به نام خدا

یک الگوریتم رگرسیون غیرخطی خوشه ای برای داده های بازه ای مقدار

مولفان :

بتول خوازه

زبیده خوازه

انتشارات ارسطو

(سازمان چاپ و نشر ایران – ۱۴۰۳)

نسخه الکترونیکی این اثر در سایت سازمان چاپ و نشر ایران و اپلیکیشن کتاب رسان موجود می باشد

chaponashr.ir

سرشناسه : خوازه، بتول، ۱۳۶۵
عنوان و نام پدیدآور : یک الگوریتم رگرسیون غیرخطی خوشه ای برای داده های بازه ای مقدار/ مولفان بتول خوازه، زبیده خوازه.
مشخصات نشر : انتشارات ارسطو (سازمان چاپ و نشر ایران)، ۱۴۰۳.
مشخصات ظاهری : ۱۰۸ ص.
شابک : ۵-۹۴۳-۳۳۹-۶۲۲-۹۷۸
وضعیت فهرست نویسی : فیپا
یادداشت : کتابنامه : ص ۱۰۸-۱۰۵ .
موضوع : الگوریتم رگرسیون غیرخطی خوشه ای - داده های بازه ای مقدار
شناسه افزوده : خوازه، زبیده، ۱۳۶۱
رده بندی کنگره : LC۴۷۲۸
رده بندی دیویی : ۳۷۱/۹۱۳۸۲
شماره کتابشناسی ملی : ۹۴۲۳۲۴۰
اطلاعات رکورد کتابشناسی : فیپا

نام کتاب : یک الگوریتم رگرسیون غیرخطی خوشه ای برای داده های بازه ای مقدار
مولفان : بتول خوازه - زبیده خوازه
ناشر : انتشارات ارسطو (سازمان چاپ و نشر ایران)
صفحه آرایی، تنظیم و طرح جلد: پروانه مهاجر
تیراژ : ۱۰۰۰ جلد
نوبت چاپ : اول - ۱۴۰۳
چاپ : زبرجد
قیمت : ۹۸۰۰۰ تومان
فروش نسخه الکترونیکی - کتاب‌رسان :
https://chaponashr.ir/ketabresan
شابک : ۵-۹۴۳-۳۳۹-۶۲۲-۹۷۸
تلفن مرکز پخش : ۰۹۱۲۰۲۳۹۲۵۵
www. chaponashr. ir

انتشارات ارسطو

چاپ و نشر ایران
Chaponashr.ir

فهرست مطالب

پیش‌گفتار: ۵
فصل ۱: مقدمات و پیش‌نیازها..... ۹
فصل ۲: روش‌های رگرسیون برای داده‌های بازه‌ای‌مقدار CRM و NLM ۲۵
فصل ۳:روش‌های‌خوشه‌بندی برای رگرسیون داده‌های بازه‌ای‌مقدار ۳۳
فصل ۴ :روش رگرسیون‌خوشه‌ای برای داده‌های بازه‌ای‌مقدار iCLR و iCNLR ۴۱
فصل ۵:تجزیه‌وتحلیل و نتایج تجربی ۵۲
مراجع ۱۰۰

پیش‌گفتار

پیشرفت‌های فن‌آوری اطلاعات، امکان جمع‌آوری و ذخیره حجم زیادی از داده‌ها را در بسیاری از زمینه‌ها مانند امور مالی، بازرگانی، زیست‌شناسی، پزشکی، شبکه‌های اجتماعی و غیره را فراهم کرده است. اطلاعات جمع‌آوری شده در این مجموعه‌داده بزرگ، ناهمگن، نادقیق و با تغییر زمان است. در حال حاضر، یک نیاز مهم برای تجمیع این مجموعه‌داده‌های بزرگ در ریزدانه‌های اطلاعاتی، وجود دارد که به‌طور همزمان تا آنجا که ممکن است اطلاعات را حفظ نموده و در عین حال می‌توانند نتایج را به راحتی توضیح داده و با شواهد تجربی پشتیبانی کنند. این ریزدانه‌های اطلاعاتی اغلب به عنوان داده‌های نمادین (سمبلیک) یا داده دانه‌ای[۱] نام‌گذاری می‌شوند. در دو دهه اخیر، تجزیه‌وتحلیل داده دانه‌ای رشد سریعی داشته است و اغلب از طریق تحلیل‌های بازه‌ای‌مقدار، مجموعه‌های فازی، مجموعه‌های تصادفی، مجموعه‌های ناهموار و مجموعه‌های پنهان در چارچوب‌های رسمی انجام می‌شود. [۴۰]. تجزیه‌وتحلیل داده‌های نمادین برای گسترش روش‌های داده‌کاوی و آمار اکتشافی برای مدیریت داده‌های نمادین استفاده می‌شود که اغلب توسط روش‌های آماری و چندمتغیره پشتیبانی می‌شوند [۱۸,۳]. مسائل منابع [۴۶,۲۸,۲۰] نیز گواه علاقه روزافزون به این زمینه‌ها است.

