دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش ارائه

درس داده‌کاوی

نگارش

امیرحسین پاشایی‌هیر، حیدر فهمی، مهدی قیاسی

استاد راهنما

دکتر مریم امیر مزلقانی

زمستان ۱۴۰۰


بسم الله الرحمن الرحيم

سپاس‌گزاری

وظیفه خود می‌دانیم که مراتب امتنان خود را نسبت به استادمان در درس داده‌کاوی، دکتر مریم امیر مزلقانی که ما را یاری نموده‌اند، اعلام بداریم.

امیرحسین پاشایی، امیرحیدر فهمی، مهدی قیاسی
زمستان ۱۴۰۰

چکیده

در این گزارش به بررسی روشی به نام تی اس ان ای می‌پردازیم که هدف آن مصور سازی داده‌هایی با ابعاد بالا می‌باشد که این هدف را با نگاشت کردن هر داده به یک نقطه در فضای دو یا سه بعدی انجام می‌دهد. این روش یک نسخه از روش تعبیه همسایه تصادفی (یا اس ان ای) می‌باشد که بهینه کردن آن بسیار راحت‌تر بوده و با کم کردن گرایش نقاط متراکم به مرکز، خروجی به مراتب بهتری تحویل می‌دهد. تی اس ان ای از روش‌های موجود، عملکرد بهتری در ارائه انواع ساختارهای داده کلی در قالب یک نگاشت، ارائه می‌دهد. این عملکرد بخصوص برای داده‌هایی که ابعاد بالایی دارند اما در اصل به روی چند منحی با ابعادی پایین قرار داده شده‌اند بسیار مهم می‌شود. برای مصور سازی ساختار داده‌های با ابعاد بالا، بررسی می‌کنیم که تی اس ان ای چگونه با استفاده از قدم زدن اتفاقی به روی گراف همسایگی می‌تواند ساختار داده‌ها را خلاصه کند و آن‌ها را مصور سازی کند، ما همچنین به مقایسه‌ی عملکرد روش تی اس ان ای نسبت به تعداد زیادی از روش‌ها مانند نگاشت سامون، ایزومپ و تعبیه خطی محلی می‌پردازیم و در نتیجه می‌بینم که تی اس ان ای، مصور سازی به مراتب بهتری نسبت به بقیه روش‌ها بر روی تمام مجموعه داده‌ها ارائه داده است.

واژه‌های کلیدی:

مصورسازی، کاهش بعد، الگوریتم‌های تعبیه سازی، مقیاس چندبعدی

فهرست مطالب

عنوان صفحه

۱ مقدمه ۱

۲ تعبیه همسایه تصادفی (اس ان ای) ۴

۳ روش تی اس ان ای ۸
۳-۱ اس ان ای متقارن ۹
۳-۲ مشکل شلوغی ۱۰
۳-۳ دم نامناسب می‌تواند ابعاد ازبین رفته را جبران کند ۱۱
۳-۴ روش‌های بهینه‌سازی برای تی-اس‌ان‌ای ۱۲

۴ آزمایش‌ها ۱۴
۴-۱ دیتاست‌ها ۱۵
۴-۲ نحوه‌آزمایش ۱۶

۵ به‌کارگیری تی‌اس‌ان‌ای بر روی مجموعه داده‌های بزرگ ۲۱

۶ بررسی نهایی ۲۵
۶-۱ مقایسه با دیگر تکنینک‌ها ۲۶
۶-۲ نقاط ضعف ۲۷
۶-۳ نتیجه گیری ۲۸

منابع و مراجع ۲۹

فهرست اشکال

شکل صفحه

۲-۱ فرمول تابع احتمال شرطی ۵
۲-۲ فرمول تابع احتمال شرطی برای داده‌های نگاشت داده شده ۵
۲-۳ فرمول تابع C ۵
۲-۴ فرمول سرگشتگی ۶
۲-۵ فرمول آنتروپی شنون ۶
۲-۶ گرادیان تابع C ۶
۲-۷ فرمول گرادیان به همراه تکانه ۷

۳-۱ تابع هزینه متقاران ۹
۳-۲ تابع شباهت در ابعاد پایین ۹
۳-۳ تابع شباهت در ابعاد بالا ۱۰
۳-۴ تابع گرادیان برای اس ان ای متقارن ۱۰
۳-۵ شبه کد بهینه سازی ۱۲

۴-۱ نمایش ۶۰۰۰ هزار تصویر اعداد ۱ تا ۹ به صورت دست‌نویس از دیتاست اول ۱۷
۴-۲ نمایش ۶۰۰۰ هزار تصویر اعداد ۱ تا ۹ به صورت دست‌نویس از دیتاست اول ۱۸
۴-۳ مقایسه الگوریتم‌ها در دیتاست دوم ۱۹
۴-۴ مقایسه الگوریتم‌ها در دیتاست سوم ۲۰

۵-۱ . . . . . ۲۲
۵-۲ . . . . . ۲۳

۶-۱ . . . . . ۲۶
۶-۲ . . . . . ۲۷

فصل اول

مقدمه

مصور سازی[1] داده‌های با ابعاد بالا مسئله‌ی مهمی در بسیاری از دامنه‌ها می‌باشد. این مسئله با بازه‌ی بزرگی از ابعاد (از سی بعد برای داده‌های پزشکی تا ده‌ها هزار بعد برای داده‌های سندها) درگیر است.
در دهه‌ی اخیر راه‌حل‌های متنوعی برای این مسئله ارائه شده است [۲] که می‌توان به روش‌های شمایل نگاری، روش‌های براساس پیکسل[2] و روش‌هایی که ابعاد را به صورت رئوس گراف نشان می‌دهند اشاره کرد.
اکثر این روش‌ها ابزاری ارائه می‌دهند که داده‌ها را بتوان به صورت دو بعدی نمایش داد و تحلیل را به مشاهده انسانی واگذار می‌کنند. این موضوع باعث می‌شود که کاربرد این روش‌ها در دنیای واقعی برای داده‌هایی که ده‌ها هزار بعد دارند کم شود.
برخلاف روش‌های بالا، روش‌های کاهش بعد سعی می‌کنند داده‌های با ابعاد بالا را به داده‌های دو یا سه بعدی خلاصه کنند.
روش‌های خطی و سنتی‌ای مانند تجزیه و تحلیل اجزای اصلی یا پی سی ای[3] یا مقایسه چند بعدی کلاسیک یا ام دی اس[4] روش‌هایی هستند که سعی دارند داده‌های غیرمشابه را پس از نگاشت تا جای ممکن از هم دور نگه دارند. در روش‌های غیرخطی سعی می‌شود که داده‌هایی که در توصیف اصلی خود به همدیگر نزدیک هستند، پس از نگاشت و خلاصه شدن نیز باز هم به همدیگر نزدیک بمانند، که پیاده‌سازی این مهم توسط روش‌های خطی امکان پذیر نمی‌باشد.
روش‌های غیرخطی زیادی برای کاهش بعد ارائه شده است که سعی دارند ساختار محلی داده را نگه دارند. از این روش‌ها می‌توان به نگاشت سامون، تحلیل و تجزیه اجزای منحنی یا سی سی ای[5] ، تعبیه همسایه تصادفی یا اس ان ای[6] [۱] ، حداکثر واریانس آشکار یا ام وی یو[7] ، تعبیه خطی محلی یا ال ال ای[8] و نگاشت ویژه لاپلاس[9] اشاره کرد. با اینکه روش‌های بر داده‌های مصنوعی با ابعاد بالا بسیار خوب عمل می‌کنند اما در مصور سازی داده‌های با ابعاد بالا، نمی‌توانند عملکرد مناسبی داشته باشند. اکثر این روش‌ها نمی‌توانند ساختار عمومی و محلی داده‌ها را در یک نگاشت به خوبی مصور کنند. برای مثال، یک مدل ام وی یو با نظارت متوسط نمی‌تواند رقم‌های با دست نوشته را به خوشه‌های طبیعی هر رقم نگاشت کند و رقم‌ها را از هم جدا کند.
در این گزارش ما به بررسی روشی می‌پردازیم که داده‌های با ابعاد بالا را به ماتریس شباهت داده‌ها تبدیل می‌کند و سپس با روشی به نام تی اس ان ای[10] [۴] [۳] این ماتریس به دست آمده را مصور سازی می‌کند.

