AICODE.IR

Artificial Intelligence & Computer Engineering Reference

عنوان

تکنیکهای داده‌کاوی در تحلیل داده‌های سری زمانی و پیش‌بینی

Data Minning Techniques in Time Series Data Analysis and Forecasting

چکیده

در بسیاری از کاربردهای واقعی، مقدار مشاهدات یک متغیر وابسته به مقدار آن متغیرهای در زمان‌های قبل است. اینگونه داده‌ها با عنوان سری زمانی مطرح هستند که پژوهش‌های مختلفی با استفاده از تکنیک‌های داده‌کاوی به پیش‌بینی و تحلیل در اینگونه داده‌ها پرداخته‌اند. در اینجا به دسته‌بندی انواع سری زمانی، کاربردها، انواع مسایل سری زمانی و نحوه حل آنها پرداخته‌ایم.

واژه‌های کلیدی:

سری زمانی، الگوی فصلی، الگوی افقی، الگوی متمایل، پیش‌بینی، رگرسیون، شبکه عصبی بازگشت‌پذیر، رگرسیون بردار پشتیبان.

فهرست عناوین

صفحه

فهرست اشکال

صفحه

۱

مقدمه

مقدمه

در این پژوهش هدف بررسی روش‌هایی است که برای تحیل و پیش‌بینی سری‌های زمانی به کار رفته است. در این فصل به تعریف مساله، انواع سری زمانی بر اساس الگو، کاربردها و چالش‌ها می‌پردازیم.

۱.۱ تعریف سری زمانی

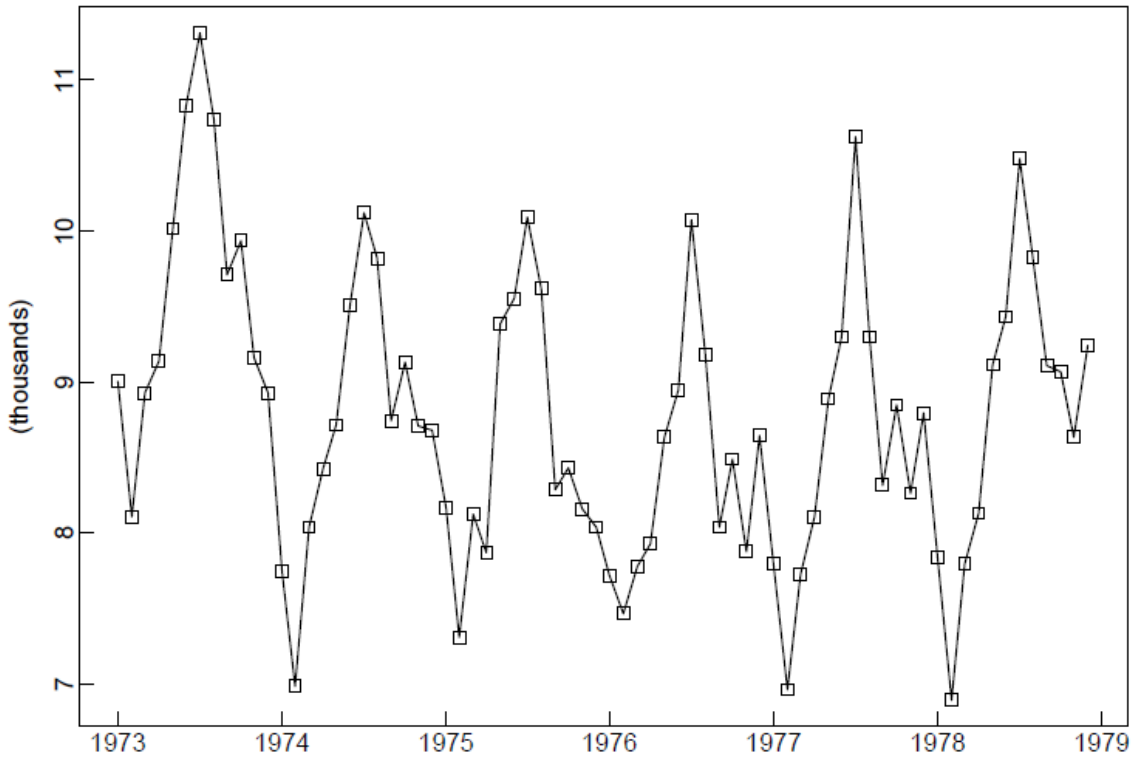
یک سری زمانی مجموعه‌ای از مشاهدات به صورت x_t می‌باشد که هرکدام از این مشاهدات در زمان مشخص t انجام شده است. در صورتی که فاصله زمانی مشاهدات مساوی باشد به آن سری زمانی گسسته و در غیر اینصورت به آن سری زمانی پیوسته می‌گویند.

