Iran J. Insur. Res., 4(4): 399-411, Autumn 2014, Serial #14

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

Insurance Research Center (IRC)

ORIGINAL RESEARCH PAPER

Dynamic tariffing in insurance industry with data mining method (case study: third person car insurance)

D. Karimzadehgan Moghaddam*, M. Behravan
Department of Information Technology Management, Faculty of Computer Engineering and Information Technology, Payame Noor University of Tehran, Tehran, Iran

ARTICLE INFO

Article History
Received: 12 March 2015
Revised: 06 November 2015
Accepted: 16 February 2016

Keywords
Car insurance; Data mining; Data warehouse; Dynamic tariff; Premium.

ABSTRACT

The third person car insurance has the biggest portion in the insurance market which makes an appropriate opportunity for data discovery and extracting unknown patterns for decision making in insurance industry. Currently premium is calculated with the minimum consideration to the risks. Based on these calculations lots of damages may occur to the insurers which affects the quality of their services which also leads to costumers' dissatisfaction. To make data mining happen, vehicles' information, police background and the insured person's information are gathered, standardized and stored in a data warehouse from over ۳۰ million policies and ۲,۷ million losses. For standardizing vehicles' specification, police data are used and insured people's identity are evaluated with national databases. Thereafter a mining structure designed and the three algorithms of clustering, neural network, and decision tree were performed on it. Finally, all models are evaluated using sample data and the results checked with confusion matrix and loss rate, which indicates the feasibility of this method in dynamically tariffing for this type of policies and leads to the decrease of the loss rate, the confusion matrix also indicates the accuracy of the evaluation.

***Corresponding Author:**
Email: *d_karimzadgan@pnu.ac.ir*
DOI: ۱۰,۲۲۰۵٦/ijir.۲۰۱٤.۰٤,۰۱

نشریه علمی پژوهشنامه بیمه

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

مقاله علمی

ارائه راهکاری برای تعرفه‌گذاری پویا در صنعت بیمه با استفاده از تکنیک داده‌کاوی (مورد مطالعه: بیمه شخص ثالث)

داود کریم زادگان مقدم*، مجید بهروان

گروه مدیریت فناوری اطلاعات، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور تهران، تهران، ایران

اطلاعات مقاله

تاریخ دریافت: ۲۱ اسفند ۱۳۹۳

تاریخ داوری: ۱۵ آبان ۱۳۹۴

تاریخ پذیرش: ۲۶ بهمن ۱۳۹۴

کلمات کلیدی

انبارداده‌ها

بیمه‌نامه اتومبیل

تعرفه‌گذاری پویا

حق بیمه

داده‌کاوی

چکیده:

بیمه‌نامه شخص ثالث بیشترین سهم از بازار بیمه کشور را دارا می‌باشد و فرصت مناسبی برای کاوش اطلاعات و استخراج الگوهای ناشناخته جهت تصمیمات کلان در صنعت بیمه را فراهم می‌نماید. در حال حاضر حق بیمه باکمترین توجه به عوامل ریسک بیمه‌گذاران محاسبه می‌گردد، که موجب زیانده شدن بیمه‌نامه شخص ثالث برای شرکت‌های بیمه و نارضایتی بیمه‌گذاران از خدمات شرکت‌های بیمه گردیده است. بدین منظور در این پژوهش اطلاعات خودرویی، سوابق بیمه‌ای و ویژگی‌های بیمه‌گذاران در بیش از ۳۰ میلیون بیمه‌نامه و ۲,۷ میلیون خسارت جمع‌آوری و استانداردسازی شده و در انبارداده‌ها ذخیره گردید. برای استانداردسازی داده‌های خودرویی با بانک اطلاعاتی نیروی انتظامی و اطلاعات هویتی بیمه‌گذاران با استفاده از داده‌های سازمان ثبت احوال کشور، اعتبارسنجی و تکمیل گردیده است. سپس ساختار کاوشی طراحی و با استفاده از سه الگوریتم خوشه بندی، شبکه عصبی و درخت تصمیم و داده‌های آموزشی مورد آموزش قرار گرفت. در نهایت مدل‌ها با استفاده از داده‌های آزمایشی مورد بررسی قرار گرفته و نتایج به دست آمده از مدل‌ها با استفاده از نمودار صعود و نسبت خسارت مورد اعتبارسنجی قرار گرفتند که نتایج به دست آمده نشانگر امکان استفاده از چارچوب ارائه شده را نشان می‌دهد، به نحوی که نسبت خسارت کاهش و نمودار صعود صحت ارزیابی را نشان می‌دهد.

***نویسنده مسئول:**

ایمیل: *d_karimzadgan@pnu.ac.ir*

DOI: ۱۰,۲۲۰۵٦/ijir.۲۰۱٤.۰٤,۰۱

مقدمه

بیمه راهکاری است که ریسک را به طرفی انتقال می‌دهد که در ریسک مزیت نسبی دارد. در مثال معمول بیمه، شخص ریسک‌پذیر (مواجه با ریسک)، تمایل دارد که قیمت ثابتی را به بیمه‌گر کمتر ریسک گریز یا بیمه‌گر دارای تنوع ریسک، که حاضر است به آن قیمت ریسک را تقبل کند، بپردازد. از آنجا که دو طرف در مورد قرارداد توافق دارند، وضع هر دو بهتر خواهد شد (Dionne,G ۲۰۰۵). همچنین حق بیمه وجهی است که، بیمه‌گذار در قبال خرید بیمه و یا به عبارت بهتر انتقال خطر به بیمه‌گر پرداخت می‌کند. با پرداخت حق بیمه و خرید بیمه مسئولیت مدنی، تعهد جبران خسارت از بیمه‌گذار منتقل‌شده و بیمه‌گر متعهد می‌گردد.

قانون بیمه اجباری مسؤولیت مدنی دارندگان وسایل نقلیه موتوری زمینی در مقابل شخص ثالث که در ایران اختصاراً قانون بیمه شخص ثالث نامیده می‌شود، اولین بار در ۱۴ ماده و سه تبصره و در تاریخ ۱۳۴۷/۹/۲۶ به تصویب مجلس شورای ملی و متعاقباً در تاریخ ۱۳۴۷/۱۰/۲۳ به تصویب مجلس سنا رسیده و از اول فروردین ماه ۱۳۴۸ نیز به اجرا گذاشته شد.