[1] Visualization
[2] Pixel
[3] PCA
[4] MDS
[5] CCA
[6] SNE
[7] MVU
[8] LLE
[9] Laplacian Eigenmaps
[10] t-SNE

روش تی اس ان ای توانایی نمایش دادن اکثر ساختار محلی داده‌ها را داشته و در کنار آن ساختار عمومی داده‌ها مانند خوشه‌های موجود در داده‌ها را نیز نمایش می‌دهد.
در این گزارش ما عملکر تی اس ان ای را با مقایسه کردن آن با هفت روش بیان شده بر روی پنج مجموعه داده‌ی بدست آمده از دامنه‌های مختلف بررسی خواهیم کرد، نتایج نشان می‌دهد که این روش در اکثر دامنه‌ها نسبت به بقیه روش‌ها دارای برتری می‌باشد.
در فصل بعدی اس ان ای را معرفی می‌کنیم که مفاهیم آن پایه تی اس ان ای می‌باشد، سپس در فصل بعدی آن تی اس ان ای را معرفی می‌کنیم که دو تفاوت اساسی با اس ان ای دارد، در فصل بعد آن شرایط آزمایش و نتایج آزمایش‌ها را بیان کرده و در فصل بعد آن نشان می‌دهیم که تی اس ان ای چطور می‌تواند تغییر پیدا کند به شکلی که داده‌هایی که ابعاد آن‌ها بسیاری بیشتر از ده هزار بعد است را نمایش دهد و در فصل بعد آن نتایج آزمایش‌ها با دقت بیشتری بررسی می‌شود، در فصل آخر نیز جمع بندی انجام می‌شود.

فصل دوم

تعبیه همسایه تصادفی (اس ان ای)

اس ان ای با تبدیل فاصله‌ی اقلیدسی برای ابعاد بالا به یک تابع احتمال شرطی که نشان دهنده‌ی شباهت بین دو نقطه است کار خود را آغاز می‌کند. شباهت داده‌ی x_i با داده‌ی x_j به صورت تابع احتمال شرطی $p_{j|i}$ نشان داده می‌شود. معنای این تابع احتمال این است که داده‌ی x_i داده‌ی x_j را به عنوان همسایه خود انتخاب کند که احتمال انتخاب داده‌ها به عنوان همسایه، یک توزیع گواسی به مرکزیت x_i می‌باشد. برای داده‌های نزدیک به x_i این احتمال نزدیک بالا و برای داده‌های دور از x_i این احتمال تقریبا بسیار کوچک است (اگر مقدار واریانس به خوبی انتخاب شود). فرمول این تابع به شکل زیر است:

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

شکل ۲-۱: فرمول تابع احتمال شرطی

به دلیل اینکه ما به دنبال شباهت به صورت دوتایی داده‌ها هستیم مقدار $p_{i|i}$ برابر با صفر در نظر گرفته می‌شود، همچنین نحوه به دست آوردن واریانس (σ_i) بعدا در این بخش توضیح داده می‌شود.
اگر فرض کنیم که داده‌ی x_i به داده‌ی y_i و داده y_i به داده‌ی y_j نگاشت شده باشد، تابع احتمال شرطی دیگری مانند p برای شباهت بین دو داده‌ی نگاشت داده شده به نام q تعریف می‌کنیم که فرمول آن به شکل زیر است:

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}.$$

شکل ۲-۲: فرمول تابع احتمال شرطی برای داده‌های نگاشت داده شده

در ایت تابع نیز باز مقدار $q_{i|i}$ برابر با صفر در نظر گرفته می‌شود و همچنین مقدار واریانس برابر با $\frac{1}{\sqrt{2}}$ در نظر گرفته شده است.
در اس ان ای سعی می‌شود که تفاوت مقدار $p_{j|i}$ و مقدار $q_{j|i}$ تا حد ممکن کمینه شود، این موضوع باعث می‌شود که شباهت بین داده‌ها حفظ شود. از تابع واگرایی کولبک-لایبلر استفاده می‌کنیم. (که در این حالت همان کراس آنتروپلی می‌باشد)
اس ان ای سعی می‌کند که با روش شیب گرادیان تابع C را مینیمم کند. تعریف تابع C به صورت زیر است:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}},$$

شکل ۲-۳: فرمول تابع C

در این فرمول P_i به معنای توزیع شرطی احتمالی به روی تمام نقاط به غیر از x_i می‌باشد و Q_i نیز به معنای توزیع شرطی احتمالی به روی تمام نقاط نگاشت شده به غیر y_i می‌باشد. به دلیل عدم تقارن تابع

واگرایی کولبک-لایبلر، انواع مختلف خطا در فواصل دوتایی در داده‌های نگاشت شده‌ی کم بعد دارای وزن یکسانی نمی‌باشند. برای مثال، تابع هزینه بسیار افزایش پیدا می‌کند اگر دو داده‌ی دور از هم به دو داده‌ی نزدیک به هم نگاشت پیدا کنند اما تابع هزینه تغییر آنچنانی نمی‌کند اگر دو داده‌ی نزدیک به دو داده‌ی دور از هم نگاشت پیدا کنند.
دلیل این اتفاق را می‌توان به این صورت بیان کرد که تابع هزینه اس ان ای تمرکز خود را بر روی حفظ ساختار محلی و انتقال آن به داده‌های نگاشت داده شده، قرار داده است.
پارامتر واریانس (σ_i) نمی‌تواند برای تمام نقاط برابر باشد، برای مثال در مکان‌هایی با چگالی بالا مقدار واریانس کوچک بهتر عمل می‌کند. با افزایش پارامتر واریانس، آنتروپلی توزیع احتمال آن نقطه نیز افزایش پیدا می‌کند.
روش اس ان ای برای پیدا کردن این مقدار، از سرچ دودویی استفاده می‌کند و به دنبال مقداری از واریانس میگردد که توزیع احتمالی درست کند که سرگشتگی آن برابر با مقدار مشخص شده توسط کاربر باشد، سرگشتی طبق فرمول زیر تعریف می‌شود:

$$Perp(P_i) = 2^{H(P_i)},$$

شکل ۲-۴: فرمول سرگشتگی

که در آن $H(P_i)$ شانون آنتروپلی می‌باشد که فرمول آن به شکل زیر می‌باشد:

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}.$$

شکل ۲-۵: فرمول آنتروپی شنون

سرگشتگی می‌تواند به صورت اندازه‌گیری تعداد همسایه‌های موثر تفسیر شود. عملکر اس ان ای نسبت به تغییرات اندازه‌ی سرگشتی بسیار مقاوم است و مقادیر معمول این پارامتر، عددی بین ۵ تا ۵۰ می‌باشد.
برای استفاده از روش شیب گرادیان برای کوچک کردن مقدار تابع C از گرادیان این تابع استفاده می‌کنیم که فرمول آن به شکل زیر است:

$$\frac{\delta C}{\delta y_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j).$$

شکل ۲-۶: گرادیان تابع C

به صورت فیزیکی، گرادیان می‌تواند به صورت برآیند نیروهای فنرهایی بین نقطه‌ی نگاشت شده‌ی y_i و بقیه نقاط تفسیر شود، نیروی فنر بین دو داده‌ی y_i و داده‌ی y_j در راستای $y_i - y_j$ واقع شده است که این نیرو بر اساس میزان نزدیکی این دو داده در فضای اصلی با ابعاد بالا، می‌تواند به صورت جذبی یا دفعی باشد.

