جدول ۳: ماتریس آشفتگی برای پیش‌بینی فروش حق‌بیمه (به درصد)

روش داده‌کاوی	رده‌بندی	پیش‌بینی		کلی
		ردهٔ ۱	ردهٔ ۲	
شبکهٔ عصبی مصنوعی				
واقعی	ردهٔ ۱	۳۱/۶	۸/۸	۴۰/۴
	ردهٔ ۲	۳۸	۲۱/۶	۵۹/۶
کلی		۶۹/۶	۳۰/۴	۱۰۰
تحلیل تفکیک‌کننده				
واقعی	ردهٔ ۱	۱۸/۶	۲۱/۸	۴۰/۴
	ردهٔ ۲	۱۳/۶	۴۶	۵۹/۶
کلی		۳۲/۲	۶۷/۸	۱۰۰
درخت تصمیم				
واقعی	ردهٔ ۱	۳۲/۴	۸	۴۰/۴
	ردهٔ ۲	۳۹	۲۰/۶	۵۹/۶
کلی		۷۱/۴	۲۸/۶	۱۰۰

همان‌گونه که از سطوح بالای درخت تصمیم قابل مشاهده است، تجربهٔ کاری و موقعیت شغلی از جملهٔ مهم‌ترین عاملها در پیش‌بینی حق‌بیمهٔ فروخته‌شده توسط یک نمایندهٔ بیمه است.

با توجه به جدول ۴، بررسی ضرایب تفکیک استاندارد نشان می‌دهد که پارامترهایی مانند شغل قبلی، درآمد سالانهٔ قبلی، و سن متقاضی (که ضرایبشان بالاتر از ۰/۴ است)، سهم به‌سزایی در پیش‌بینی حق‌بیمهٔ فروخته‌شده دارند.

با توجه به تحلیل روشهای DT و DA مشخص شد که تأثیرگذارترین عاملها در پیش‌بینی حق‌بیمهٔ فروخته‌شده عبارت‌اند از: ۱. شغل قبلی، ۲. درآمد سالیانهٔ قبلی، ۳. سن متقاضی، ۴. سابقهٔ شغلی و ۵. موقعیت شغلی.

جدول ۴: ضرایب تابع تفکیک متعارف استاندارد (حق‌بیمهٔ فروخته‌شده)

متغیر	ضریب استاندارد
موقعیت	-۰/۲۷۰
جنسیت	-۰/۱۴۲
سن	-۰/۵۱۰
وضعیت تأهل	
مجرد	-۰/۱۳۹
متأهل	۰/۰۴۱
مطلقه	۰/۰۲۶
تعداد افراد تحت تکلف	۰/۰۹۸
سطح علمی	-۰/۰۸۱
ماهیت شغل قبلی	
دفتری/فنی	۰/۵۴۳
بخش فروش/بخش خدمات	۰/۵۵۱
تخصصی	۰/۱۳۵
مدیریتی	۰/۶۳۳
وضعیت استخدام	۰/۳۰۱

درآمد ماهیانهٔ گذشته	۰/۵۷۴
سابقهٔ کاری	۰/۱۸۶
سابقهٔ مدیریتی	-۰/۰۶۶
پورتفوی بیمه‌ای آورده شده	۰/۲۶۵

نشریه علمی پژوهشنامه بیمه دوره ۲، شماره ۱، زمستان ۱۳۹۶، شماره پیاپی ۲۳، ص ۲۲-۴۸

پیش‌بینی تداوم و ماندگاری نمایندگی

تداوم نمایندگی، شاخصی است برای اندازه‌گیری نسبت حق‌بیمهٔ ریزشی به کل حق‌بیمهٔ ورودی و شامل دو رده می‌شود: ٪۰-٪۷۵ (ردهٔ ۱) و ٪۷۵ -٪۱۰۰ (ردهٔ ۲). هرچه ماندگاری نماینده بیشتر باشد، احتمال ریزش بیمه‌نامه‌هایش کمتر می‌شود؛ بنابراین، نمایندگانی که به ردهٔ ۲ تعلق دارند، نمایندگان با کارایی بالا هستند. برای داده‌های آموزشی و داده‌های آزمون، دقت DA به‌ترتیب برابر است با ٪۶۹/۸۹ و ٪۵۷/۸۰ (٪۱۴ + ٪۴۳/۸ مطابق جدول ۵)، دقت متناظر برای شبکه‌های عصبی مصنوعی ٪۷۲/۵۶ و ٪۵۸/۴۰ (٪۱۳ +٪۴۵/۴ مطابق جدول ۵)، و این دقت برای درخت تصمیم برابر است با ٪۸۷/۲۰ و ٪۵۸/۲۰ (٪۱۰ + ٪۴۸/۲ مطابق جدول ۵). دقیق‌ترین ابزار برای پیش‌بینی تداوم و ماندگاری یک نماینده مدل شبکه‌های عصبی مصنوعی است. این مدل بالاترین دقت مشروط (٪۵۶/۳۹) را برای پیش‌بینی ردهٔ ۲ تولید می‌کند؛ اما برای پیش‌بینی ردهٔ ۱، درخت تصمیم بهترین دقت مشروط (٪۶۹/۴) را داراست.