اما با گذشت چهل سال و در تاریخ ۱۳۸۷/۴/۱۶ این قانون اصلاح شده و با عنوان جدید "قانون اصلاح قانون بیمه اجباری مسؤولیت مدنی دارندگان وسایل نقلیه موتوری زمینی در مقابل شخص ثالث" در ۳۰ ماده و ۲۵ تبصره به تصویب کمیسیون اقتصادی مجلس شورای اسلامی رسیده و به اجرا گذاشته شده است (قانون شخص ثالث،۱۳۸۷).

اما پس از اصلاح این قانون، همچنان نحوه محاسبه میزان حق بیمه اتومبیل بسیار نامتوازن و براساس مسائل سیاسی کشور نرخ‌گذاری گردیده است و کمتر جنبه فنی نرخ‌گذاری در آن رعایت شده است که موجب زیانده شدن این رشته در صنعت بیمه گردیده است.

بیمه شخص ثالث، یکی از مهمترین خدمات بیمه‌ای بوده و برخی از آن به عنوان ویترین شرکت‌های بیمه یاد می‌کنند. در واقع این بیمه‌نامه به دلیل اجباری بودن می‌تواند به عنوان دروازه ورود و جذب مشتری به یک شرکت بیمه برای فروش سایر پوشش‌های بیمه‌ای نیز باشد. بنابراین رضایت بیمه‌گذاران در این شاخه از خدمات بیمه‌ای، می‌تواند اثر قابل توجهی بر حضور دائم بیمه‌گذار گذاشته و موجب سودآوری بیشتر گردد (رنجبرفرد، ۱۳۹۳).

ورشکستگی شرکت‌های بیمه در مقایسه با شرکت‌ها در سایر صنایع می‌تواند اثرات مخرب‌تری بر مشتریان این صنعت و جامه به صورت کلی داشته باشد (دقیقی، ۱۳۹۲)، لذا بایستی با ایجاد بستر تعرفه‌گذاری صحیح ریسک شرکت‌های بیمه را مدیریت کرد.

مرور ادبیات تحقیق

حق بیمه

از دید اقتصاد بیمه، میزان ارزش افزوده ناشی از تجمیع ریسک با استفاده از مفهوم حق بیمه، قابل اندازه گیری است. بدین معنا که اگر فرض کنیم فرد ریسک پذیری با درآمد Y مواجه است و یا این امکان را دارد که درآمدی معادل Y۰ را با اطمینان انتخاب کند، حاضر است Y۰ را که کمتر از میانگین Y یعنی E(Y) است برگزیند، در این صورت تفاوت بین Y۰ و E(Y) در واقع حکم حق بیمه را پیدا می‌کند. به عبارت دیگر حق بیمه مبلغی است که تصمیم گیرنده را بین پذیرش ریسک و بیمه کردن آن بی تفاوت می‌نماید (خرمی، ۱۳۸۲).

حق بیمه (π) عبارت از حل رابطه زیر است:

فرمول شماره ۱: $$U(W-\mu L-\pi)=\int U(W-L)+f(L)dL$$

که در آن:

U(W)= تابع مطلوبیت U'>۰, U''<۰

W= ثروت اولیه (غیر تصادفی)

L= خسارت (تصادفی) L≥۰

f(L)= تابع توزیع احتمال خسارت

$E(L)= \mu L$= میانگین خسارت احتمالی (خرمی، ۱۳۸۲).

بنابراین می‌توان گفت که ارزش افزوده ناشی از انعقاد قرارداد بیمه، عبارت است از حداکثر ارزش حاصله بیش از خسارت انتظاری که بیمه‌گذار حاضر به پرداخت آن است، زیرا بیمه‌گذار مختار است که بیمه خریداری نکند و در عوض ثروت ریسکی را در اختیار داشته باشد که ارزش انتظاری است که بیمه‌گذار حاضر به پرداخت آن بوده و مبین ارزش بیمه است (خرمی، ۱۳۸۲).

در یک بازار رقابتی بیمه، میزان کل رفاه بیمه‌گذار ممکن است قابل مشاهده نباشد. به این معنا که شاید بیمه‌گذار بتواند بیمه‌ای را با هزینه-های اضافی کمتر از π خریداری کند، اما غیر ممکن است که مازاد بیمه‌گذار را به علت غیر قابل مشاهده بودن اندازه‌گیری کرد. درحالی که واضح است که مبلغ پرداختی بیمه‌گذار بیش از مبلغ انتظاری ارزش افزوده قابل اندازه‌گیری ناشی از تجمیع ریسک است (خرمی، ۱۳۸۲).

حق بیمه ناخالص عبارت است از مجموع هزینه‌هایی که بیمه‌گذار در قبال دریافت خدمات بیمه‌ای به شرکت‌های بیمه پرداخت می‌نماید. علاوه بر حق بیمه خالص هزینه‌هایی مانند هزینه اداری، کارمزدها و میزان سود شرکت بیمه به ازای بیمه‌نامه حق بیمه ناخالص را تشکیل می‌دهند. در مقابل میزان حق بیمه‌ای که بیمه‌گذار به ازای ریسک‌های وارده به شرکت بیمه در طول مدت اعتبار بیمه‌نامه پرداخت می‌نماید، حق بیمه خالص نامیده می‌شود.

برای تعیین حق بیمه ابتدا می‌بایست مقدار مورد انتظار ادعای خسارت برای هر قرارداد بیمه را برآورد کرد. بدین منظور تابع شرطی $P(x)$که در آن متغیر ورودی x، اطلاعات موجود هر بیمه‌گذار می‌باشد، میزان خسارت هر بیمه‌گذار را بر اساس شرایط خاص او محاسبه می‌نماید، مبنای محاسبه حق بیمه خالص برای بیمه‌گذار خواهد بود (Chapados, ۲۰۱۰).

فرمول شماره ۲ (Chapados, ۲۰۱۰):

$$P(X) = E[A|X]$$

متغیر A میزان خسارت بیمه‌گذار به ازای شرایط X را نشان می‌دهد. در واقع میزان خسارت تمامی بیمه‌گذاران با شرایط یکسان خروجی یکسانی را خواهد داشت.

داده‌کاوی

موضوع داده‌کاوی شناخت چیزهای جدید و با ارزش، بالقوه مفید، رابطه‌های منطقی و الگوهای موجود در داده‌ها است. در جوامع مختلف یافتن الگوهای مفید در داده‌ها با عناوین متعدد بیان می‌شود. برای مثال از عنوان‌هایی نظیر استخراج دانش، کشف اطلاعات، برداشت اطلاعات، پردازش الگوهای داده‌ها (Fayyad,۱۹۹۶) می‌توان نام برد.