نیروی هر فنر وابسته به طول آن (فاصله‌ی دو نقطه‌ی نگاشت شده) و سختی آن است که این سختی بر اساس فرمول $(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$ محاسبه می‌شود.
روش شیب گرادیان کار خود را با سمپل کردن نقاط نگاشت پیدا شده به صورت اتفاقی از یک توزیع گوسی که حول مرکز واقع شده است شروع می‌کند. برای افزایش سرعت و همچنین پرهیز کردن از مینیمم‌های محلی، یک مقدار نسبتا بزرگ تکانه به گرادیان اضافه می‌شود، به عبارتی دیگر مقدار کنونی گرادین با مقادیر گرادیان‌های قبلی که به صورت تصاعدی و نزولی ضریب داده شده‌اند، جمع بسته می‌شود. به صورت کلی، فرمول بروز رسانی مقدار گرادیان به همراه تکانه به صورت زیر می‌باشد:

$$\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t) \left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)} \right),$$

شکل ۲-۷: فرمول گرادیان به همراه تکانه

که در آن $\gamma^{(t)}$ نماد نگاشت در مرحله‌ی tام، η نماد ضریب یادگیری و $\aleph(t)$ نماد تکانه در مرحله‌ی tام می‌باشد.
همچنین در مراحل اولیه الگوریتم، یک نویز گوسی به هر نقطه‌ی نگاشت شده، پس از هر مرحله اضافه می‌شود. کاهش به تدریج مقدار واریانس این نویز، باعث می‌شود که روش بتواند از مینیمم‌های محلی ضعیف، فرار کند. در اس ان ای متاسفانه این روند نزول واریانس نویز و انتخاب مقدار اولیه آن، نیاز به تنظیم دستی توسط کاربر دارد و همچنین این مقادیر تاثیر زیادی بر روی خروجی خواهند داشت.
همچنین پارمترهایی مانند تکانه و اندازه‌ی هر مرحله نیز نیاز به تنظیم دستی دارند. به همین دلیل به صورت معمول، روش اس ان ای چندین بار بر روی داده با پارمترهای مختلف اجرا می‌شود و از بین آن‌ها خروجی بهتر انتخاب می‌شود.
به طول کلی روش اس ان ای، نسبت به روش‌های دیگری که خروجی مناسبی را بدون نیاز به تنظیم دستی و زمان زیاد اجرای چند باره‌ی الگوریتم ارائه می‌دهند در سطح پایین‌تری قرار دارد.

فصل سوم

روش تی اس ان ای

در فصل قبلی به توضیح روش اس ان ای پرداختیم، با اینکه روش اس ان ای مصور سازی خوبی ارائه می‌دهد اما عملکرد این روش توسط تابع هزینه‌ای که به سختی بهینه می‌شود و مشکلی به نام مشکل شلوغی، کاهش یافته است. در این بخش ما به بررسی روش تی اس ان ای می‌پردازیم که سعی دارد این مشکلات را حل کند.
تابع هزینه این روش دو تفاوت اساسی با تابع هزینه‌ی روش اس ان ای دارد، اولین تفاوت این است که در این روش از نسخه‌ی متقارن تابع اس ان ای برا گرادیانی ساده‌تر استفاده می‌کنیم و دومین تفاوت این است که برای محاسبه‌ی شباهت بین داده‌های کم بعد، به جای توزیع گوسی از توزیع استیودنت[1] استفاده می‌شود
روش تی اس ان ای از توزیع دم سنگین[2] برای ابعاد کوچک استفاده می‌کند که دو مشکل بهینگی و مشکل شلوغی در اس ان ای را حل نماید.
در این فصل ابتدا به نسخه‌ی متقارن اس ان ای می‌پردازیم، سپس مشکل شلوغی را مطرح کرده و پس از آن نحوه استفاده از توزیع‌های دم سنگین برای حل این مشکل را بیان می‌کنیم، در نهایت روش‌های مطرح شده برای حل مشکل بهینگی اس ان ای بیان می‌شود.

۳-۱ اس ان ای متقارن

برای یک جایگزین برای مینیمم کردن تابع واگرایی کولبک-لیبلر بین توابع احتمالی شرطی $pj|i$ و $q_{j|i}$ می‌توان سعی کرد که یک تابع واگرایی کولبک-لیبلر بین توزیع P در ابعاد بالا و توزیع Q در ابعاد پایین استفاده نمود. فرمول تابع هزینه جدید به این شکل می‌شود:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

شکل ۳-۱: تابع هزینه متقارن

که باز در این تابع نیز مقادیر p_{ii} و q_{ii} برابر با صفر هستند. این نوع اس ان ای، اس ان ای متقارن شمرده می‌شود زیرا برای تمام i و j ها داریم که $p_{ij} = p_{ji}$ و $q_{ij} = q_{ji}$.
در اس ان ای متقارن، شباهت بین دو نقطه نگاشت شده در ابعاد پایین با فرمول زیر محاسبه می‌شود:

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq l} \exp\left(-\|y_k - y_l\|^2\right)},$$

شکل ۳-۲: تابع شباهت در ابعاد پایین

و همچنین تابع شباهت بین دو نقطه در ابعاد بالا نیز به فرمول زیر است:

[1] Student-t distribution
[2] heavy-tailed distribution

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)}{\sum_{k\neq l}\exp\left(-\|x_k - x_l\|^2/2\sigma^2\right)},$$

شکل ۳-۳: تابع شباهت در ابعاد بالا

اما این فرمول در زمانی که یک داده در ابعاد بالا، داده‌ی پرت[۳] باشد مشکل زا می‌شود زیرا مقدار p_{ij} برای تمام j های به غیر از این داده بسیار کم می‌شود و نگاشت آن به هر جایی از صفحه آنچنان تاثیری در تابع هزینه نخواهد داشت و در نتیجه مکان این نقطه نگاشت شده، به خوبی توسط بقیه نقاط مشخص نمی‌شود.
برای دور زدن این مشکل، تابع شباهت در ابعاد بالا را به صورت میانگین دو تابع احتمال شرطی تعریف می‌کنیم، یعنی مقدار p_{ij} را به صورت $\frac{p_{j|i}+p_{i|j}}{2n}$ تعریف می‌کنیم. این تعریف تضمین می‌دهد که $\sum_j p_{ij} > \frac{1}{2n}$ برای تمام x_i ها و در نتیجه هر نقطه، تاثیر بالایی در تابع هزینه C خواهد داشت.
تابع شباهت برای ابعاد پایین را تغییر نمی‌دهیم زیرا مشکل قبلی پیش نخواهد آمد.
برتری اصلی اس ان ای متقارن نسبت به اس ان ای معمولی این است که تابع گرادیان آن بسیار ساده‌تر است که این موضوع باعث می‌شود که محاسبه‌ی آن بسیار سریع‌تر باشد. تابع گرادیان برای اس ان ای متقارن به شکل زیر است:

$$\frac{\delta C}{\delta y_i} = 4\sum_j (p_{ij} - q_{ij})(y_i - y_j).$$

شکل ۳-۴: تابع گرادیان برای اس ان ای متقارن

نتایج آزمایش‌ها نشان می‌دهد که اس ان ای متقارن قالبا همان نگاشت اس ان ای را خروجی می‌دهد و گاها عملکرد بهتری نیز ارائه می‌دهد.

۳-۲ مشکل شلوغی

مجموعه‌ای از نقاط دو بعدی را در نظر بگیرید، که بر روی یک منحی دو بعدی نامتقارن و دارای پستی و بلندی قرار دارند که در به صورت تقریبی می‌توان با یک خط در اندازه‌های کوچک تقریب زده شوند، و تمام این نقاط در فضایی با ابعاد بالا تعبیه شده‌اند. این امکان وجود دارد که فاصله‌ی دو به دوی نقاط را در یک نگاشت دو بعدی مدل کرد.
حال فرض کنید که پستی بلندی‌های واقعی منحنی که در نگاه جزئی دیده نمی‌شدند دارای ده بعد دیگر جز این دو بعد باشند که دیگر نمی‌توان آن‌ها را به خوبی در دو بعد مدل کرد. دلایل متنوعی برای این موضوع وجود دارد که برای مثال می‌توان گفت که در ۱۰ بعد، می‌توان ۱۱ نقطه را جوری در صفحه چید

