یازدهمین سمینار بین‌المللی
جبر خطی و کاربردهای آن
بهمن ۱۴۰۰

کاربرد تجزیه مقادیر منفرد تصادفی در تحلیل مولفه‌های اصلی به منظور افزایش سرعت آموزش در روش‌های یادگیری ماشین

زهرا بهروزه، [۱،*] مسعود حجاریان، [۲] علیرضا افضل آقائی نائینی [۱]
[۱] گروه علوم داده ها و کامپیوتر، دانشکده ریاضی، دانشگاه شهید بهشتی، تهران، ایران
[۲] گروه علوم ریاضی، دانشکده ریاضی، دانشگاه شهید بهشتی، تهران، ایران

چکیده

الگوریتم‌های یادگیری ماشین بر اساس مجموعه داده مسئله، یک مدل ریاضیاتی ایجاد کرده و پس از آموزش مدل، به پیش‌بینی داده‌های جدید می‌پردازند. افزایش ابعاد مجموعه داده، یادگیری این مدل‌ها را با مشکلاتی مواجه کرده است. تحلیل مولفه‌های اصلی یکی از راهکارهای رفع این مشکل می‌باشد. این الگوریتم بر اساس تجزیه مقادیر منفرد با ترکیب ویژگی‌های مجموعه داده یک تقریب از فضای اصلی مسئله ارائه می‌کند. محاسبه این تجزیه در مسائل با ابعاد بالا خود دارای مشکلاتی است. در این تحقیق پس از معرفی تجزیه مقادیر منفرد تصادفی به بررسی کاربردهای این الگوریتم در افزایش سرعت یادگیری مدل ماشین بردار پشتیبان در چند مجموعه داده شناخته شده می‌پردازیم.

واژه‌های کلیدی: یادگیری ماشین، تحلیل مولفه‌های اصلی، تجزیه مقادیر منفرد تصادفی، ماشین بردار پشتیبان
کد موضوع‌بندی ریاضی [۲۰۱۰]: 15A23، 68T10

۱ مقدمه

آرتور ساموئل در سال ۱۹۵۹ لفظ یادگیری ماشین را ابداع نمود. او معتقد بود یادگیری ماشین، حوزه‌ای از تحقیقات است که در آن کامپیوترها بدون برنامه‌نویسی توانایی یادگیری دارند. واقعیت امر به این صورت است که دانش یادگیری ماشین سعی بر این دارد تا با استفاده از الگوریتم‌ها، یک مدل را به گونه‌ای طراحی کند که این مدل بتواند آموزش ببیند و به درستی عمل کند. یادگیری ماشین به دسته‌های مختلفی همچون یادگیری نظارت شده، یادگیری بدون نظارت و یادگیری نیمه نظارتی تقسیم می‌شود. در این روش‌ها رایج است که یک مجموعه داده، مشخص شده و سپس الگوریتم‌هایی جهت یادگیری الگوهای پنهان در میان این داده‌ها استفاده می‌شود. این مجموعه شامل تعدادی نمونه بوده که هر کدام از آن‌ها دارای ویژگی‌هایی هستند. بدین ترتیب این مجموعه داده دارای دو شاخصه مهم، به نام تعداد نمونه‌ها و تعداد ویژگی‌ها می‌باشد.تعداد ویژگی در برخی مسائل بسیار بالا بوده و برای یافتن الگو در این داده‌ها نیازمند الگوریتم‌های خاصی هستیم تا بتوانیم مشکلاتی همچون کمبود حافظه و پیچیدگی محاسباتی را کاهش دهیم. علاوه بر مشکلات محاسباتی، در این مجموعه داده‌ها پدیده مشقت بُعدچندی یا نفرین ابعاد نیز ظاهر می‌گردد. این پدیده بیانگر آن است که با افزایش ابعاد، حجم فضا آنقدر سریع افزایش می‌یابد که داده‌های موجود پراکنده و تُنُک می‌شوند. همچنین با افزایش ابعاد لازم است داده‌های مورد نیاز به‌طور نمایی افزایش یابند تا نتیجه حاصله از نظر آماری معقول و معتبر باشد. برای رفع این مشکل دانشمندان بر اساس رویکردهای نظارت شده و بدون نظارت ایده‌های مختلفی جهت کاهش تعداد ویژگی‌ها ارائه کرده‌اند. برای حل این مشکل دانشمندان دو رویکرد استخراج و انتخاب ویژگی را ارائه دادند. در حالی که انتخاب ویژگی هدف انتخاب زیرمجموعه‌ای از ویژگی داده‌ها را دارد، استخراج ویژگی با ترکیب ویژگی‌های موجود، ویژگی‌های جدید را تولید می‌کند. در تمامی روش‌های انتخاب و استخراج ویژگی، رویکرد های جبرخطی به کاربرده می‌شود. روش تحلیل مولفه‌های اصلی یکی از پرکاربردترین روش‌های استخراج ویژگی می‌باشد. این رویکرد با یافتن یک پایه متعامد جدید برای فضای ویژگی‌ها، تعداد ویژگی‌های مدل را کاهش می‌دهد. یافتن این پایه متعامد به کمک روش تجزیه مقادیر منفرد انجام می‌گردد. تجزیه مقادیر منفرد یک رویکرد ماتریسی است که ماتریس را به سه ماتریس دیگر تجزیه