عبارت "داده کاوی" توسط آمارشناسان، محققان پایگاه‌های داده‌ها و سیستم‌های اطلاعات مدیریتی و جوامع بازرگانی به کار برده می‌شود عبارت کشف دانش در پایگاه داده‌ها عموماً برای اشاره به فرآیند کلی کشف دانش مفید از داده‌هایی که داده‌کاوی گام مهمی در این فرآیند است، مورد استفاده قرار می‌گیرد (Fayyad, ۱۹۹۶). هر نویسنده، محقق و کاربر با توجه به دیدگـــاه و نوع نگرش خود تعریف‌های مختلفی از داده‌کاوی ارائه کرده است.

مزایا و ویژگی‌های بهره‌گیری از داده‌کاوی در تعرفه‌گذاری

در روش‌های رایج تعرفه‌گذاری، بررسی تاثیر عوامل موثر بر ریسک به صورت محدود و منظم فراهم می‌شود که جامعت، صحت و دقت محاسبات را کاهش می‌دهند.

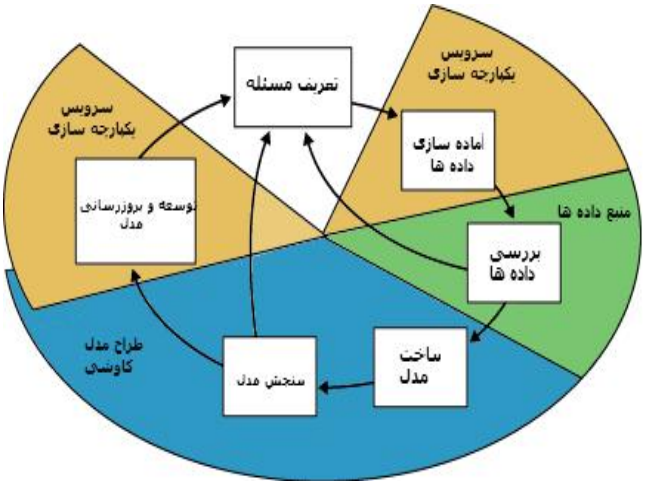
به عنوان مثال عواملی که هم اکنون در محاسبه حق بیمه شخص ثالث به کار گرفته می‌شود، بسیار محدود و به صورت کلی مورد بررسی قرار گرفته‌اند.

در حالی که الگوریتم‌های داده‌کاوی با بررسی میزان تاثیر فاکتورهای متنوع بر عامل نهایی، به صورت پویا و غیر خطی رفتار عامل هدف را تخمین می‌زنند.

فرآیند داده‌کاوی

فرآیند داده‌کاوی شامل تعریف مسئله، آماده‌سازی داده‌ها، بررسی و اعتبارسنجی داده‌ها، تهیه مدل‌ها، بررسی و ارزیابی مدل‌ها، اجرا و به-روزرسانی مدل‌ها می‌باشد.

شکل ۱: فرآیند و تکنولوژی‌های موجود در SQL Server به منظور داده‌کاوی[۱]



این فرآیند به صورت چرخه است ولی هر مرحله لزوما به صورت مستقیم به مرحله بعدی منجر نمی‌شود؛ بلکه باید در نظر داشت که تهیه یک مدل داده‌کاوی، فرآیندی پویا و تکرار پذیر است؛ چرا که پس از بررسی داده‌ها ممکن است دریابیم که داده‌های موجود برای تهیه مدل مناسب داده کاوی، کافی نیستند و بنابراین لازم است در جستجو داده‌های بیشتری باشیم. ممکن است مجبور باشیم مدل‌ها را پس از اجرای آنها بروز نماییم چون داده‌های بیشتری به دست آوریم. بنابراین درک اینکه یک مدل داده‌کاوی یک فرآیند چرخشی است و ممکن است لازم باشد هریک از مراحل فرآیند آنقدر تکرار شود تا مدل مناسبی تهیه گردد؛ از اهمیت ویژه‌ای برخوردار است (MicrosoftMSDN,۲۰۱۴).

تجزیه و تحلیل

تعریف مسئله

تعداد فاکتورهای موثر بر خسارت در قالب فاکتورهای خودرویی، بیمه‌ای و انسانی قابل بررسی است. در بررسی میزان تاثیر فاکتورهای بر خسارت از توابع وابستگی مایکروسافت استفاده خواهد شد.

پس از مشخص شدن فاکتورهای موثر از توابع خوشه‌بندی برای تفکیک ریسک بیمه‌گذاران و پیش‌بینی میزان خسارت هریک از خوشه‌ها استفاده می‌گردد.

الگوریتم‌های درخت تصمیم امکان نمایش میزان اثرگذاری با جزئیات بیشتر به نمایش در خواهد آورد و امکان تصمیم‌گیری بهتر از میزان تاثیر فاکتورهای ریسک را فراهم می‌نماید.

ساختار انبارداده‌ها بر اساس داده‌های مورد نیاز پژوهش طراحی و سپس توسط ابزارهای یکپارچه سازی جمع‌آوری و پس از استانداردسازی در انبارداده‌ها قرار می‌گیرند.

[۱]. Microsoft, ۲۰۱۴,Data Mining Algorithms (Analysis Services - Data Mining)
From: https://msdn.microsoft.com/en-us/library/ms۱۷۴۹۴۹.aspx

در بخش ابتدا می‌بایست فیلدهای اطلاعاتی و نوع آنها برای انجام این پژوهش تعیین گردد. اطلاعتی که بیمه گر بایستی در مورد یک قرارداد بیمه داشته باشد: برخی از این اطلاعات در جهت رعایت حسن نیت و با هدف افشای اطلاعاتی که اموری در اختیار بیمه‌گر گذاشته شود که وی نسبت به آنها جاهل است و به آن نیاز دارد. اما پاره‌ای از اطلاعات است که بیمه‌گر نسبت به آن عالم بوده و نیازی به افشای آن نیست. همچنین پاره‌ای از اطلاعات وجود دارد که بیمه‌گر به اقتضای حرفه خود باید نسبت به آنها عالم باشد (انصاری،۱۳۹۲).

مجموعه داده‌هایی که در این پژوهش مورد کاوش قرار می‌گیرند در سه گروه عوامل خودرویی، انسانی و اطلاعات بیمه نامه‌ای قابل تفکیک می‌باشند.