[3] outlier

که فاصله‌ی بین آن‌ها را به هیچ صورتی نمی‌توان به خوبی در دو بعد نشان داد.
یک مسئله مربوط به این موضوع، مسئله‌ی نگاشت یک ابر کره‌ی m بعدی به دو بعد است که توزیع نقاط در آن به این صورت است که یک مرکز در نظر گرفته شده و احتمال وجود یه نقطه در مکانی که فاصله‌ی r از مرکز دارد متناسب با r^m می‌باشد، که برای حل این مسئله نیز به مشکل شلوغی می‌خوریم.
مسئله شلوغی رو می‌توانیم به این صورت تعریف کنیم: فضای دو بعدی در دسترس برای نمایش فاصله‌ی نقاط نسبتا دور از هم به اندازه‌ی کافی بزرگتر از فضای دو بعدی در دسترس برای نمایش فاصله‌ی نقاط نزدیک به هم نیست، به همین دلیل ما اگر بخواهیم فاصله‌ی نقاط نزدیک به هم را به خوبی نگاشت دهیم، نقاط نسبتا دور از هم در نگاشت، بسیار دور از هم نگاشت داده می‌شوند.
در اس ان ای، فنری که بین این داده‌ی i و هر یک از این داده‌های بسیار دور نگاشت داده شده، قرار داده شده است، بسیار سختی پایینی دارد و به همین دلیل نیروی جذبی ضعیفی از خود نشان می‌دهد.
اگرچه این فنرها نیروهای بسیار ضعیفی از خود نشان می‌دهند اما تعداد زیادی از این نیروها داده‌ها را به سمت مرکز حرکت می‌دهد که باعث می‌شود فاصله‌ی طبیعی بین خوشه‌ها از بین برود.
دقت کنید که این مشکل فقط مربوط به اس ان ای نمی‌باشد و در روش‌های دیگر نیز پیش می‌آید.
یک تلاش برای حل مشکل شلوغی، اضافه کردن یک دافعه خفیف به هر فنر می‌باشد.
این دافعه خفیف به این شکل ساخته می‌شود که یک مدل پس‌زمینه یکنواخت با نسبت اختلاط ρ معرفی می‌شود. که در آن نقاط نگاشت داده شده‌ی بسیار دور از هم مقدار q_{ij} کمتر از $\frac{2p}{n(n-1)}$ نمی‌توانند داشته باشند که در نتیجه‌ی آن نقاطی که در ابعاد بالا بسیار دور هستند مقدار q_{ij} آن‌ها همواره بیشتر از مقدار p_{ij} آن‌ها می‌شود که این موضوع باعث یک دافعه خفیف می‌شود.
به این روش یونی اس ان ای[۴] می‌گویند که گرچه بسیار بهتر از روش اس ان ای عمل می‌کند اما تابع هزینه‌ی آن بسیار پیچیده است.

۳-۳ دم نامناسب می‌تواند ابعاد ازبین رفته را جبران کند

از آنجایی که روش اس‌ان‌ای متقارن تلاش می‌کند تا توزیع توأم در ابعاد بالا و پایین را برابر یک‌دیگر قرار دهد، راه حل مناسبی برای حل مشکل گزارش شده در زیر بخش قبل وجود دارد که به شرح زیر است. در ابعاد بالا برای تبدیل فاصله به احتمال از توزیع گوسی استفاده‌شده‌است اما در ابعادپایین می‌توان از توزیع احتمالی استفاده کرد که فاصله خطی بیش‌تری را نسبت به توزیع گوسی ایجاد کند. با این‌روش فاصه‌های که متوسط هستند نیز در ابعاد کوچک به فاصله‌های بزرگ‌تر مپ می‌شوند و به طبع باعث می‌شود که آن‌های که در یک خوشه نیستند شباهت احتمالی کمتری داشته‌باشند و نمایش بهتری داشته باشیم.

[4] UNI-SNE

Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.

Data: data set $X = \{x_1, x_2, ..., x_n\}$,
cost function parameters: perplexity $Perp$,
optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$.
Result: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, ..., y_n\}$.
begin
  compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ (using Equation 1)
  set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
  sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, ..., y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$
  for t=1 **to** T **do**
    compute low-dimensional affinities q_{ij} (using Equation 4)
    compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
    set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$
  end
end

شکل ۳-۵: شبه کد بهینه سازی

۳-۴ روش‌های بهینه‌سازی برای تی-اس‌ان‌ای

در ابتدا روش تی-اس‌ان‌ای با استفاده از الگوریتم گرادیان‌دسنت روی تابع‌هزینه بهینه‌سازی شد. شبه کد نحوه‌ی بهینه سازی آن درشکل ۳-۵ آورده شده است.

این الگوریتم ساده این قابلیت‌را دارد تا با استفاده از نرخ‌یادگیری‌انطباقی که در سال ۱۹۸۸ توسط جکابس [5]معرفی شد، که در آن نرخ یادگیری را درجهتی که گرادیان ثابت بماند، نرخ یادگیری را زیادمیکرد ،سریع‌تر شود.
هرچند این الگوریتم تصاویری ایجاد می‌کند که بسیار بهتر از دیگر الگوریتم‌های غیرپارامتریک هستند، می‌توان از طریق دو ایده‌ای که در این بخش معرفی‌می‌کنیم تصاویر بسیار بهتری ایجاد کرد. اولین ایده که آن‌را فشرده‌سازی‌اولیه [6] می‌نامیم تلاش می‌کند تا در همان‌ابتدای بهینه‌سازی نقاط را نزدیک به‌هم نگه‌دارد. با این روش فاصله نقاط به یکدیگر کوچک می‌شوند و به این‌ترتیب نقاط داخل یک خوشه راحت همدیگررا پیمایش می‌کنند. برای انجام فشرده‌سازی‌اولیه یک پارامتر به‌نام $L2 - penalty$ را به تابع هزینه اضافه می‌کنیم. که متناسب است با مجذور مجوع جمع فاصله‌نقاط در فاصله اصلی.
روش دوم به‌نام مبالغه‌اولیه [7] که کمی پیچیده‌تر از روش اول است به این ترتیب می‌باشد که در آن همه پارامترهای P_{ij} را در یک عددی مانند ۴ درهمان گام‌های نخست ضرب می‌کنیم. . با انجام این‌کار q_{ij} ها که هنوز جمع آن‌ها برابر یک است بسیار کوچک تر از P_{ij} ها می‌باشند و این باعث می‌شود که مدل تلاش کند اعداد بسیار بزرگ p_{ij} را به اعداد بزرگ‌ترین اعداد q_{ij} نگاشت کند و به‌تبع باعث خوشه‌بندی بهتری برای درابعاد پایین می‌شود.
در تمامی آزمایش‌های انجام شده در این مقاله از بهینه‌سازی ذکرشده همراه با مبالغه‌اولیه با عدد ۴ برای

[5] Jcobs
[6] early compression
[7] early exaggeration

۵۰ تکرار اول بهینه‌سازی استفاده شده است.نرخ یادگیری نیز همانطور که ذکر شد به روش انطباقی در طول تکرار تغییر کرد.

فصل چهارم

آزمایش‌ها

برای بررسی عملکرد الگوریتم تی-اس-ای [1] با دیگر الگوریتم‌های زیر مورد مقایسه قرار داده شده است.

- $Sammonmapping$
- $Isomap$
- LLE
- CCA
- SNE
- MVU
- $LaplacianEigenmaps.$

این مقایسه برای پنج دیتاست انجام شده که در مقاله به سه مجموعه داده اشاره می‌شود که در زیر آورده شده است.

۴-۱ دیتاست‌ها

پنج دیتاستی که مقایسه روی آن‌ها انجام شده است عبارتند از :

- دیتاست $MNIST$
- دیتاست $Olivettifaces$
- دیتاست $COIL20$
- دیتاست مربوط به کلمات.
- دیتاست نتفلیکس.

.

در این بخش سه دیتاست اول را مورد بررسی قرار میدهیم.
دیتاست اول دارای شصت‌هزار عکس از اعداد دست‌نوشته می‌باشد. برای این آزمایش شش هزار تصویر به صورت تصادفی برای بار محاسبتی کمتر انتخاب شده اند. هر تصویر شامل ۷۶۸ پیکسل می‌باشد.دیتاست دوم شامل ۴۰۰ تصویر از ۴۰ لیوان مختلف از هر لیوان ۱۰ تصویر و هر تصویر دارای ۱۰۳۰۴ پیکسل می‌باشد که هر تصویر با توجه به مشخصاتی که دارا بود برچسب خورده است. دیتاست سوم نیز شامل ۱۴۴۰ تصویر با ابعاد ۳۲ در ۳۲ می‌باشد که از ۲۰ شی مختلف تصویر برداری شده است.