هدف پیش‌بینی در سری‌های زمانی گاهی پیش‌بینی مقدار x در زمان‌هایی است که هنوز اتفاق نیفتاده است. انواع مسایل سری زمانی و نحوه برخورد با آنها در فصل دوم توضیح داده شده است.

۱.۱.۱ مثال سری زمانی: تعداد مرگ‌ها در تصادفات ماهانه

شکل ۱-۱ تصادفات ماهانه در سال‌های ۱۹۷۳-۱۹۷۸ در آمریکا را نشان می‌دهد [۱]. همانطور که مشاهده می‌کنید یک الگوی تکرارشونده فصلی[1] در طول سال قابل مشاهده است و در اواسط هر سال یک پیک داریم.

[1] Seasonal Pattern



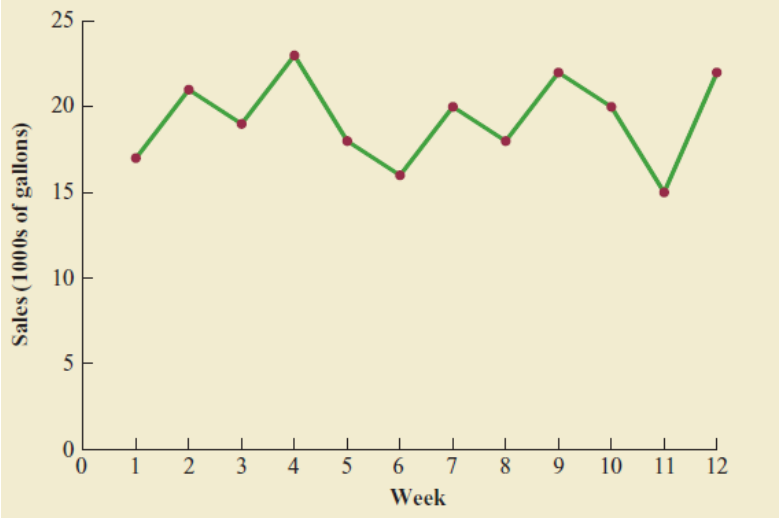
شکل ۱-۱ تعداد مرگ‌ها در تصادفات ماهانه در سال‌های ۱۹۷۳-۱۹۷۸

یکی از نکات مهم در پیش‌بینی سری زمانی در نظر گرفتن تمام ویژگی‌های مرتبط با مشاهدات است. برای مثال در مثال قبل ماه یک متغیر موثر در تصادفات است که باید به گونه‌ای در مدل‌سازی در نظر گرفته شود. در واقع سری زمانی بالا رفتار فصلی و تکرار شونده دارد.

۲.۱ انواع الگوهای سری زمانی

الگوی مشاهدات در یک سری زمانی را می‌توان به صورت زیر دسته‌بندی کرد [۲]:

۱. الگوی افقی[۲]: در این نوع الگو مشاهدات حول یک مقدار میانگین نوسانات تصادفی دارد (شکل ۱-۲).



شکل ۱-۲ مثالی از یک الگوی افقی. نمودار فروش بنزین

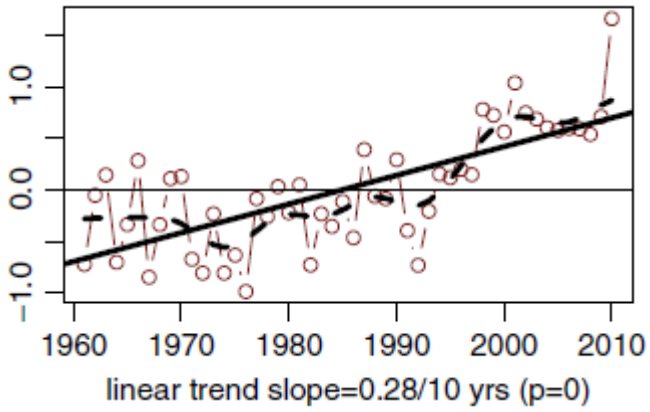
در صورتی که نوسات حول یک مقدار میانگین واقعا تصادفی باشد بهترین مدل برای برخود با این مشاهدات مدل کردن آنها به صورت یک توزیع گوسی با میانگین و یک واریانس است که از روی مشاهدات تخمین می‌زنیم. در واقع این نوع الگو رفتار ساده‌ای دارد که نیازی به مدل‌های پیچیده برای پیش‌بینی در آنها نیست.