معماری مورد نظر در جمع‌آوری و استانداردسازی داده‌ها و استفاده از بانک‌های اطلاعاتی نیروی انتظامی و سازمان ثبت احوال جهت اعتبارسنجی، طراحی و انتقال داده‌های گردآوری شده در ساختار انبارداده‌ای، مطابق شکل شماره ۲ زیرساخت فرآیند داده‌کاوی در صنعت بیمه را می‌تواند با دقت بسیار بالایی فراهم می‌نماید. بکارگیری داده‌کاوی در تعرفه‌گذاری اساس ارزیابی مناسب از ریسک بیمه‌گذاران را فراهم می‌نماید و شرایط رشد شرکت‌های بیمه و رضایت مشتریان این صنعت را شکل می‌دهد. میزان توسعه بیمه در کشورهای مختلف ارتباط مستقیمی با میزان توسعه یافتگی و شکوفایی اقتصادی کشور را نشان می‌دهد به نحوی که با تحلیل دقیق اطلاعات موجود و انتقال صحیح ریسک صنایع و افراد به ساختار بیمه‌ای، ریسک در اقتصاد توزیع و رشدی متوازن در جامعه را پدید خواهد آورد.

ویژگی های خودرو

با توجه به نظر کارشناسان صنعت بیمه کشور یکی از عوامل موثر بر میزان ریسک هر یک از بیمه‌گذاران نوع و ویژگی‌های وسیله نقلیه می‌باشد. از این رو جهت تعیین تعرفه حق بیمه، می‌بایست اطلاعات خودرویی مورد کاوش قرار گیرند.

مجموع فاکتورهایی که در این بخش مورد استفاده قرار گرفته‌اند، با استفاده از بانک اطلاعاتی نیروی انتظامی اعتبارسنجی و پالایش شده‌اند.

جدول ۱: ویژگی‌های خودرو در مدل کاوشی

نام فیلد	نوع	واحد	پیوستگی
سال ساخت	عدد صحیح[۱]	سال	گسسته
کاربری	متن[۲]	نوع	گسسته
رنگ اصلی	متن	نوع	گسسته
نوع سوخت	متن	نوع	گسسته
برند	متن	نوع	گسسته
مدل برند	متن	نوع	گسسته
ظرفیت	متن	نفر و کیلوگرم	گسسته
محور	عدد صحیح	تعداد	گسسته
مسافت پیموده شده	ده‌دهی[۳]	کیلومتر	پیوسته
چرخ	عدد صحیح	تعداد	گسسته
تعداد سیلندر	عدد صحیح	تعداد	گسسته

ویژگی‌های بیمه‌گذار

[۱]. int

[۲]. nvarchar

[۳]. decimal

رفتار رانندگان متاثر از عوامل بسیاری است که در مطالعات قابل جمع‌آوری نیستند (Higgs, B,۲۰۱۵) و به دلیل عدم ذخیره صحیح این اطلاعات در بانک‌های اطلاعاتی و عدم امکان صحت‌سنجی اطلاعات در زمان انجام این پژوهش با سایر بانک های اطلاعاتی از آنها صرف نظر شده است. فیلدهای اطلاعاتی مربوط به بیمه‌گذاران به شرح جدول شماره ۲ می‌باشد:

جدول۲: ویژگی‌های بیمه‌گذاران در مدل کاوشی

نام فیلد	نوع	واحد	پیوستگی
جنسیت	بیت[۱]	-	گسسته
تحصیلات	متن	دوره	گسسته
تعداد خودرو	عدد صحیح	تعداد	گسسته
سن	عدد صحیح	سال	گسسته
اعضای خانواده تعداد	عدد صحیح	تعداد	گسسته

میزان تحصیلات افراد بر اساس اطلاعات معتبر مرکز آمار ایران به صورت نسبی تکمیل گردیده است.

جدول ۳: موقعیت جغرافیایی

نام فیلد	نوع	واحد	پیوستگی
استان محل صدور بیمه نامه	متن	-	گسسته
شهر محل صدور	متن	-	گسسته

محل صدور بیمه‌نامه بر اساس موقعیت جغرافیایی نمایندگی و یا شعبه صادر کننده بیمه‌نامه تعیین می‌گردد.

سوابق بیمه‌نامه‌ای

از مهمترین عوامل تعیین کننده ریسک بیمه‌گذاران، فاکتورهای مالی و بیمه‌ای افراد می‌باشد. از جمله دلایل صحت بیشتر این پژوهش می‌توان به یکپارچه سازی اطلاعات تمامی شرکت‌های بیمه‌ای در انبارداده‌ها جهت بررسی کامل و رصد رفتار بیمه‌گذاران در کل صنعت اشاره کرد، نکته‌ای که تا قبل از این پژوهش امکان‌پذیر نبوده است.

جدول ۴: سوابق بیمه‌ای در مدل کاوشی

نام فیلد	نوع	واحد	پیوستگی
تعداد سنوات عدم خسارت	عدد صحیح	تعداد	گسسته
جریمه صندوق تامین خسارتهای بدنی	ده‌دهی	ریال	پیوسته
خسارت جانی	عدد صحیح	تعداد	گسسته
خسارت مالی	عدد صحیح	تعداد	گسسته
دفعات زیان‌دیدگی	عدد صحیح	تعداد	گسسته
میزان جرائم رانندگی	ده‌دهی	ریال	پیوسته
تخلفات حادثه‌ساز	عدد صحیح	تعداد	گسسته

[۱]. bit

فرآیند بررسی و اعتبارسنجی داده‌ها

با توجه به اینکه داده‌های جمع‌آوری شده از بانک‌های اطلاعاتی شرکت‌های بیمه از منابع اطلاعاتی گوناگون با ساختار و فرمت‌های داده‌ای متفاوت جمع‌آوری شده‌اند، نیازمند همگون‌سازی نوع و محتوای داده‌ها می‌باشند، تا استخراج مفاهیم از داده‌های با تعاریف متفاوت امکان‌پذیر شود.

جمع‌آوری اطلاعات از شرکت‌های بیمه‌ای که هریک از نرم‌افزار مستقل خود برای عملیات‌های بیمه‌گری استفاده می‌کنند، مستلزم فرآیند استاندارد و پاکسازی داده‌ها جهت ایجاد انبارداده‌ها می‌باشد.