[1] t-SNE

۴-۲ نحوه‌آزمایش

در این آزمایش در ابتدا همه دیتاست‌ها با استفاده از pca به ۳۰ بعد کاهش یافته اند. این کار به دلیل بار محساباتی کمتر صورت گرفته است. همچنین همه نقاط در سه دیتاست رنگ‌بندی شده اند. این رنگ بندی صرف فهمیدن بهتر نحوه عملکرد الگوریتم‌ها صورت گرفته است.
پارامترهای تابع هزینه هر الگوریتم در زیر معرفی شده است.

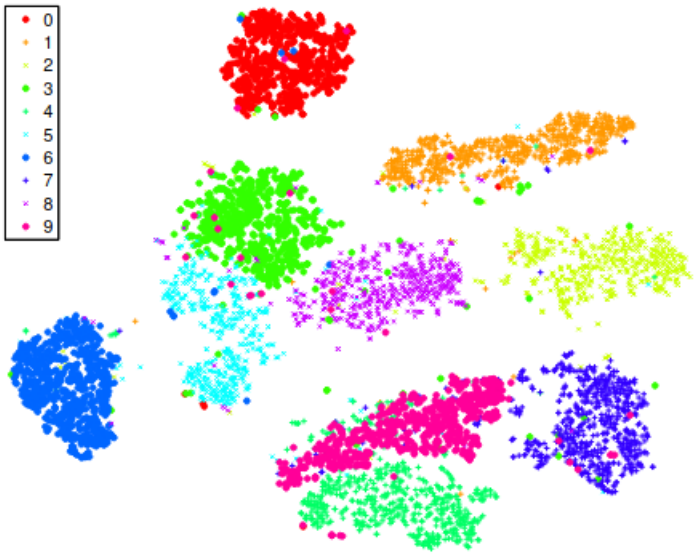
- الگوریتم $tSNE$: در این الگوریتم پارامتر سرگشتی توزیع شرطی احتمال که توسط کرنل گوسی استفاده شده برابر ۴۰ قرار داده شده است.
- الگوریتم $sammonmapping$: در این الگوریتم متد نیوتون با ۵۰۰ بار تکرار انجام شده است..
- الگوریتم $Isomap$ و LLE : در این دو الگوریتم عدد k مربوط به تعداد همسایگان نزدیک در گراف همسایگی ۱۲ در نظر گرفته شده است.

.

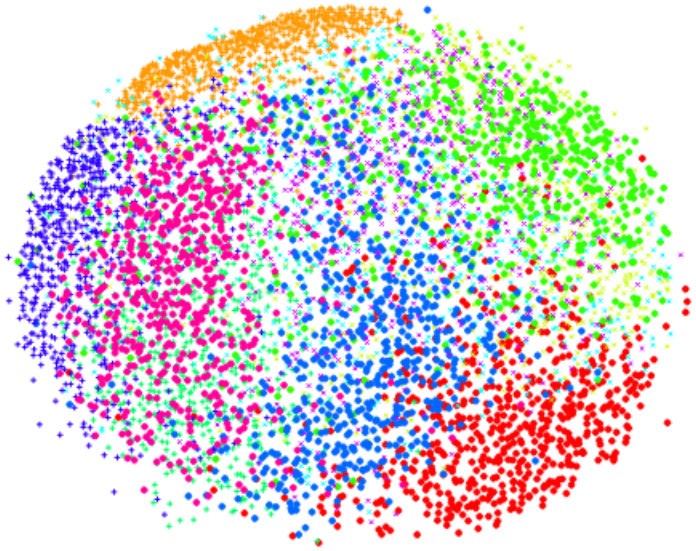
نتایج

در زیر تصاویر مربوط به نتایج آورده شده است .تصاویر به وضوح عملکرد مناسب$tSNE$ را نشان‌میدهد در به‌صورتی که تصویر مربوط به دیتاست اول دو الگوریتم $ISomap$ و LLE باعث شده‌اند داده‌ها در بعد پایین هم‌پوشانی بالایی داشته‌باشند اما در الگوریتم $sammonmapping$ تنها سه کلاس از داده های کلاستر شده ازهم جدا باشند و دیگر داده‌ها مشابه شوند ولی در مقابل الگوریتم معرفی شده توانسته است تا حد قابل قبولی داده‌های مشابه را نسب به دیگر داده‌ها نزدیک تر قرار دهد.

سه مقایسه بالا نشان داد دیگر الگوریتم‌ها به خوبی $tSNE$ معرفی شده کار نمی‌کنند همچنین لازم به ذکر است که با توجه به محاسبات الگوریتم‌های $Isomap$ و LLE در دیتاست دوم ، اعداد حاصل در نزدیک بودن کلاس‌های معرفی شده بسیار بزرگ بوده و باعث شده‌است که حتی نتواند انها در داخل یک دسته نگه دارد.

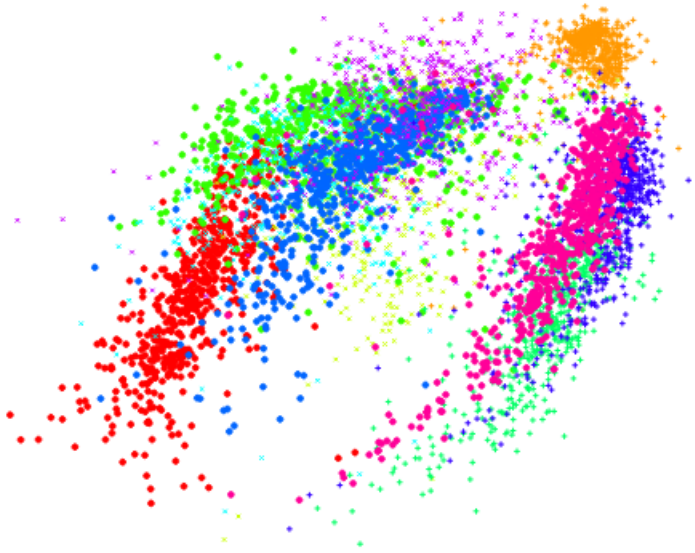


(a) Visualization by t-SNE.

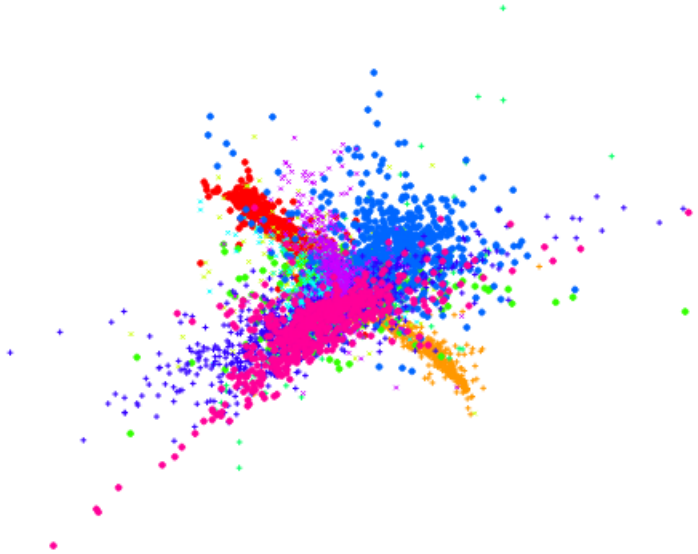


(b) Visualization by Sammon mapping.