*سخنران. آدرس ایمیل: zahrabehrouzeh@gmail.com

می‌کند. محاسبه این تجزیه برای مسائل در مقیاس بزرگ از نظر زمانی بهینه نیست. الگوریتم تجزیه مقادیر منفرد تصادفی جهت رفع این مشکل ارائه شده است. در ادامه به بررسی روش‌های تحلیل مولفه‌های اصلی، تجزیه مقادیر منفرد و تجزیه مقادیر منفرد تصادفی می‌پردازیم و کاربرد تجزیه مقادیر منفرد تصادفی در تحلیل مولفه‌های اصلی برای چند مجموعه داده مختلف را به کمک ماشین بردار پشتیبان بررسی می‌نماییم.

۲ پیش نیازها

در این قسمت به بررسی مفاهیم به کاربرده شده در این پژوهش از جمله تحلیل مولفه‌های اصلی، تجزیه مقادیر منفرد، تجزیه مقادیر منفرد تصادفی و ماشین بردار پشتیبان می‌پردازیم.

۱.۲ تحلیل مولفه‌های اصلی (PCA)

تحلیل مولفه‌های اصلی یکی از روش‌های پرکاربرد استخراج ویژگی است. در این رویکرد تلاش می‌شود به کمک ترکیب خطی ویژگی‌های موجود، ویژگی‌های جدیدی تولید شود. این شیوه با یافتن جهت‌های بیشترین واریانس در داده‌ها، یک دستگاه مختصات جدید ارائه می‌کند. در این دستگاه جدید اولین بُعد متناظر با جهت بیشترین واریانس و اخرین بُعد بیانگر جهت کمترین واریانس در میان داده‌ها است. از این رو با انتخاب k بُعد ابتدای این دستگاه می‌توان ویژگی های استخراج شده و مفید کمتری بدست آورد. در ادامه به شرح این روش می‌پردازیم . [۴]

قضیه ۱.۲. *فرض کنید مجموعه داده X با ابعاد $n \times d$ با میانگین ستونی صفر داده شده است. در این صورت جهت‌های بیشترین واریانس بردارهای داده‌ها، توسط بردارهای ویژه ماتریس کواریانس ارائه می‌شود.*

اثبات. با فرض صفر بودن میانگین ویژگی‌های مجموعه داده، ماتریس کواریانس به صورت

$$C = \frac{1}{n}\sum_{i=0}^{n} x_i x_i^T$$

نمایش داده خواهد شد. فرض کنید بردار u_1 جهت بیشترین واریانس داده ها را نمایش دهد. برای یافتن این بردار مسئله بهینه‌سازی زیر را تشکیل می‌دهیم.