در مجامع علمی معتبر معماری سرویس‌گرا جهت اعتبارسنجی داده‌های بین سازمانی در بستر تکنولوژی‌های امن ارتباطات شبکه‌ای، به عنوان بهترین راهکار برای این فرآیند توصیه گردیده است و لذا در این فرآیند مورد استفاده قرار گرفت. بدین منظور داده‌های استاندارد شده در مرحله استانداردسازی امکان فراخوانی سرویس‌های سایر سازمانها برای تکمیل و اصلاح اطلاعات جمع‌آوری شده را فراهم می‌نماید.

امکان اعتبارسنجی داده‌های مورد نیاز بیمه‌گذاران در این پژوهش بر اساس کد ملی و تاریخ تولد افراد از سازمان ثبت احوال کشور و اعتبارسنجی اطلاعات خودرویی مورد بیمه نیز از طریق سرویس‌های نیروی انتظامی، به عنوان مرجع ثبت اطلاعات خودرویی کشور فراهم می‌باشد.

سرویس ارائه شده نیروی انتظامی با دریافت شماره پلاک اطلاعات خودروی مربوطه به پلاک ارسالی را از بانک اطلاعاتی نیروی انتظامی فراخوانی و باز می‌گرداند به همین دلیل، کلیه اطلاعات دریافتی به نوعی فقط از یک منبع خواهد بود و نیاز به نگاشت مجدد اطلاعات نیست.

با انجام این مرحله داده‌های مربوط به خودروهای پژوهش، در انبارداده‌ها برای ادامه فرآیند داده‌کاوی آماده می‌گردد.

ساخت مدل کاوش

داده‌کاوی تا حد بسیاری با ساخت مدل سر و کار دارد. مدل کاوشی شامل مجموعه‌ای از قواعد و یا یک الگوریتم می‌باشد که به مجموعه‌ای از ورودی ها متصل و خروجی و اهداف خاصی را نشان می‌دهد. مدل برای بازنمایی و بصری‌سازی یک واقعیت به منظور درک بهتر آن به کار می رود.

یک مدل داده‌کاوی یا مدل کاوش می‌تواند به صورت یک جدول رابطه‌ای تصور گردد. هر مدل کاوش در ارتباط با یک یا چند الگوریتم است که مدل بر اساس آن آموزش دیده است.همان طور که یک جدول رابطه ای حاوی رکوردهایی از داده‌ها می‌باشد، یک مدل کاوش نیز حاوی الگوریتم‌ها است.

آموزش مدل کاوش به معنای یافتن الگوها در مجموعه داده آموزشی با استفاده از الگوریتم‌های مشخص شده و تنظیم پارامترهای مناسب هر الگوریتم می‌باشد. پس از آموزش مدل، الگوهای استخراج شده توسط الگوریتم‌ها برای هر مجموعه داده در مدل‌های کاوشی ذخیره می‌شوند. سه مرحله پایه برای ساخت مدل عبارتند از: ایجاد ساختار کاوشی، آموزش مدل، پیش‌بینی مدل، که در ادامه مورد تحلیل قرار خواهند گرفت.

ایجاد ساختار کاوش

اولین مرحله ایجاد یک ساختار کاوشی می‌باشد. ایجاد این ساختار، شبیه ایجاد یک جدول رابطه‌ای در پایگاه داده می‌باشند. تعریف یک ساختار کاوشی شامل تعیین تعداد ستون‌های ورودی، تعداد ستونهای قابل پیش‌بینی و الگوریتم وابسته آن است.

مدل کاوش برای ذخیره نمودن الگوهای استخراج شده توسط الگوریتم‌های داده‌کاوی بکار می‌رود.

ساختار کاوش، یک ساختار داده است و محدوده داده‌هایی را که از روی آنها مدل‌های کاوش ساخته می‌شود را تعریف می‌نماید. یک ساختار کاوش می‌تواند شامل چندین مدل باشد که از یک مجموعه داده به طور اشتراکی استفاده می‌کنند. یک مدل کاوشی، الگوریتم‌های کاوش را به داده‌هایی که ساختار کاوش ارائه می‌نماید، اعمال می‌کند.

عناصر سازنده ساختار کاوش، ستون‌های ساختار کاوشی هستند که داده‌هایی را که منبع اصلی داده فراهم می‌کنند، توصیف می‌نمایند. این ستون ها شامل اطلاعاتی از قبیل نوع داده، نوع محتوا، ماهیت داده و اینکه داده چطور توزیع شده است، می‌باشند.

مدلهای کاوش باید هم شامل ستون کلیدی تعریف شده در ساختار کاوش باشند و هم شامل زیر مجموعه‌ای از ستون‌های باقیمانده باشند.

آموزش مدل کاوش با داده‌های انبارداده‌ها

این مرحله شامل فرآیند تفکیک‌سازی داده‌ها به مجموعه آموزشی و آزمایشی می‌باشد.

بنابراین شما می‌بایست در ابتدا داده‌ها را آماده‌سازی و یک بانک اطلاعاتی ویژه برای این منظور ایجاد نمایید؛ به گونه‌ای که داده‌ها را به دو بخش مجموعه آموزشی و آزمایشی تفکیک نموده و سپس مدل‌های داده‌کاوی را با استفاده از این دو مجموعه داده، پیاده‌سازی نمایید.

پیش‌بینی مدل

پیش‌بینی مهمترین مرحله در پردازش داده‌کاوی می‌باشد و اغلب هدف نهایی در پروژه‌های داده‌کاوی می‌باشد. پیش‌بینی به روشی برای بستن چرخه تحلیل اشاره می‌کند، اجرای آن ساده است و یک وظیفه موثر برای الگوریتم‌های داده‌کاوی می‌باشد.

در پیش‌بینی به یک مدل کاوشی آموزش دیده و یک مجموعه داده جدید نیاز است. در طول پیش‌بینی، موتور داده‌کاوی، قواعد به دست آمده در مرحله یادگیری را در مورد مجموعه داده جدید ورودی تخصیص می‌دهد.

بررسی و ارزیابی صحت مدل‌های کاوش

پس از ساخت مدل و ذخیره الگوها، می‌بایست نتایج به دست آمده ارزیابی و تفسیر شوند. میزان دقت مدل‌ها به تنهایی برای انتخاب بهترین مدل کافی نیست و نیازمند آگاهی در خصوص انواع خطاها و هزینه‌های وابسته به آنها نیز می‌باشد. در این مرحله مواردی از قبیل عملکرد درست مدل، توانایی توصیف داده‌ها، اعتماد به مدل‌های پیش‌بینی و جامعیت مدل مورد بررسی قرار می‌گیرد.