[۱] Granular data

تجزیه‌وتحلیل داده‌های کلاسیک، متغیرهای (عددی یا گروهی) معمولاً برای توصیف اهداف تک مقداری استفاده می‌شوند، به این معنی که برای یک هدف معین، یک متغیر نیز یک مقدار کمی یا کیفی واحد را می‌گیرد. با این‌حال، در بسیاری از موقعیت‌های حقیقی استفاده از متغیرهای تک مقداری می‌توانند بسیار محدودکننده باشد، به‌ویژه هنگام آنالیز گروهی به جای آنالیز فردی، که در آن تنوع ذاتی گروه باید در نظر گرفته شود. بنابراین، تجمیع مشاهدات تک‌مقداری امکان تولید انواع متغیرهای جدید مانند متغیرهای مجموعه‌مقدار، متغیرهایی با بازه‌ای‌مقدار یا حتی متغیرهایی با مقدار ستونی را فراهم می‌کند. در چارچوب آمار و آنالیز داده‌های چندمتغیره، تجزیه‌وتحلیل داده‌های نمادین (SDA)[1]با نوع خاصی از اطلاعات دانه‌ای که به عنوان داده‌های نمادین شناخته می‌شوند، سروکار دارد [۳,۱۸].

این پایان‌نامه بر روی روش‌های‌رگرسیون‌خوشه‌ای برای داده‌های بازه‌ای‌مقدار تمرکز دارد. این نوع داده‌ها می‌توانند عدم‌دقت و یا عدم‌قطعیت موجود در اندازه‌گیری خطا را نشان دهد، و می‌توانند تغییرپذیری و تنوع طبیعی داده‌ها را نیز نشان دهند. این پایان‌نامه فقط تغییرپذیری و تنوع طبیعی داده‌ها را بررسی می‌کند. داده‌های بازه‌ای‌مقدار در ثبت دماهای ماهانه در ایستگاه‌های هواشناسی، قیمت‌های روزانه سهام و غیره مورد استفاده قرار می‌گیرند. منبع دیگری از داده‌های بازه‌ای‌مقدار را می‌توان در تجمیع پایگاه داده‌های عظیم در تعداد کم‌تری از گروه‌ها، یافت که ویژگی‌های آن‌ها توسط متغیرهای بازه‌ای‌مقدار توصیف شده است. بنابراین، ابزارهایی برای تجزیه‌وتحلیل داده‌های بازه‌ای‌مقدار بسیار مورد نیاز است. این پایان‌نامه یک روش رگرسیون‌غیرخطی خوشه‌ای با مرکز و محدوده متغیرهای بازه‌ای‌مقدار را پیشنهاد می‌کند که از این پس در پایان‌نامه با نام اختصاری ICNLRعنوان خواهد شد (رگرسیون‌غیرخطی با مرکز و محدوده بازه‌ای‌مقدار). کارهای انجام شده در این پایان‌نامه به شرح زیر است:

[1] Symbolic Data Analysis

پیش‌گفتار

این پایان‌نامه، یک روش رگرسیون غیرخطی خوشه‌ای با مرکز و محدوده را برای متغیرهای بازه‌ای‌مقدار پیشنهاد می‌کند که در این پایان‌نامه بنام (iCLR)[۱] (رگرسیون‌غیرخطی خوشه‌ای با مرکز و محدوده بازه‌ای‌مقدار) نام‌گذاری شده است. روش پیشنهادی ترکیبی از روش‌های رگرسیون‌خطی[۲](CRM)[۱۴]، روش غیرخطی[۳] (NLM) [۳۹] و روش رگرسیون برای داده‌های بازه‌ای‌مقدار با یک الگوریتم خوشه‌بندی‌پویا[۴] [۱۹] است. این یک الگوریتم خوشه‌بندی مرتبط با الگوریتم میانگین k است که هدف آن شناسایی افراز داده‌ها و مدل‌های‌رگرسیون مربوطه (در مرکز و محدوده بازه‌ها) به‌طور همزمان یکی برای هر خوشه است. روش جدید پیشنهادی (iCLR) در مقایسه با روش iCLR دی‌کاروالو[۵] و همکاران [۱۴]، قادر است تا افراز داده‌های بازه‌ای‌مقدار را به تعداد ثابتی از خوشه‌ها انجام دهد. همچنین می‌توانند بهترین جفت توابع‌خطی و/یا غیرخطی را در مرکز و محدوده بازه‌ها برای هر خوشه (از بین مجموعه‌ای از مدل‌های‌غیرخطی (خطی)) انتخاب کند.