$$\begin{array}{ll} \max\limits_{u_1} & \frac{1}{n} u_1 C u_1 \\ \text{s.t.} & \|u_1\| = 1 \end{array} \tag{1}$$

محاسبه فرم دوگان ۱ مسئله مقدار ویژه‌ی زیر را نتیجه می‌دهد.

$$C \cdot u_1 = \lambda \cdot u_1 \tag{2}$$

بنابراین برای محاسبه جهت بیشترین واریانس، بزرگترین مقدار ویژه و بردارهای ویژه متناظر آن از ماتریس کورایانس انتخاب می‌شوند. به منظور یافتن دومین جهت بیشترین واریانس، دومین بزرگترین مقدار ویژه محاسبه میگردد. لازم به ذکر است محاسبه مسئله مقدار ویژه ماتریس کواریانس برابر با تجزیه مقادیر منفرد ماتریس داده X می‌باشد. [۴] □

۲.۲ تجزیه مقادیر منفرد (SVD)

در بسیاری از مسائل یادگیری ماشین ما با ماتریس‌هایی سر و کار داریم که دارای سطر و ستون‌های بسیاری می‌باشند در واقع انجام عملیات برروی آن‌ها بسیار هزینه بالای دارد. روش تجزیه مقادیر منفر در جبر خطی پاسخی برای ساده‌سازی این مشکل می‌باشد. تجزیه مقادیر منفرد ماتریس A به روشی گفته می‌شود که ماتریس مذکور را به سه ماتریس بالا مثلثی، قطری و پایین مثلثی تجزیه می‌کند. [۱]حال به شرح مختصر این روش می‌پردازیم. فرض کنید ماتریس $A_{m\times n}$ را داریم و مقادیر m, n بسیار بزرگ می‌باشند. آنگاه تجزیه مقادیر منفرد ماتریس A به صورت زیر است:

$$A_{m\times n} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}$$

U یک ماتریس $m \times m$ متعامد، Σ یک ماتریس قطری $m \times n$ و V^T یک ماتریس $n \times n$ می‌باشد. از آن‌جایی که ماتریس U و V متعامد می‌باشند داریم:

$$U^T U = V V^T = I$$

و هم‌چنین به سادگی ثابت می‌شود که مقادیر روی قطر اصلی Σبرابر است با مقادیر ویژه ماتریس $A^T A$ درنتیجه با معلوم بودن ماتریس Σو A می‌توان U،V را به سادگی با جای‌گذاری در فرمول‌های زیر محاسبه نمود :

$$A^T A V = V \Sigma^2 \qquad\qquad A A^T U = U \Sigma^2$$

۳.۲ تجزیه مقادیر منفرد تصادفی

رویکرد اصلی در روش تجزیه مقادیر منفرد تصادفی. [۳] به این صورت می‌باشد که ماتریس با ابعاد بزرگ‌تر را به ماتریس با ابعاد کوچک‌تر تقریب می‌زنیم و سپس این ماتریس تقریب را به کمک SVD تجزیه می‌کنیم و در نهایت مقدار SVD ماتریس بزرگتر را به کمک SVD ماتریس کوچک تقریب می‌زنیم. ماتریس $A_{m\times n}$ را درنظر بگیریم. می‌خواهیم ماتریس‌های $B_{m\times k}$ و$C_{k\times n}$ را به شرطی تعریف کنیم که $k \ll n$ و A≈BC باشند. برای تقریب این محاسبه با استفاده از الگوریتم های تصادفی، دانشمندان یک محاسبه دو مرحله ای را پیشنهاد می کنند:

قدم اول: یک تقریب پایه با برد ماتریس A محاسبه می‌کنیم. می‌خواهیم یک ماتریس Q متعامد و نرمال با l ستون تعریف کنیم به طوری که عملA را به تصویر می‌کشد. در در واقع $QQ^*A \simeq A$ و Q را به گونه‌ای تعریف می‌کنیم که نامساوی زیر به ازای ϵ مشخص برقرار باشد.