نمودار صعود

این نمودار دقت پیش‌بینی های ایجاد شده را در مورد یک مدل داده‌کاوی مشخص و محاسبه می‌نماید. این کار را توسط پیش‌بینی یک ستون از داده‌های مجموعه آزمایشی و سپس مقایسه‌ی آنها با مقدار واقعی انجام می‌دهد. سپس مقدار پیش‌بینی شده و مقدار واقعی به صورت گرافیکی نمایش داده می‌شوند.

این نمودار میزان اثر بخشی مدل را به صورت مقایسه نتایج حاصل شده از مدل پیش‌بینی کننده با نتایج حاصل شده بدون پیش‌بینی، مورد سنجش قرار می‌دهد.

بوسیله رسم تفاوت بین داده‌های تصادفی و داده‌های هدفمند، می‌توانیم یک نمودار صعود ایجاد کنیم.

برای پیاده‌سازی نمودار صعود موارد زیر نیاز است :

یک مدل آموزش دیده،

یک مجموعه داده آزمایشی برای استفاده در مدل پیش بینی،

نگاشتی از داده‌های آزمایشی به عنوان داده‌های ورودی به ساختار مدل کاوشی.

یک نمودار استاندارد صعودی شامل یک منحنی به ازای هر مدلی که انتخاب شده است و دو منحنی دیگر، یکی منحنی ایده‌آل و دیگری منحنی حدس تصادفی است.

نمودار صعود کمک شایانی در ارزیابی مفید بودن یک مدل می‌نماید و نشان می‌دهد چگونه پاسخ ها با بکار بستن مدل‌ها تغییر می‌کنند. این نسبت تغییر، لیفت[۱] نامیده می‌شود.

شاخص نسبت خسارت

نسبت خسارت در صنعت بیمه یکی از شاخص های کلیدی ارزیابی پرتفوی رشته‌های بیمه‌ای به حساب می‌آید. نسبت خسارت عبارت است از حاصل تقسیم خسارت بر حق بیمه دریافتی که با استفاده از این شاخص می‌توان میزان خسارت پیش‌بینی شده را با وضعیت موجود در صنعت ارزیابی کرد.

۵-۳. خوشه‌بندی ریسک مشتریان با استفاده از الگوریتم خوشه‌بندی

یک خوشه، به مجموعه‌ای از اشیاء داده‌های اطلاق می‌شود که از جهاتی شبیه به یکدیگرند.

تحلیل خوشه‌ای، خوشه‌های موجود در داده‌ها را تعیین می‌کند. الگوریتم خوشه‌بندی بهینه، خوشه‌هایی با کیفیت بالا تهیه می‌کند؛ این بدان معنی است که الگوریتم خوشه‌بندی را به نحوی انجام می‌دهد که شباهت بین خوشه‌ای کم و شباهت درون خوشه‌ای، زیاد می‌باشد؛ به بیان دیگر، اعضای یک خوشه بسیار شبیه تر به یکدیگرند تا هریک از اعضای سایر خوشه‌های دیگر.

اغلب اطلاعات و داده‌های موجود در پایگاه‌های داده‌ها توزیع‌های ناشناخته یا پیچیده‌ای دارند که به راحتی نمی‌توان آن توزیع‌ها را شناسایی نمود و مورد استفاده قرار داد. بنابراین برای تحلیل داده‌ها و اطلاعات موجود در پایگاه‌های داده‌ها استفاده از روش‌هایی که نیاز به دانستن توزیع متغیرها ندارد از اهمیت خاصی برخوردار است. خوشه‌بندی یکی از روش‌هایی است که با توزیع داده‌های موجود سر و کار نداشته و اغلب با استفاده از معیارهای تشابه و عدم تشابه به خوشه‌بندی داده‌ها می پردازد (حائری مهریزی، ۱۳۸۲).

الگوریتم خوشه‌بندی تمام خوشه‌ها را براساس شباهت ها ایجاد می‌کند و در نهایت خوشه‌ها را بر اساس میزان شباهت به صورت پراکنده به نمایش می‌گذارد. فاصله بین خوشه‌ها اندازه گیری و به گونه ای چیده می‌شوند که خوشه‌هایی که به هم شبیه هستند نزدیکتر و فاصله زیاد به معنی حداقل شباهت بین داده‌های موجود در این خوشه‌ها می‌باشد. در واقع ابتدا خوشه‌ها براساس شباهت‌های بین مشخصه‌ها ایجاد و سپس چیدمان خوشه‌ها بر اساس فاصله آنها شکل می‌گیرد.

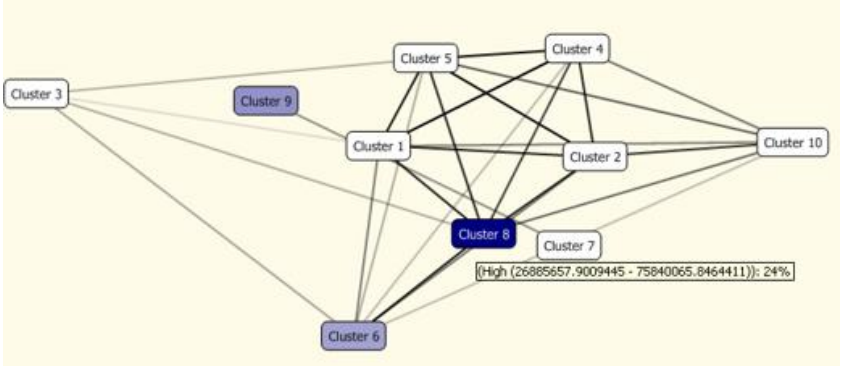
الگوریتم خوشه‌بندی مایکروسافت، ابتدا روابط را در یک مجموعه داده شناسایی و خوشه‌ها را بر اساس آن روابط تولید می‌نماید.

الگوریتم این پردازش را آنقدر تکرار می‌کند تا دیگر نتواند نتایج را با توصیف مجدد خوشه‌های بیشتر، بهبود بخشد.

همانطور که در ۵ قابل مشاهده است، علاوه بر مشاهده خوشه‌ها بر اساس فیلد قابل پیش‌بینی، امکان مشاهده میزان پراکندگی ریسک بیمه‌گذاران در خوشه‌های ده‌گانه نمایش داده شده است.