جدول ۵: ماتریس آشفتگی برای پیش‌بینی تداوم و ماندگاری نمایندگی (به درصد)

روش داده‌کاوی	رده‌بندی	پیش‌بینی		کلی
		< ٪۷۵	=> ٪۷۵	
شبکهٔ عصبی مصنوعی				
واقعی	< ٪۷۵	۱۳	۳۴/۴	۴۷/۴
	=> ٪۷۵	۷/۲	۴۵/۴	۵۲/۶
کلی		۲۰/۲	۷۹/۸	۱۰۰
تحلیل تفکیک‌کننده				
واقعی	< ٪۷۵	۱۴	۳۳/۴	۴۷/۴
	=> ٪۷۵	۸/۸	۴۳/۸	۵۲/۶
کلی		۲۲/۸	۷۷/۲	۱۰۰
درخت تصمیم				
واقعی	< ٪۷۵	۱۰	۳۷/۴	۴۷/۴
	=> ٪۷۵	۴/۴	۴۸/۲	۵۲/۶
کلی		۱۴/۴	۸۵/۶	۱۰۰

سابقهٔ کاری و موقعیت شغلی دو مورد از مهم‌ترین عاملهایی هستند که در سطوح بالای درخت تصمیم ظاهر می‌شوند. درخت تصمیم نشان می‌دهد که شغل قبلی نماینده و پس از آن سن و تجربهٔ شغلی، بیشترین اهمیت را در فرایند پیش‌بینی ماندگاری نمایندگان دارند.

جدول ۶ نشان می‌دهد که تمام ضرایب بالا بزرگتر از ۰/۳۵ هستند؛ بنابراین، تحلیل هم‌زمان روشهای درخت تصمیم و تحلیل تفکیک‌کننده نشان می‌دهد که عاملهای ۱. تجربهٔ کاری، ۲. موقعیت شغلی، ۳. سن، و ۴. شغل قبلی، تأثیرگذارترین عاملها در تعیین تداوم و ماندگاری پورتفوی یک نمایندهٔ بیمه است.

جدول ۶: ضرایب تابع تفکیک متعارف استاندارد

متغیر	ضریب استاندارد
موقعیت	-۰/۱۵۰
جنسیت	-۰/۰۶۹
سن	-۰/۴۸۵
وضعیت تأهل	
مجرد	-۰/۲۵۲
متأهل	-۰/۱۴۱
مطلقه	-۰/۰۶۴
تعداد افراد تحت تکلف	۰/۱۸۴
سطح علمی	۰/۱۳۲
ماهیت شغل قبلی	
دفتری/فنی	۱/۰۲۲
بخش فروش/بخش خدمات	۱/۰۲۶
تخصصی	۰/۴۲۴
مدیریتی	۰/۸۵۸
وضعیت استخدام	۰/۲۶۰
درآمد ماهیانهٔ گذشته	۰/۱۱۵
سابقهٔ کاری	۰/۳۷۶
سابقهٔ مدیریتی	۰/۰۷۰
پورتفوی بیمه‌ای آورده شده	۰/۱۹۹

جدول ۷ نشان می‌دهد که کارایی و کاربرد سه روش مذکور در موقعیتهای مختلف متفاوت است. برای مثال، شبکه‌های عصبی مصنوعی اجرای مناسب‌تری در پیش‌بینی طول خدمت و تداوم پرتفوی دارد؛ در حالی که تحلیل تفکیک‌کننده دقیق‌ترین پیش‌بینی را از فروش حق‌بیمه دارد. نتایج نشان می‌دهد که ارتباطی خطی بین حق‌بیمهٔ فروخته‌شده و عاملهای تأثیرگذار برقرار است؛ این در حالی است که ارتباط بین طول مدت‌زمان خدمت و ماندگاری نمایندگی با این عاملها، غیرخطی است (Curram and Mingers, 1994). یادآوری این نکته لازم است که شبکه‌های عصبی مصنوعی به زمان محاسباتی بیشتری نیاز دارند، اما توانایی بالایی برای پیش‌بینی دو پارامتر دیگر یعنی طول مدت‌زمان خدمت و تداوم ماندگاری نمایندگان دارند.