$$||I - QQ^*A|| \leq \epsilon$$

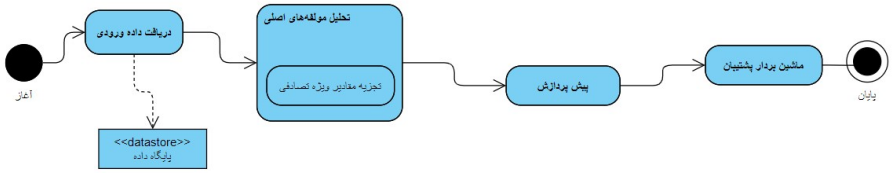
الگوریتم ۲.۲. الگوریتم قدم اول

- تعریف یک ابرپارامتر l برای نمونه‌برداری بیش از اندازه
- ساخت یک ماتریس Ω گاوسی تصادفی با ابعاد $n \times l$
- ایجاد ماتریس $Y = A\Omega$ با ابعاد $m \times l$
- ساخت یک ماتریس متعامد نرمال Q برای مثال با استفاده از روش فاکتورگیری QR ، $Y = QR$

قدم دوم: در این مرحله قصد داریم تجزیه مقادیر منفرد تصادفی ماتریس Q که از ماتریس A ابعاد کوچک‌تری دارد را محاسبه می‌کنیم. سپس به وسیله تجزیه مقادیر منفرد ماتریس Q به تجزیه مقادیر منفرد ماتریس A را به دست می‌آوریم.

الگوریتم ۳.۲. الگوریتم قدم دوم

- تعریف ماتریس B به صورت $B = Q^*A$
- محاسبه تجزیه مقادیر منفرد ماتریس $B = S\Sigma V^T$
- از آن‌جایی که A≈QQ^*A پس $A = QB = QS\Sigma V^T$
- تعریف U به طوری که: $U = QS$
- در نهایت U, Σ, V با تجزیه مقادیر منفرد ماتریس A .



شکل ۱: نمودار الگوریتم

۴.۲ ماشین بردار پشتیبان

ماشین بردار پشتیبان یکی از شناخته شده ترین الگوریتم های یادگیری ماشین در حوزه یادگیری نظارت شده و یادگیری بدون نظارت است. [۲]. این الگوریتم، رویکردی افتراقی دارد. مدل ماشین بردار پشتیبان که بر پایه ی تئوری یادگیری آماری و قضیه VC بنا شده است یکی از قدرتمند ترین مدل‌های یادگیری ماشین است.در این مدل فرض می شود هر نمونه داده ورودی به صورت $x_i \in \mathbb{R}^d$ و خروجی مطلوب آن برای مسائل رگرسیون $y \in \mathbb{R}$ و برای طبقه‌بندی $y \in \{-1, +1\}$ است. در مدل دسته بند، الگوریتم تلاش می کند معادله بهترین ابرصفحه جدا کننده را به گونه ای به دست آورد که جداکننده حاصل از داده های دو دسته بیشترین حاشیه را داشته باشد.. [۲] در ادامه ماشین بردار پشتیبان، با استفاده از ترفند هسته این ابرصفحه را به یک رویه تبدیل کرده و عملیات پیش بینی را انجام می دهد. به این منظور دسته بند به صورت:

$$y(x) = \text{sign}\left[w^T\varphi(x) + b\right]$$

در نظر گرفته شده و جهت به دست آوردن ابرصفحه بهینه، مسئله ی بهینه سازی محدب زیر تشکیل می‌شود:

$$\begin{aligned} \min w, b, \xi \quad & \frac{1}{2}w^T w + C\sum_{k=1}^{n}\xi_k \\ \text{s.t.} \quad & y_k\left[w^T\varphi(x_k) + b\right] \geq 1 - \xi_k, \qquad k = 1, \ldots, n. \\ & \xi_k \geq 0, \qquad k = 1, \ldots, n. \end{aligned}$$