شکل ۴-۱: نمایش ۶۰۰۰ هزار تصویر اعداد ۱ تا ۹ به صورت دست‌نویس از دیتاست اول

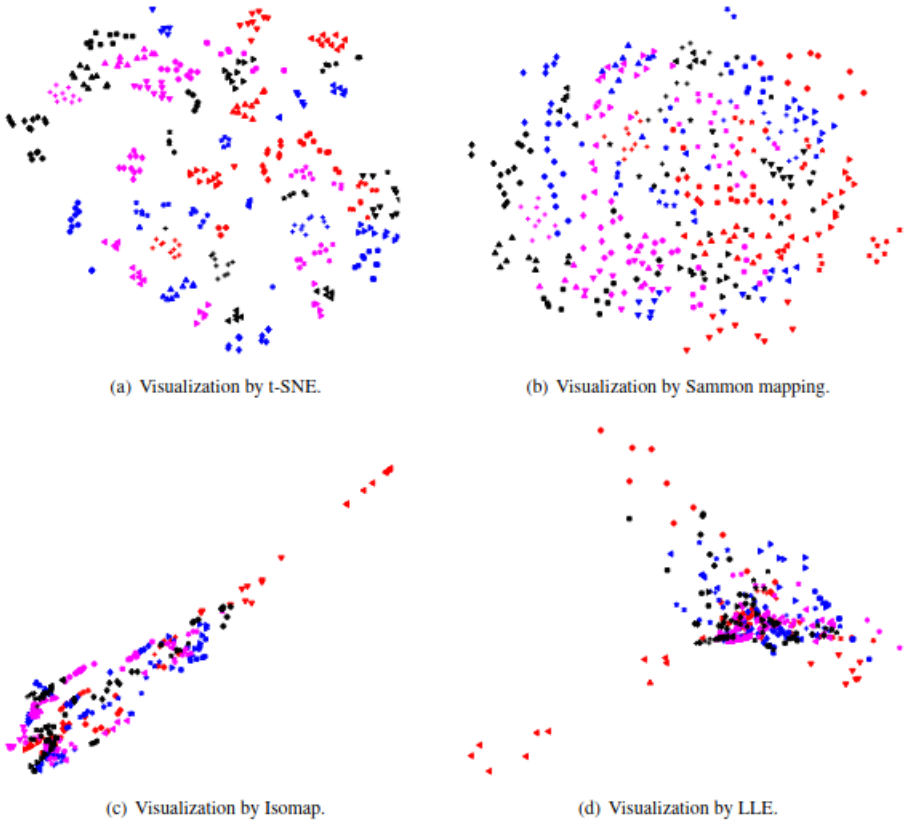


(a) Visualization by Isomap.



(b) Visualization by LLE.

شکل ۴-۲: نمایش ۶۰۰۰ هزار تصویر اعداد ۱ تا ۹ به صورت دست‌نویس از دیتاست اول



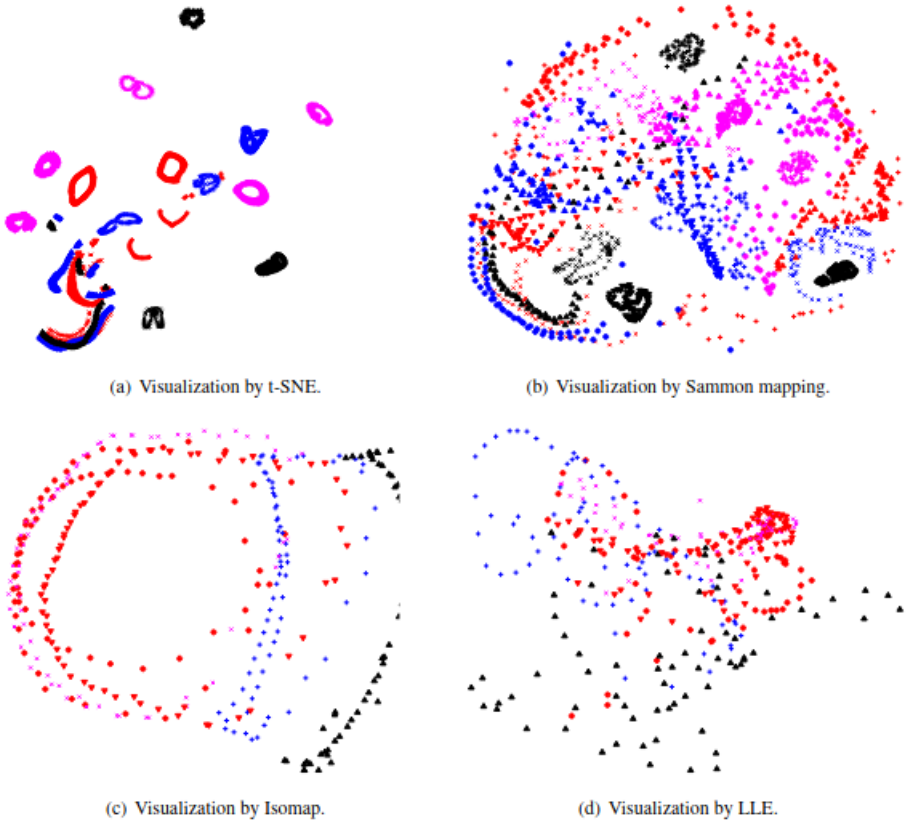
(a) Visualization by t-SNE.

(b) Visualization by Sammon mapping.

(c) Visualization by Isomap.

(d) Visualization by LLE.

شکل ۴-۳: مقایسه الگوریتم‌ها در دیتاست دوم



(a) Visualization by t-SNE.

(b) Visualization by Sammon mapping.

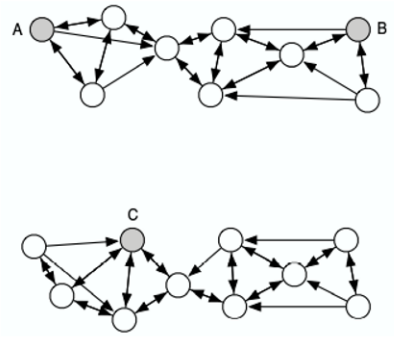
(c) Visualization by Isomap.

(d) Visualization by LLE.

شکل ۴-۴: مقایسه الگوریتم‌ها در دیتاست سوم

فصل پنجم

به‌کارگیری تی‌اس‌ان‌ای بر روی مجموعه داده‌های بزرگ

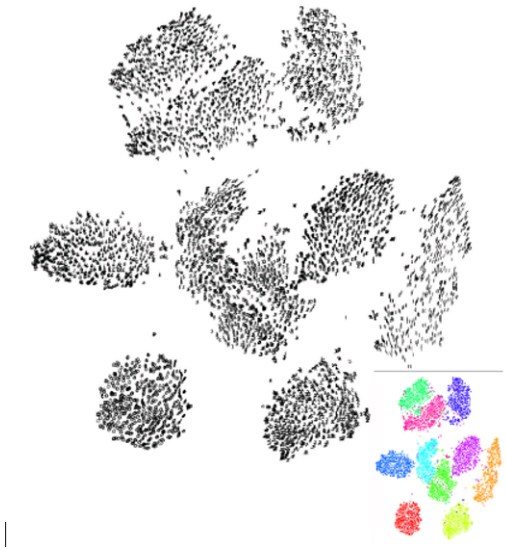


شکل ۵-۱:

جمع‌بارگیری تی‌اس‌ان‌ای بر روی مجموعه داده‌های بزرگ

مانند بسیاری دیگر از تکنیک‌های مصورسازی داده، تی‌اس‌ان‌ای از نظر پیچیدگی زمان و حافظه به نسبت تعداد داده‌های درجه ۲ است. این امر باعث می‌شود تا اعمال این الگوریتم بر روی‌مجموعه داده‌هایی که بیش از ۱۰۰۰۰ دارند غیر ممکن شود. مشخصا نمونه برداری از روی داده‌ها یک راه حل ممکن برای رفع این مشکل است؛ اما این روش نمی‌تواند از اطلاعاتی که داده‌های انتخاب نشده در نمونه برداری در مورد گوناگونی داده‌های دیتاست در اختیار ما می‌گذارند بهره ببرد. به عنوان مثال فرض کنید داده‌های A، B، C در فضای چند بعدی به صورت دو به دو فاصله‌ی یکسانی از یکدیگر داشته باشند. اگر تعداد زیادی داده‌ی نمایش داده نشده در بین A و C وجود داشته باشد اما تمام داده‌های بین A و B در نمونه برداری موجود باشند، آنگاه احتمال اینکه A و B بخشی از یک خوشه باشند بسیار بیش‌تر از احتمال هم خوشه بودن A و C است. این امر در شکل ۵-۱ نشان داده شده است. در این بخش نشان خواهیم داد که تی‌اس‌ان‌ای چگونه می تواند تغییر کند تا با نشان دادن یک زیرمجموعه تصادفی از داده‌ها بتواند از تمام اطلاعات مربوط به تنوع داده‌های دیتاست اصلی استفاده نماید.