۲. الگوی متمایل[۳]: یک سری زمانی اگرچه عموما شامل نوسانات تصادفی می‌باشند، بعضی از آنها به صورت تدریجی متمایل به حرکت به سمت مقادیر بالاتر یا پایین‌تر می‌باشند. این نوع رفتار را

[2] Horizontal Pattern

[3] Trend Pattern

الگوی متمایل می‌گویند. این نوع الگو متاثر از عوامل بلند مدت[۴] همچون افزایش یا کاهش جمعیت، پیشرفت تکنولوژی، تغییر تمایلات مصرف‌کنندگان می‌باشد. نمونه‌ای از این نوع سری زمانی را که مربوط به دمای یک منطقه می‌باشد را در شکل ۱-۳ می‌بینند [۳].



شکل ۱-۳ حداکثر دمای منطقه Arab در بین سال‌های ۱۹۶۰ تا ۲۰۱۰

۳. الگوی فصلی: در اینگونه سری‌های زمانی الگویی از مشاهدات به صورت وابسته به فصل (و یا هفته و ...) به صورت تناوبی تکرار می‌شوند.

۴. الگوهای ترکیبی: یک الگوی سری زمانی می‌تواند برای مثال ترکیبی از تمایل به سمت افزایش و فصلی باشد برای مثال میزان تصادفات هم وابسته به فصل و هم افزایش تعداد وسایل نقلیه است.

[4] Long Term

۳.۱ اهمیت موضوع و کاربردها

در بسیاری از کاربردهای واقعی همچون پیش‌بینی وضعیت آب و هوا، تعداد فروش کالا، قیمت سهام، پیش‌بینی معدل دانشجویان و ... مشاهدات متغیرهای وابسته‌ای از زمان می‌باشند. برای مثال یک واحد تجاری با پیش‌بینی میزان تقاضای کالا در هر دوره می‌تواند در آن جهت برنامه‌ریزی مناسبی انجام دهد. گاهی با تحلیل یک سری زمانی همچون نمودار تغییرات معدل دانشجویان می‌توان پیش‌بینی‌ها و برنامه‌ریزی‌های مناسبی انجام داد. یکی دیگر از کاربردهای تحلیل سری زمانی یافتن متغیرهای تاثیرگذار در یک پدیده زیست محیطی مثل دمای هوا است که می‌توانند معیارهای بسیار مهمی برای جلوگیری از پدیده‌های مخرب زیست محیطی شود. پژوهشگران در [۳] نشان داده‌اند که دمای هوا در منطقه عرب ارتباط بسیار زیادی با نوسانات اطلس شمالی دارد. از نتایج این پژوهش [۳] می‌توان برای پیش‌بینی دوا و وضع هوا در این منطقه استفاده کرد.

۴.۱ چالش‌ها

از چالش‌های مطرح در زمینه سری زمانی وجود مقادیر جاافتاده[5] در مجموعه داده آموزش است. معمولا در بعضی بازه‌های زمانی به علت مشکلاتی مقدار متغیر x یا وجود نداشته و یا اندازه گیری نشده است. برای اینکه مدل یادگیر ما بتواند از اطلاعات به خوبی استفاده کند، باید راهکاری برای حل این چالش داشته باشد. از دیگر چالش‌ها درنظر گرفتن تمام ویژگی‌های موثر در مدل‌سازی است.

[5] Missing Value

۲

تکنیک‌های داده‌کاوی در تحلیل داده‌های سری زمانی

۱.۲ انواع مسایل در سری زمانی و روش‌های موجود در آنها

دو نوع مساله متداول در سری زمانی وجود دارد:

۱. تشخیص نوع سری زمانی (دسته‌بندی سری زمانی)

۲. تحلیل و پیش‌بینی سری‌های زمانی

۱.۱.۲ تشخیص نوع سری زمانی

در این نوع مساله هدف دسته‌بندی نوع الگوی داده‌های سری زمانی است. در این نوع مسایل "دنباله سری زمانی" ورودی مدل نهایی است که به یکی از دسته های مورد نظر دسته‌بندی می‌شود. این مساله در اکثر پژوهش‌ها با عنوان "تشخیص الگوی چارت کنترل[۶]" کار شده است [۴، ۵]. برای مثال پژوهشگران در [۴] با استفاده از شبکه عصبی SOM الگوهای چارت کنترل را براساس الگوی موجود در آنها به ۶ دسته تقسیم بندی کرده‌اند. از مجموعه داده‌های در این زمینه می‌توان به Synthetic Control Chart Time Series Data Set[۷] اشاره کرد. این مجموعه داده شامل ۶۰۰ الگو از ۶ دسته زیر می‌باشد.