فرم دوگان این مسئله بهینه سازی نیز به کمک ضرایب لاگرانژ محاسبه می‌شود که در آن Ω همان ماتریس هسته می‌باشد.

$$\begin{aligned} \max_{\alpha} \quad & \frac{1}{2}\alpha^T\Omega\alpha - 1^T\alpha \\ \text{s.t.} \quad & 0 \leq \alpha \leq C \\ & y^T\alpha = 0 \end{aligned}$$

$$\Omega_{i,j} = y_i y_j \varphi(x_i)^T\varphi(x_j), \quad i, j = 1, 2, \ldots, n.$$

در این صورت دسته بند مدل در فرم دوگان به صورت زیر خواهد بود:

$$y(x) = \text{sign}\left[\sum_{k=1}^{n}\alpha_k y_k \varphi(x_k)^T\varphi(x) + b\right]$$

۳ کاربرد تجزیه مقادیر منفرد تصادفی در الگوریتم تحلیل مولفه‌های اصلی

در این پژوهش برای محاسبه تجزیه مقادیر منفرد در الگوریتم تحلیل مولفه‌های اصلی از تجزیه مقادیر منفرد تصادفی استفاده می‌کنیم.در این بخش به معرفی این مدل و نتایج حاصل از به کارگیری این مدل بر روی مجموعه داده‌های انتخابی می‌پردازیم.

مجموعه داده	تعداد داده‌های آموزشی	تعداد داده‌های آزمایشی	تعداد ویژگی
Scania Trucks	۶۰۰۰۰	۱۶۰۰۰	۱۷۱
Fashion-MNIST	۶۰۰۰۰	۱۰۰۰۰	۷۸۴
MNIST	۶۰۰۰۰	۱۰۰۰۰	۷۸۴

جدول ۱: جدول داده‌ها

۱.۳ مدل پیشنهادی

حال می‌خواهیم به بررسی مدل به کار برده شده در این پژوهش بپردازیم. نمودار فعالیت الگوریتم این مدل در نمودار الگوریتم آمده است. ابتدا مجموعه داده ورودی از پایگاه داده خوانده می‌شود. سپس در مرحله بعدی الگوریتم تحلیل مولفه‌های اصلی که پیش از این بررسی نمودیم بر روی پایگاه داده اعمال می‌شود. الگوریتم تحلیل مولفه‌های اصلی که در این مدل به کار می‌بریم پایه متعامد را به کمک تجزیه مقادیر تصادفی محاسبه می‌نماید. پس از محاسبه فضای داده جدید توسط تحلیل مولفه‌های اصلی پیش پردازش‌های مورد نیاز برای عملکرد هرچه بهتر الگوریتم بر روی مجموعه داده فعلی اعمال می‌شود. حال پس از پیش‌پردازش الگوریتم SVM بر روی مجموعه داده اعمال می‌کنیم.

۲.۳ نتایج

به منظور سنجش عملکرد صحیح الگوریتم پیشنهادی در این تحقیق از سه مجموعه داده معتبر در سایت Kaggle استفاده نمودیم. ابتدا به شرح مختصر این مجموعه داده‌ها می پردازیم.

مجموعه داده Scania Trucks : این مجموعه داده شامل داده‌های جمع آوری شده مربوط به سیستم فشار هوای کامیون‌های اسکانیا می‌باشد. این سیستم در خودرو هوای مورد نیاز برای عملکردهای مختلف مانند ترمزگیری و تعویض دنده و غیره را تولید می‌کند. کلاس مثبت با ۱۰۰۰ رکورد شامل کامیون‌هایی با خرابی اجزا برای یک جزء خاص از سیستم فشار هوا است. کلاس منفی با ۵۹۰۰۰ رکورد شامل کامیون‌هایی با خرابی قطعات غیر مرتبط با سیستم فشار هوا است. هر رکورد نیز شامل ۱۷۱ ویژگی می‌باشد.