بقیه پایان‌نامه به شرح زیر سازماندهی شده است:

در فصل ۲، روش رگرسیون برای داده‌های بازه‌ای‌مقدار CRM و NLM بیان شده است که به ترتیب در بخش‌های ۲,۲ و ۳,۲، روش‌های خطی (CRM) و غیرخطی (NLM) برای داده‌های بازه‌ای‌مقدار بیان گردیده است. در فصل ۳، روش‌های خوشه‌بندی برای رگرسیون داده‌های بازه‌ای‌مقدار ارائه شده است.

در فصل ۴، روش رگرسیون خوشه‌ای برای داده‌های بازه‌ای‌مقدار iCLR و iCNLR آورده شده است. ابتدا در بخش ۲,۴، روش iCLR ارائه و سپس در بخش ۳,۴، روش پیشنهادی برای رگرسیون غیرخطی خوشه‌ای داده‌های بازه‌ای‌مقدار ارائه گردیده است که این الگوریتم پیشنهادی با یک مقداردهی اولیه آغاز می‌شود و دو مرحله را انجام می‌دهد که مرحله اول افراز را در تعداد خوشه‌های از پیش تعریف‌شده به‌روزرسانی می‌کند، و

---

[۱] Interval center and range clusterwise linear regression
[۲] center and range linear regression method
[۳] nonlinear method
[۴] Dynamic clustering algorithm
[۵] De Carvalho

مرحله دوم با مدل‌های رگرسیون خطی/غیرخطی در مرکز و محدوده بازه‌ای‌مقدار، یکی برای هر خوشه مطابقت دارد. سپس، در پیوست اثبات همگرایی[1] الگوریتم iCNLR نشان داده شده است. فصل ۵، بحثی را در مورد آزمایش‌های انجام شده با داده‌های بازه‌ای‌مقدار مصنوعی ارائه می‌کند که در ادامه، قابلیت‌های برآورد و پیش‌بینی روش پیشنهادی را برای طرح‌های مختلف نشان می‌دهد. سپس، عملکرد پیش‌بینی روش پیشنهادی را در مجموعه‌داده‌های با مقادیر حقیقی ارائه خواهد داد. سرانجام، نتیجه‌گیری و برخی کارهای تحقیقاتی آتی ارائه شده است.

[1] Convergence

فصل ۱

مقدمات و پیش‌نیازها

۱.۱ مقدمه

ابتدا در این فصل به طور مختصر به بررسی مفاهیم مقدماتی و اساسی می‌پردازیم. در طول نگارش این فصل از دو منبع [۵۳] و [۵۶] استفاده شده است.

۲.۱ رده‌بندی‌داده‌ها

رده‌بندی داده‌ها فرآیندی است که در دو گام انجام می‌شود:

۱- گام یادگیری که در آن یک مدل یا الگوریتم رده‌بند ساخته می‌شود. در این گام که با نام گام آموزشی نیز شناخته می‌شود، به کمک تحلیل و یادگیری از مجموعه داده‌های آموزشی، یک الگوریتم رده‌بندی ایجاد می‌شود.

۲- گام رده‌بندی که در آن از مدل ساخته شده در گام یادگیری (آموزشی)، برای پیش‌گویی برچسب‌های رده استفاده می‌شود.

۱.۲.۱ مجموعه‌داده‌های آموزشی و آزمایشی

۱- مجموعه داده‌هایی که برای گام یادگیری از آن‌ها استفاده می‌شود و شامل داده‌ها و برچسب‌های رده‌ای متناظر با آن است، مجموعه‌ی آموزشی نام دارد. مجموعه آموزشی معمولاً به صورت تصادفی از پایگاه داده‌های مورد نظر انتخاب می‌شود.

۲- مجموعه داده‌هایی که برای بررسی صحت و درستی الگوریتم رده‌بندی از آن استفاده می‌شود و مستقل از مجموعه‌ی آموزشی است و بدان معنی است که از آن‌ها برای ساخت الگوریتم رده‌بندی استفاده نشده است.