به کارگیری تی‌اس‌ان‌ای بر روی مجموعه داده‌های بزرگ برای این کار، ابتدا یک تعداد مشخص از نقاط همسایه را در نظر می‌گیریم و یک گراف همسایگی برای تمام داده‌ها می‌سازیم. اگرچه این کار بار محاسباتی بسیار زیادی دارد، اما قرار است تنها یک بار این فرایند را انجام دهیم. سپس از هر نقطه برجسته ، به طور تصادفی شروع به قدم زدن بر روی گراف همسایگی می‌کنیم تا به یک نقطه برجسته دیگر برسیم. در حین قدم زدن تصادفی، احتمال انتخاب هر یال که از نقطه x_i شروع و به نقطه xj ختم می‌شود، متناسب است با $e^{x_i-x_j}{}^2$. مقدار $P_{j|i}$ برابر است با بزرگی نسبی تعداد قدم‌های تصادفی که از نقطه‌ی برجسته‌یx_i شروع شده وبه نقطه‌ی برجسته‌یx_j ختم می‌شود. این تعریف شباهت‌هایی با روش $Isomap$ دارد که شباهت دو به دوی نقاط را اندازه‌گیری می‌کند. با این حال، همانند نقشه‌های انتشار، به جای اینکه به دنبال کوتاه ترین مسیر در گراف همسایگی باشیم، مقیاس نزدیکی بر مبنای قدم زدن تصادفی برای تمام مسیرها با یکدیگر ادغام می‌شوند. بنابراین، مقیاس نزدیکی بر مبنای قدم زدن تصادفی،



شکل ۵-۲:

در برابر مسیرهای کوتاهی که توسط داده‌های نویز ساخته شده‌اند یعنی "$shortcircuiting$" مقاوم است. (تعدادی داده‌ی نویز می‌توانند یک مسیر کوتاه بین دو خوشه جدا از هم بسازند که باعث متصل شدن دو خوشه می‌شود.)

بدیهی ترین روش برای محاسبه میزان شباهت بر مبنای قدم زدن تصادفی این است که به طور مستقیم قدم زدن تصادفی را بر روی گراف همسایگی اجرا کنیم که این روش در عمل بسیار کابردی است، چراکه می‌توان در هر ثانیه حدود یک میلیون بار قدم زدن را اجرا کرد. در روشی دیگر، یک راه حل تحلیلی برای محاسبه‌ی دو به دوی شباهت‌ها ارائه می‌شود که شامل حل کردن یک مدل خطی غیر متراکم است. در تجربه‌های اولیه، چندان تفاوتی میان اجرای مستقیم قدم زدن تصادفی و روش‌های تحلیلی مشاهده نمی‌شود. در آزمایشی که در این بخش به آن می‌پردازیم، برای محاسبه شباهت‌ها به اجرای مستقیم قدم زدن تصادفی روی می‌آوریم چراکه بار محاسباتی کمتری دارد. اما در نظر داشته باشید که برای دیتاست‌های بسیار بزرگ که ممکن است نقاط استراژیک بسیار پراکنده باشند، روش‌های تحلیلی کارآمد تر خواهند بود. شکل ۵-۲ نتیجه یک آزمایش را نشان می‌دهد که در آن برای محاسبه دو به دوی شباهت‌ها، قدم زدن تصادفی را بر روی یک مجموعه‌ی شش هزارتایی از داده‌ها که به طور تصادفی از میان یک دیتاست شصت هزارتایی انتخاب شده‌اند اجرا شده است. در این آزمایش، از یک گراف همسایگی استفاده شده است که به ازای مقدار $K = 20$ ساخته می‌شود؛ یعنی بیست تا از نزدیکترین همسایه‌های هر نقطه انتخاب می‌شوند. قسمت داخلی شکل، کل دیتاست را نشان می‌دهد که هر کلاس با یک رنگ خاص و متفاوت نشان داده شده است. در نقشه تی‌اس‌ان‌ای کاملا می‌توان مشاهده کرد که کلاس‌های مختلف به خوبی از یکدیگر قابل تفکیک می‌باشند و علاوه بر آن، واریانس و شکل خوشه‌ها به خوبی حفظ شده است. عملکرد بسیار خوب تی‌اس‌ان‌ای را می‌توان از خطای تعمیم [1] محاسبه شده برای

[1] generalization error

الگوریتم KNN نیز مشاهده کرد. در حالیکه خطای تعمیم برای الگوریتم KNN به ازای $K = 1$ که بر روی دیتاست اصلی آموزش دیده است برابر ۷۵.۵ درصد است، مقدار این خطا برای همین الگوریتم که بر روی دیتاست‌ای که توسط تی‌اس‌ان‌ای تولید شده است آموزش دیده، برابر ۱۳.۵ درصد است. بار محاسباتی برای اجرای قدم زدن تصادفی تی‌اس‌ان‌ای بسیار معقول است. به طوریکه تنها یک ساعت زمان برای تولید شکل ۵-۲ لازم است.

فصل ششم

بررسی نهایی

$$C=\frac{1}{\sum_{ij}\|x_i-x_j\|}\sum_{i\neq j}\frac{(\|x_i-x_j\|-\|y_i-y_j\|)^2}{\|x_i-x_j\|},$$

شکل ۶-۱:

نتیجه‌های بدست آمده در دو بخش قبلی، کارایی تی‌اس‌ان‌ای را بر روی چندین مجموعه مختلف از دیتاست‌ها به خوبی نشان داد. در این بخش قصد داریم تا تی‌اس‌ان‌ای را با چندین تکنیک مطرح دیگر مقایسه کنیم و تفاوت‌های آن را بیان نماییم. همچنین به برخی از نقاط ضعف تی‌اس‌ان‌ای اشاره کرده و مواردی را برای ارتقای این روش ارائه می-کنیم.

۶-۱ مقایسه با دیگر تکنینک‌ها

امروزه پیروش مقیاس بندی کلاسیک[۱] که بسیار نزدیک به pca است، به دنبال یک انتقال خطی می‌گردد که مجموع مربعات خطاها[۲] را مینیمم کند. خطاها به صورت فاصله‌ی دو به دوی داده‌ها در فضای بالاتر و تبدیلشان در فضای پایین تر تعریف می‌شوند. یک مدل خطی مانند مقیاس بندی کلاسیک، نمی‌تواند خوشه‌های منحنی مانند را به خوبی مدل کند، چراکه تمرکز آن بیش‌تر بر روی حفظ فاصله‌ی داده‌های جدا از هم است و توجهی به حفظ فاصله‌ی داده‌ی نزدیک به هم ندارد. یک روش مهم که سعی در بر طرف کردن مشکلات مقیاس بندی کلاسیک دارد، روش نقشه برداری سامون[۳] می‌باشد که تابع هزینه را به صورت تقسیم فاصله‌ی اقلیدسی دو به دوی داده‌های انتقال داده شده به فاصله‌ی اقلیدسی آن‌ها در فضای اصلی تعریف می‌کند. بنابراین تابع هزینه برابر خواهد بود با فرمول شکل ۶-۱

که کسر بیرون از علامت سیگما برای ساده سازی گرادیان تابع نوشته شده است. نقطه ضعف اصلی این تابع هزینه این است که حفظ داده‌های با فاصله نزدیک تا حد زیادی به اندازه فاصله دو به دوی آن‌ها وابسته است. به نحوی که اگر یک خطای کوچک در مدل سازی دو داده‌ای که بسیار به یکدیگر نزدیک هستند رخ دهد، منجر به افزایش تابع هزینه به مقدار زیادی خواهد شد. از آنجایی که تمام فواصل دو به دو در شکل دهی فرم‌های محلی نقش دارند، بهتر است تا اهمیتی تقریبا برابر به فواصل به اندازه کافی کوچک داده شود.