مجموعه داده Fashion-MNIST : این مجموعه داده عکس‌های مقاله زولاندو شامل ۶۰۰۰۰ داده آموزشی و ۱۰۰۰۰ داده آزمایشی می‌باشد. هر رکورد یک عکس با ابعاد ۲۸ × ۲۸ پیکسل است که هر کدام متعلق به یکی از ۱۰ کلاس برچسب‌ها هستند. این برچسب‌ها نام نوعی از لباس مانند : تیشرت، پلیور و ... هستند.

مجموعه داده MNIST : این مجموعه داده شامل داده‌های MNIST می‌باشد. شامل ۶۰۰۰۰ داده آموزشی و ۱۰۰۰۰ داده آزمایشی می‌باشند. هر رکورد تصویری ۲۸ × ۲۸ پیکسل می‌باشد که متعلق به کلاس یکی از اعداد ۰ الی ۹ می‌باشد.

نتایج حاصل از عملکرد مدل در سه حالت بدون استفاده از تحلیل مولفه‌های اصلی، با استفاده از تحلیل مولفه‌های اصلی بدون به کارگیری تجزیه مقادیر منفرد و با استفاده از تحلیل مولفه‌های اصلی با به کارگیری تجزیه مقادیر منفرد تصادفی را در جدول نتایج مشاهده می‌فرمایید. .همان‌طور که مشاهده می‌کنید در هر سه مجموعه داده بعد از استفاده از PCA با کاهش زیاد زمان رو به رو هستیم. هم‌چنین در زمان استفاده از تجزیه مقادیر منفرد تصادفی در PCA برای داده‌های با ابعاد بزرگتر این زمان باز هم کاهش می‌یابد و تغییر ناچیزی در دقت اتفاق می‌افتد که قابل چشم‌پوشی است.

۴ نتیجه‌گیری

هنگامی که تعداد ویژگی‌ها در برخی از مسائل بسیار بالا بوده دچار مشکلات زمان، حافظه و پدیده‌ای مانند مشقت چندبعدی یا نفرین ابعاد می‌شویم. پدیده مشقت چندبعدی نیز به همراه خود مشکلاتی همچون کاهش سرعت و دقت مدل‌های یادگیری ماشین را خواهد داشت. در این تحقیق تلاش نمودیم با استفاده از روش تجزیه مقادیر منفرد تصادفی در الگوریتم تحلیل مولفه‌های اصلی کاهش ابعاد انجام دهیم. نتایج بدست آمده حاکی از کارایی بالای این روش در سرعت و زمان با حفظ دقت می‌باشد..

مجموعه داده	بدون اعمال PCA		PCA بدون اعمال SVD تصادفی		PCA با اعمال SVD تصادفی	
	زمان	دقت	زمان	دقت	زمان	دقت
Trucks Scania	۵۴.۳۳	۹۸.۸.۹۱۸	۰.۹۲	۹۸.۷۷۵۰	۰.۹۹	۹۸.۸۶۸۸
MNIST	۶۷۹.۵۹	۹۱.۱۹۰۰	۷.۰۱	۹۱.۳۲۰۰	۱.۷۴	۹۰.۶۲۰۰
MNIST Fashion	۷۳۸.۴۹	۸۳.۷۴۰۰	۷.۹۶	۸۴.۹۲۰۰	۱.۷۲	۸۳.۳۹۰۰

جدول ۲: جدول نتایج

مراجع

[۱] م. حجاریان. ،نخستین درس در جبر خطی عددی، انتشارات شهید بهشتی، تهران، ۱۳۹۷.

[2] Cortes, Corinna, and Vladimir Vapnik. ”Support-vector networks.” Machine learning 20.3 (1995): 273-297.

[3] N. Halko, P. G. Martinsson, J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. SIAM Review.2011.

[4] S Wold, K Esbensen, P Geladi. Principal component analysis. Chemometrics and Intelligent Laboratory Systems.1987 August.