بیشترین بیمه‌گذاران به ترتیب در خوشه‌های ۱، ۲، ۳ و ۴ قرار دارند.

شکل ۶: نمودار خوشه‌بندی حاصل از پردازش داده‌های بیمه‌ای شخص ثالث



۱. Lift

درخت تصمیم

درخت تصمیم یکی از رایج‌ترین تکنیک‌های داده‌کاوی است. معمول ترین کاربرد درخت تصمیم، دسته‌بندی می‌باشد. الگوریتم درخت تصمیم مایکروسافت، تکنیک‌های دسته‌بندی و رگرسیون را پشتیبانی می‌نماید و برای مدل‌های پیشگویانه بسیار خوب عمل می‌کند. با استفاده از این الگوریتم، مشخصه های گسسته و پیوسته قابل پیش‌بینی هستند. هدف اصلی در درخت تصمیم، تقسیم داده‌ها به صورت بازگشتی به زیر مجموعه‌هایی است که هر زیر مجموعه در برگیرنده وضعیت همگنی از متغیر هدف می‌باشد. زمانی که پردازش بازگشتی کامل شد، درخت تصمیم شکل گرفته است.

در هر تقسیم درخت، تاثیر تمام مشخصه‌های ورودی بر روی مشخصه‌های قابل پیش‌بینی بصورت بازگشتی ارزیابی می‌شوند. به عبارتی دیگر در ساخت یک مدل، الگوریتم ارزیابی می‌کند که هر مشخصه ورودی در مجموعه داده‌ها چگونه در نتیجه مشخصه پیش‌بینی شده تاثیر می‌گذارد و سپس از این مشخصه های ورودی، قوی‌ترین رابطه را به منظور ساخت مجموعه‌ای از تقسیم‌ها که گره نامیده می‌شوند، استفاده می‌کند. گره‌های جدید به مدل اضافه می‌شوند و به این ترتیب یک ساختار درخت‌واره شکل می‌گیرد. گره بالای درخت تفکیک مشخصه‌های پیش‌بینی شده را در سطح کل جمعیت توصیف می‌نماید. هر گره‌ای که افزوده می‌شود، بر پایه توزیع وضعیت هایی از مشخصه های پیش‌بینی شده در مقایسه با مشخصه‌های ورودی ایجاد می‌شود.

فرمول پیشنهادی

فرمول پیشنهادی پژوهش برای محاسبه حق بیمه دریافتی از بیمه‌گذاران به شرح زیر می‌باشد:

فرمول۵: $P=E[A|X]+Sc+Sd+Pf(P)$

که در آن :

P: حق بیمه دریافتی از بیمه‌گذار

E[A|X]: میزان خسارت پیش‌بینی شده هر بیمه نامه با استفاده از داده‌کاوی

متغیر A میزان خسارت بیمه‌گذار به ازای شرایط X را نشان می‌دهد

Sc: هزینه‌های ثابت

Sd: هزینه‌های متغیر

Pf(P): سود بیمه‌گری

در فرمول فوق با استفاده از پرس‌وجو از مدل کاوشی پردازش شده، براساس پارامترهای ورودی مدل کاوش شامل اطلاعات بیمه‌نامه، بیمه‌گذار و اتومبیل، میزان خسارت (A) قابل پیش‌بینی می‌باشد.

سایر اجزای فرمول محاسبه حق بیمه، شامل: هزینه‌های ثابت و متغیر براساس هزینه‌های شرکت‌های بیمه قابل محاسبه می‌باشد و میزان سود بیمه‌گری براساس سیاست‌های کلان اقتصادی تبیین می‌گردد.

به عنوان نمونه خسارت برآوردی برای بیمه‌نامه‌ای خودرو پژو ۲۰۶ مدل ۱۳۸۶ با مسافت پیموده شده ۱۴۰ هزار کیلومتر برای آقایی ۵۷ ساله در شهرستان شیراز و یدون خسارت، در مدل کاوشی ۲۵۳۰۰۰۰ ریال می‌باشد که مبالغ هزینه‌های ثابت و متغیر و میزان سود بیمه‌گری به آن افزوده و در نتیجه حق بیمه خالص محاسبه می‌گردد.

نسبت خسارت داده‌های موجود در انبارداده‌های پژوهش به شرح زیر است:

فرمول شماره ۳ (سامانه سنهاب بیمه مرکزی):

$$\frac{۴۹۰۶۸۶۸۸.۸۹}{۶۸۹۴۳۶۱۸.۱۴} * ۱۰۰ = ۷۷.۱۷$$

نسبت خسارت در فرمول پیشنهادی پژوهش

فرمول شماره ۴ (حاصل از فرمول پیشنهادی):

$$\frac{۴۶۵۷۵۴۷۸.۴}{۷۱۲۶۰۴۸۱.۹۵} * ۱۰۰ = ۶۵.۳$$

نتیجه‌گیری

در این پژوهش بیش از ۳۰ میلیون بیمه‌نامه از سال ۱۳۸۵ تا ۱۳۹۳ جمع‌آوری، استانداردسازی و از سایر بانک‌های اطلاعاتی معتبر در سطح کشور از جمله بانک اطلاعاتی نیروی انتظامی و سازمان ثبت احوال، جهت اعتبارسنجی داده‌ها استفاده گردید. اطلاعات مربوط به خسارات این بیمه‌نامه‌ها نیز به در ساختار بانک اطلاعاتی رابطه‌ای به بیمه‌نامه‌های خود متصل گردید. اطلاعات مربوط به منطقه جغرافیایی بیمه‌نامه‌ها، بر اساس اطلاعات شبکه فروش شرکت‌های بیمه تکمیل گردید.

برخی از اطلاعات به خاطر عدم وجود در سامانه‌های اطلاعاتی سطح کشور، مانند میزان مسافت پیموده شده هر خودرو، بر اساس آمارهای موجود به صورت توزیع تصادفی فراوانی تهیه گردید. پس از پالایش، داده‌ها به انبارداده‌های طراحی شده که منبع اطلاعاتی مدل کاوشی است، انتقال پیدا کرد.