در مقابل روش $Sammon$، روش کرنل گوسی [۴] وجود دارد که توسط تی‌اس‌ان‌ای در ابعاد بالا به کار گرفته می‌شود. این روش یک حاشیه نرم بین فرم‌های محلی ایجاد می‌کند و برای جفت‌هایی که بر مبنای انحراف معیار مدل گوسی به یکدیگر نزدیک می‌باشند، اهمیت بیش‌تری در نظر گرفته می‌شود. بنابراین اهمیت حفظ هر فاصله، مستقل از بزرگی آن است. علاوه بر آن، تی‌اس‌ان‌ای تعداد همسایگان محلی را برای هر داده به طور جداگانه و بر مبنای چگالی محلی مشخص می-کند. برتری کارایی تی‌اس‌ان‌ای در

[1] classical scaling

[2] SSE

[3] sammon

[4] Gaussian kernel

شکل ۶-۲:

مقایسه با $Isomap$ در بحث مقاوم بودن در برابر "$shortcircuiting$" به خوبی نشان داده شد. علاوه بر آن، $Isomap$ بر روی حفظ ساختارهای بزرگ داده تمرکز دارد در حالی که تی‌اس‌ان‌ای به حفظ فرم‌های محلی بیش‌تر اهمیت می‌دهد. عملکر قوی تی‌اس‌ان‌ای در مقایسه با LLE، به یک نقطه ضعف بزرگ در LLE باز می‌گردد: تنها عاملی که باعث می‌شود تا تمام داده‌ها در یک خوشه قرار نگیرند محدودیتی است که بر روی کوواریانس داده‌ها انتقال داده شده اعمال می‌شود. در عمل این محدودیت می‌تواند با قرار دادن تمام داده‌ها حول یک مرکز و چندین داده‌ی متراکم که فاصله‌ی آن‌ها از مرکز به نسبت بقیه داده‌ها بسیار زیاد است به سادگی ارضا شود. محدودیت کوواریانس در گراف همسایگی نیز می‌تواند با وجود چند ناحیه متراکم و جدا از هم ارضا شود. بنابراین روش LLE نمی‌تواند ناحیه‌های متراکم که شامل چند خوشه هستند ولی از یکدیگر دور اند را به خوبی مدل کند. به عنوان مثال شکل ۶-۲ را در نظر بگیرید، LLE قادر به تفکیک خوشه‌های قرمز و آبی و همچنین خوشه‌های زرد و سبز نخواهد بود و کل دیتاست را به دو خوشه دسته بندی می‌کند. همانند LLE ، قدم زدن تصادفی در تی‌اس‌ان‌ای از گراف همسایگی استفاده می‌کند اما مشکلات LLE را نخواهد داشت چراکه شباهت دو به دوی داده‌ها در فضای اصلی، با در نظر گرفتن ترکیبی از تمام مسیرهای بین دو داده در گراف همسایگی مشخص می‌-شود.

۶-۲ نقاط ضعف

اگرچه در قسمت قبل مشاهده کردیم که تی‌اس‌ان‌ای نسبت به برخی روش‌های دیگر برتری هایی دارد اما این روش دارای نقطه ضعف‌هایی نیز می‌باشد که اصلی ترین آن‌ها عبارت‌اند از: (۱) مشخص نیست که تی‌اس‌ان‌ای در کاهش ابعاد دقیقا چگونه عمل می-کند. (۲) کاهش ابعاد بر مبنای ویژگی‌های محلی داده‌ها، باعث می‌شود تا تی‌اس‌ان‌ای نبت به ابعاد ذاتی داده‌ها حساس باشد. (۳) تابع هزینه تی‌اس‌ان‌ای محدب نیست و بنابراین تضمینی وجود ندارد که به بهینه ترین حالت برسیم. در ادامه به طور خلاصه به شرح این نقاط ضعف می‌پردازیم.

- کاهش ابعاد برای اهداف دیگر: مشخص نیست که روش تی‌اس‌ان‌ای در حالت کلی دقیقا چگونه ابعاد را کاهش می‌دهد. این روش برای کاهش ابعاد داده‌ها به دو یا سه بعد مناسب است اما در حالتی که نیاز داشته باشیم ابعاد داده‌ها بیش‌تر از سه بعد باشد، بهتر است از آن استفاده نکنیم

چراکه نمی‌تواند به خوبی داده‌ها را مدل کند. بنابراین اگر شرایط به گونه ای باشد که برای حفظ فرم و اطلاعات کلی داده‌ها به بیش‌تر از سه بعد نیاز باشد، تی‌اس‌ان‌ای کارایی لازم را نخواهد داشت.

- نفرین ابعاد ذاتی: تی‌اس‌ان‌ای ابعاد داده‌ها بر مبنای ویژگی‌های محلی دیتاست کاهش می‌دهد که این امر می‌تواند آن را در مقابل دیتاست‌هایی که ابعاد ذاتیشان بالا است و تنوع نواحی مختلف آن‌ها زیاد است، آسیب پذیر کند. روش‌های LLE و $Isomap$ هم دارای همین مشکل هستند.

- غیر محدب بودن تابع هزینه تی‌اس‌ان‌ای: تابع هزینه در تی‌اس‌ان‌ای محدب نمی‌باشد بنابراین تضمینی وجود ندارد که به بهینه ترین حالت ممکن برسیم چراکه باید پارامترهای بهینه سازی زیادی را بیابیم که در هر اجرا و آزمایش متفاوت خواهند بود. اما این نقطه ضعف نمی‌تواند دلیلی برای کنار گذاشتن تی‌اس‌ان‌ای و استفاده از روش‌های LLE و $Isomap$ که تابع هزینه آن‌ها محدب است باشد. چراکه رسیدن به مینیمم نسبی یک تابع هزینه که ویژگی‌های داده‌ها را به وطور قابل قبولی منعکس می‌کند، بهتر از رسیدن به مینیمم مطلق تابعی است که نمی‌تواند اطلاعات مد نظر دیتاست را مدل کند. به علاوه، محدب بودن تابع هزینه به این معنا نیست که حتما می‌توان به بهینه ترین حالت ممکن رسید. چراکه محاسبه مینیمم مطلق توابع هزینه در بسیاری از تجربه‌های واقعی از نظر بار محاسباتی غیر ممکن است.

۶-۳ نتیجه گیری

در این مقاله، یک تکنیک جدید برای مصور سازی داده‌ها معرفی شد که می‌تواند ویژگی-های محلی داده‌های نزدیک به هم را به خوبی منعکس کند و علاوه بر آن بخشی از اصلی ترین ویژگی‌های سراسری دیتاست را نیز حفظ نماید. پیچیدگی حافظه و زمان روش تی‌اس‌ان‌ای برابر $O(n^2)$ است. اما رویکرد شاخصی در این مقاله مطرح شد که اجرای تی‌اس‌ان‌ای را بر روی دیتاست‌های بزرگ ممکن می‌کند. تجربه‌های بدست آمده از اجرای تی‌اس‌ان‌ای بر روی چندین دیتاست مختلف نشان داد که این روش به نسبت برخی روش‌های مطرح دیگر در زمینه مصور سازی داده بهتر عمل می‌کند. برای ارتقای این روش در آینده قصد داریم تا بهینه سازی را بر روی تعداد درجات آزادی توزیع $tStudent$ (که در تی‌اس‌ان‌ای استفاده می‌شود) پیاده سازی کنیم. این کار به برطرف کردن اولین نقطه ضعف مطرح شده در قسمت قبل کمک می‌کند. همچنین گسترش تی‌اس‌ان‌ای را به گونه‌ای که هر داده در فضای بالاتر بتواند به چندین داده در فضا پایین‌تر مدل شود، بررسی خواهیم کرد. به علاوه، هدف ما توسعه‌ی یک نسخه پرامتری از تی‌اس‌ان‌ای است که در آن با استفاده از تابع هدف تی‌اس‌ان‌ای ، به آموزش یک شبکه عصبی چند لایه‌ای که یک نگاشت مستقیم از فضای اصلی به یک فضای با ابعاد کمتر را فراهم می‌کند، می‌پردازیم.

منابع و مراجع

[1] Hinton, Geoffrey and Roweis, Sam T. Stochastic neighbor embedding. In NIPS, volume 15, pages 833–840. Citeseer, 2002.

[2] Lee, John A and Verleysen, Michel. Nonlinear dimensionality reduction. Springer Science & Business Media, 2007.

[3] Rauber, Paulo E, Falcao, Alexandre X, Telea, Alexandru C, et al. Visualizing time-dependent data using dynamic t-sne. 2016.

[4] Van der Maaten, Laurens and Hinton, Geoffrey. Visualizing data using t-sne. Journal of machine learning research, 9(11), 2008.