1. Normal
2. Cyclic
3. Increasing trend

[6] Control-Chart Pattern-Recognition

[7] https://archive.ics.uci.edu/ml/datasets/Synthetic+Control+Chart+Time+Series

4. Decreasing trend
5. Upward shift
6. Downward shift

هرکدام از داده‌ها یک دنباله

از روش‌های داده‌کاوی موجود در پژوهش‌های جدید برای دسته‌بندی داده‌های چارت کنترلی می‌توان به ماشین بردار پشتیبان وزن‌دار [۶]، شبکه‌های عصبی [۷] و سیستم‌های استنتاج نروفازی [۸] اشاره کرد.

۲.۱.۲ تحلیل و پیش‌بینی سری‌های زمانی

در مسایل پیش‌بینی سری زمانی، هدف پیش‌بینی مشاهده بعدی در دنباله مشاهدات سری زمانی است. معیار ارزیابی که در این مسایل کاربرد دارد معیار میانگین مربعات خطا[۸] است که اختلاف مقدار واقعی با مقدار پیش‌بینی شده را اندازه‌گیری می‌کند [۲]. از مدل‌های مختلف رگرسیون و اتو رگرسیون[۹] (AR) به منظور مدل پیش‌بینی کننده استفاده می‌شود. تفاوت رگرسیون در مسایل سری زمانی این است که برای پیش‌بینی مشاهدات از مقدار مشاهدات در زمان‌های قبل استفاده می‌کنیم. در مسایل رگرسیون معمولی هدف پیش‌بینی متغیر وابسته y برحسب ورودی x است.

فرض اصلی در رگرسیون سری زمانی این است که مشاهدات با نقاط قبل از خود همبستگی دارند و هدف تخمین این خود همبستگی است. هدف از رگرسیون تخمین تابعی است که زمان را به عنوان یک

[8] Mean Square Error

[9] Auto Regression

متغیر مستقل دریافت و مقدار مشاهده را تخمین می‌زند. انواع توابع خطی، مکعبی و ... را می‌توان فرض کرد [۲].

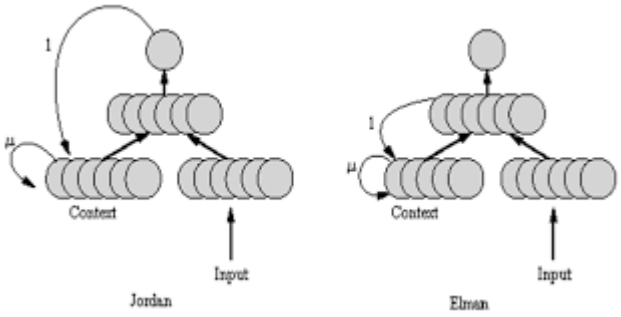
دو نوع مدل رگرسیون محلی و سراسری را می‌توان برای داده‌های سری زمانی به کار برد. در مورد رگرسیون محلی مدل رگرسیون را به صورت محلی برای پیش‌بینی آموزش می‌دهیم. پژوهشگران در [۹] مدل رگرسیون بردار پشتیبان محلی را برای پیش‌بینی سری‌های زمانی مالی مدنظر قرار داده‌اند.

یافتن متغیرهای وابسته به یک مشاهده سری زمانی نیز یکی از رویکردهای مهم در زمینه پیش‌بینی بهتر یک متغیر سری زمانی است. در بسیاری از کاربردها علاوه بر اینکه مقدار متغیر مشاهده شده را در یک دنباله زمانی داریم تعدادی ویژگی نیز از زمان جاری در دسترس است. برای مثال در پیش بینی معدل ترم جاری یک دانشجو علاوه بر دنباله معدل‌های ترم قبل، ویژگی‌هایی هم چون تعداد واحد عمومی اخذ شده، تعداد واحد تخصصی اخذ شده، تعداد کل واحدها و ... به عنوان ویژگی‌های هر ترم را داریم که مدل‌سازی باید به گونه‌ای باشد که هم ویژگی‌ها و هم دنباله تغییرات را در نظر بگیرید. برای مثال پژوهشگران در [۹] متغیر پیش‌بینی یعنی y_t را وابسته‌ای به مقادیر قبلی و ویژگی‌های آن لحظه در نظر گرفته اند.