مدل کاوشی بر اساس فاکتورهای ریسک طراحی، مدل پردازش و با داده‌های آموزشی، با استفاده از الگوریتم‌های خوشه بندی، درخت تصمیم و شبکه عصبی آموزش داده شد. سپس از مدل‌های پردازش شده برای پیش‌بینی میزان خسارت‌های بیمه‌گذاران بر اساس فاکتورهای ریسک استفاده و نتایج بدست آمده با داده‌های آزمایشی مقایسه و از شاخص نسبت خسارت برای برآورد میزان صحت نتایج استفاده شد و به فرمول مناسبی برای تعرفه‌گذاری پویا رسیدیم.

با توجه به اجزا سازنده حق بیمه شامل: هزینه‌های ثابت، هزینه های متغیر، میزان سود و حق‌بیمه خالص تنها امکان محاسبه حق‌بیمه خالص بر اساس پیش‌بینی میزان خسارت هر بیمه‌گذار میسر می‌باشد. پس از محاسبه حق بیمه با فرمول ارائه شده و مقایسه میزان نسبت خسارت در حق‌بیمه‌های برآوردی و داده‌های پژوهش اعداد بدست آمده نشانگر بهبود وضعیت سودآوری شرکت‌های بیمه می‌باشد.

برای تحقیقات بعدی پیشنهاد می‌شود از الگوریتم‌های سری زمانی و وابستگی برای ادامه تحقیق و تکمیل پژوهش در حوزه تعرفه‌گذاری و کشف تقلبات غیر از شخص ثالث استفاده گردد.

منابع و ماخذ

رنجبرفرد، مینا، (۱۳۹۳)، "رضایت بیمه‌گذاران از خدمات بیمه شخص ثالث"، پژوهشنامه بیمه، سال بیست و نهم، شماره ۱، صص ۸۱.

دقیقی اصل، علیرضا، (۱۳۹۲)،"اولویت‌بندی سیستم‌های مختلف نظارت بر توانگری شرکتهای بیمه با استفاده از تکنیک فرایند تحلیل سلسله مراتبی (AHP)"، پژوهشنامه بیمه، سال بیست و هشتم، شماره ۱، صص ۱۰۹.

انصاری، علی، (۱۳۹۲)،"مسئولیت مدنی بیمه گذارنقض کننده حسن نیست در مرحله انعقاد قرارداد بیمه"،پژوهشنامه بیمه، سال بیست و هشتم، شماره ۱، صص ۱۱۱.

خرمی، فرهاد، (۱۳۸۲)، "معرفی روشی نظری – کاربردی برای اندازه‌گیری داده‌ها و ستاده‌ها و قیمت آنها در شرکت‌های بیمه"، فصلنامه صنعت بیمه، پژوهشکده بیمه، سال هیجدهم، شماره ۱، صص ۶۵

حائری مهریزی، علی اصغر، (۱۳۸۲)، داده کاوی: مفاهیم و روش‌ها و کاربردها، پایان‌نامه کارشناسی ارشد، دانشگاه علامه طباطبایی، تهران.

ایزدپرست، محمود، (۱۳۸۹)، ارائه چارچوبی برای پیش‌بینی سطح خسارت مشتریان بیمه بدنه اتومبیل با استفاده از راهکار داده‌کاوی، تهران: دانشگاه پیام‌نور، پایان نامه کارشناسی‌ارشد مدیریت فناوری اطلاعات.

اصلاح قانون بیمه اجباری مسؤولیت مدنی دارندگان وسایل نقلیه موتوری زمینی در مقابل شخص ثالث مصوب (۱۳۸۷) مجلس شورای اسلامی، اقتباس از کتاب مجموع قوانین و مقررات بیمه اجباری مسئولیت مدنی دارندگان وسایل نقلیه موتوری زمینی در مقابل شخص ثالث.

Higgs; B.; Dept; Jr., (۲۰۱۵), "Segmentation and Clustering of Car-Following Behavior: Recognition of Driving Patterns", Intelligent Transportation Systems, IEEE Transactions on (Volume:۱۶ , Issue:۱).

Dionne; G., (۲۰۰۵), Foundations of Insurance Economics, London: KAP.

Microsoft MSDN, June (۲۰۱۴), Data Mining Algorithms (Analysis Services - Data Mining) From: https://msdn.microsoft.com/en-us/library/ms۱۷۴۹۴۹.aspx.

Han; J., (۲۰۰۶), “Data Mining Concept and Techniques”, San Francisco.Morgan Kaufman.

Fayyad; U.M; Piatetsky Shapiro; Smyth P and thurusamy.(eds.),(۱۹۹۶), "Advances in Knowledge Discovery and Data Mining", Menlo Park, California, AAAI Press.

Hand; D.; Mannila; M.; Padhraic; S., (۲۰۰۱), Principle of Data Mining, MIT Press.

George; H. John, (۱۹۹۷), "Enhancements to the Data Mining Process", PhD Thesis, Computer Science Department, School of Engineering, Stanford University.

R.L. Grossman; C. Kamath; P. Kegelmeyer; V. Kumar, (۲۰۰۱), “Data Mining for Scientific and Engineering Applications (Massive Computing)”, Springer.

Tantan Liu; M.S, (۲۰۱۲), ”Data Mining over Hidden Data Sources”, The Ohio State University.

Renuka Panchagavi, (۲۰۱۲), “Survey of Electronic Health Records Data for Developing a Predictive Model of Pressure Ulcers in Critical Care Patients”, The Ohio State University.

Dake Zhanga; Kang Jianga, (۲۰۱۲), “Application of Data Mining Techniques in the Analysis of Fire Incidents”, International Symposium on Safety Science and Engineering in China.

Weimin Chen; Guocheng Xiang; Youjin Liu; Kexi Wang, (۲۰۱۲), “Credit risk Evaluation by hybrid data mining technique”, Hunan University of Science and Technology.

Chapados; N., (۲۰۱۰), Data Mining Algorithms for Actuarial Ratemaking, ApSTAT Technologies Inc, P. ۷.

Werner; G., (۲۰۱۰), Basic Ratemaking, Casualty Actuarial Society.

V.L. Miguéis, (۲۰۱۲), “Customer data mining for lifestyle segmentation”, Expert Systems with Applications.

Roung-Shiunn; Wua, (۲۰۱۱), “Customer segmentation of multiple category data in e-commerceusing a soft-clustering approach”, Electronic Commerce Research and Applications.

Chui-Yu Chiu, (۲۰۰۹), “An intelligent market segmentation system using k-means and particle swarm optimization”, Expert Systems with Applications.

Inna Kolyshkina; Richard Brookes, (۲۰۰۲), “Data mining approaches to modelling insurance risk”, PricewaterhouseCoopers.