جدول ۷: مزیت هر یک از روش‌های مذکور در کاربردهای مختلف

	شبکه‌های عصبی مصنوعی	روش‌های تحلیل تفکیک‌کننده	درخت‌های تصمیم
طول مدت‌زمان خدمت	✓		
حق‌بیمهٔ فروخته‌شده		✓	
تداوم و ماندگاری نمایندگان	✓		

نتایج و بحث

در این پژوهش با ترکیب سه روش داده‌کاوی، سیستمی برای انتخاب نمایندگان کیفی در صنعت بیمه طراحی شده است. تصمیم‌گیری در خصوص انتخاب یک نمایندۀ بیمه بر اساس مشخصه‌های موجود در کاندیدهای بالقوه صورت می‌پذیرد.

در میان روشهای داده‌کاوی، ساده‌ترین ابزار برای طبقه‌بندی درخت تصمیم است که در پژوهش حاضر، از این روش برای پیش‌بینی حق‌بیمۀ فروخته‌شده استفاده شد. نتایج کلی نشان می‌دهد که ۱. تجربۀ شغلی، ۲. موقعیت شغلی، ۳. سن، ۴. وضعیت تأهل، ۵. شغل قبلی، ۶. درآمد سالیانه از محل شغل قبلی، و ۷. بیمه‌نامه‌های فروخته‌شده، از جملۀ مهم‌ترین فاکتورها در تعیین مدت‌زمان فعالیت نمایندگان جدید، بیمه‌نامه‌های فروخته‌شده و تمدید بیمه‌نامه‌های صادرشده (جلوگیری از ریزش بیمه‌نامه‌های صادره) هستند؛ بنابراین می‌توان با استفاده از این ویژگیها فرایند انتخاب نمایندگان را بهبود بخشید. با استفاده از این سه روش می‌توان طول ماندگاری و کارایی متقاضیان اخذ کد نمایندگی را با دقت معقولی پیش‌بینی کرد. درضمن به دلیل راحتی کار با این سیستم، مدیران نیز اقبال خوبی به آن نشان خواهند داد.

منابع و ماخذ

Blanco, C.; De Guzmán, I,G-R.; Fernández-Medina, E.; Trujillo., J., (2015). An architecture for automatically developing secure OLAP applications from models. Information and Software Technology, 59, pp. 1-16.

Bellatrechea, L.; Cuzzocreab, A.; Songc, I-Y., (2015). Advances in data warehousing and OLAP in the big Data Era. Information Systems, 53, pp. 39-40.

Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J., (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

Brockett, P.L.; Cooper, W.W.; Golden, L.L.; Xia, X., (1997). A case study in applying neural networks to predicting insolvency for property and casualty insurers. Journal of the Operational Research Society, 48, pp. 1153-1162.

Cho, V.; Wüthrich B., (2002). Distributed mining of classification rules. Knowledge and Information Systems, 4, pp. 1–30.

Cho, V.; Wüthrich, B.; Zhang, J., (1999). Text processing for classification. Journal of Computational Intelligence in Finance, 7(2), pp. 6-22.

Codd, E.; Codd, S.; Salley, C., (1993). Providing OLAP (on-line analytical processing) to user-analysts: an IT mandate, Technical Report. E.F. Codd & Associates.

Curram, S.P.; Mingers, J., (1994). Neural networks, decision tree induction and discriminant analysis: an empirical comparison. Journal of the Operational Research Society, 45(4), pp. 440-450.

Dehne, F.; Kong, Q.; Rau-Chaplin, A.; Zaboli, H.; Zhou, R., (2015). Scalable real-time OLAP on cloud architectures. Journal of Parallel and Distributed Computing, 79-80, pp. 31-41.

Delmater, R.; Hancock, M., (2001). Data Mining Explained: A Manager's Guide to Customer-centric Business Intelligence. Boston, MA: Digital Press.

Duda, R.O.; Hart, P.E.; Stork, D.G., (2001). Pattern Classification. 2nd edn, New York: Wiley.

Houaria, R.; Bounceur, A.; Kechadi, M.T.; Tari, A.K.; Euler, R., (2016). Dimensionality reduction in data mining: A copula approach. Expert Systems with Applications, 64, pp. 247-260.

Inmon, W.H., (2002). Building the Data Warehouse. 3rd edn, NewYork: Wiley.

Kass, G.V., (1980). An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29, pp. 119-127.

Kimball, R., (1996). The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. NewYork: Wiley.

Kohonen, T., (2001). Self-organizing Maps. New York: Springer.

Quinlan, J.R., (1987). Generating production rules from decision trees. International Joint Conference on Artificial Intelligence, San Mateo, CA: Morgan Kaufmann, pp. 304-307.

Quinlan, J.R., (1993). Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

Ripley, B.D., (1994). Neural networks and related methods for classification. Journal of the Royal Statistical Society, Series B (Methodological), 56 (3), pp. 409-56.

Rumelhart, D.E.; Hinton, G.E.; Williams, R.J., (1986). Learning representations by back-propagating errors. Nature, 323, p. 533.

Werbos, P.J., (1994). The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting. New York: Wiley.