$$\hat{y}_t = f(y_{t-1}, y_{t-2}, y_{t-3}, y_{t-11}, y_{t-12}, y_{t-13}, y_{t-24}, x_{t,8}, x_{t,9}), \qquad (17)$$

where $x_{t,8}$, $x_{t,9}$ are the month and order of y_t respectively, namely $x_{t,8} \in \{1,2,\ldots,12\}$, $x_{t,9} \in \{1,2,\ldots,96\}$. For D1-SP and D2-SP, we set

از مدل‌های شبکه عصبی بازگشتی[10] به شدت در زمینه پیش‌بینی متغیرهای سری زمانی استفاده شده است. روش‌های شبکه عصبی بازگشتی هم‌چون المن، جردن و المن-جردن از روش‌های مطرح در این زمینه می‌باشند [۱۰]. در این مدل‌ها خروجی شبکه (متغیر سری زمانی) در کنار ویژگی‌های دیگر با تاخیرهای مختلف به ورودی شبکه عصبی وارد می‌شود. شکل زیر یک شبکه عصبی جردن را نشان می‌دهد. همانطور که در شکل ۲-۱ مشاهده می‌شود در مدل جردن متغیرهای خروجی با تاخیر به عنوان ورودی به شبکه داده می‌شود (در کنار ویژگی‌های وضعیت جاری).



شکل ۲-۱ شبکه عصبی المن و جردن

از دیگر تکنیک‌های داده کاوی که در زمینه پیش‌بینی سری‌های زمانی استفاده شده است می‌توان به الگوریتم ژنتیک و درخت تصمیم نیز اشاره کرد [۱۱]. پژوهشگران در [۱۱] از درخت تصمیم برای یادگیری الگوها در داده‌های قبلی و انتخاب بهترین متد پیش‌بینی استفاده کرده‌اند. در واقع آنها از درخت تصمیم برای انتخاب بهترین متد پیش‌بینی بر اساس ویژگی‌های سری زمانی استفاده کرده‌اند.

[10] Recurrent Neural Networks

۳ منابع و مراجع

مقالات اصلی:

[1] Brockwell, Peter J., and Richard A. Davis. *Introduction to time series and forecasting*. Springer Science & Business Media, 2006.

[2] Time Series Analysis and Forecasting Chapter 15.

[3] Donat, M. G., T. C. Peterson, M. Brunet, A. D. King, M. Almazroui, R. K. Kolli, Djamel Boucherf et al. "Changes in extreme temperature and precipitation in the Arab region: long-term trends and variability related to ENSO and NAO."*International Journal of Climatology* 34, no. 3 (2014): 581-592.

[4] Pham, D. T., and A. B. Chan. "Control chart pattern recognition using a new type of self-organizing neural network." *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering* 212, no. 2 (1998): 115-127.

[5] Hachicha, Wafik, and Ahmed Ghorbel. "A survey of control-chart pattern-recognition literature (1991–2010) based on a new conceptual classification scheme." Computers & Industrial Engineering 63, no. 1 (2012): 204-222.

[6] Xanthopoulos, Petros, and Talayeh Razzaghi. "A weighted support vector machine method for control chart pattern recognition." Computers & Industrial Engineering 70 (2014): 134-149.

[7] El Farissi, O., A. Moudden, and S. Benkachcha. "Using Artificial Neural Networks for Recognition of Control Chart Pattern." *International Journal of Computer Applications* 116, no. 3 (2015).

[8] Nikpey, Abdolhakim, Somayeh Mirzaei, Masoud Pourmandi, and Jalil Addeh. "Identification of the Control Chart Patterns Using the Optimized

Adaptive Neuro-Fuzzy Inference System." *International Journal of Modern Education and Computer Science (IJMECS)* 6, no. 7 (2014): 16.

[9] Jiang, Hui, and Wenwu He. "Grey relational grade in local support vector regression for financial time series prediction." *Expert Systems with Applications* 39, no. 3 (2012): 2256-2262.

[10] Li, Penghua, Yinguo Li, Qingyu Xiong, Yi Chai, and Yi Zhang. "Application of a hybrid quantized Elman neural network in short-term load forecasting."*International Journal of Electrical Power & Energy Systems* 55 (2014): 749-759.

[11] Gerdes, Mike. "Decision trees and genetic algorithms for condition monitoring forecasting of aircraft air conditioning." *Expert systems with applications* 40, no. 12 (2013): 5021-5026.