درس : کلان داده

نام و نام خانوادگی : فاطمه توکلی

۴۰۰۱۳۱۰۱۶

گزارش پایانی

ابتدا به صورت خلاصه به بیان مسئله و روش ارائه شده در آن میپردازیم:

در اینجا به بیان مسئله دسته بندی برای داده‌های غیر متوازن

خوشه بندی داده‌ها نامتوازن یک مشکل چالش برانگیز در یادگیری ماشین است. مشکل اصلی ناشی از عدم تعادل در اندازه خوشه و توزیع چگالی داده است. برای پرداختن به این مشکل، یک الگوریتم خوشه‌بندی جدید به نام LDPI را بر اساس پیک‌های چگالی محلی در این مقاله پیشنهاد شده‌است.

الگوریتم پیشنهادی LDPI شامل سه مرحله است:

۱) یک طرح تولید زیر خوشه اولیه طراحی شده است که می تواند به طور خودکار نقاط نویز و مراکز زیر خوشه اولیه را شناسایی کند. خوشه‌های فرعی با طبقه بندی نقاط در مراکز زیر خوشه بر اساس اصل نزدیکترین همسایه ایجاد می شوند. ۲) یک روش به روز رسانی زیرخوشه معرفی شده است که می تواند مراکز زیر خوشه کاذب را از مراکز زیر خوشه اولیه شناسایی کند. متعاقباً، طرح آنها را حذف می کند تا مراکز فرعی به روز شده و زیرخوشه های به روز شده مربوطه را به دست آورد.

۳) یک استراتژی ادغام زیرخوشه برای ادغام زیرخوشه‌های به روز شده و ایجاد خوشه‌های نهایی اعمال می شود.

۱) تعیین آستانه فاصله مناسب dc و چگالی

همانطور که قبلاً بحث شد، تخصیص دستی d_c در DPC معمولاً منجر به خوشه‌های نادرست، به ویژه با مجموعه داده‌های نامتعادل می‌شود. در راستای حل این مشکل برای یک مجموعه داده با n نقطه،اگر D_i فاصله بین نقطه i و نزدیکترین همسایه آن در نظر بگیریم، سپس مقدار حداکثر تمام فواصل بین هر نقطه و نزدیکترین همسایه آن به صورت زیر محاسبه می‌شود:

$$d_c = \max_{i=1}^{n}\{D_i\}$$

که این را به عنوان آستانه فاصله d_c در نظر میگیرد. موضوع دیگر، تعریف چگالی مناسب نقاط است.

در DPC چند نقطه ممکن است چگالی یکسانی داشته باشند و در نسخه بهبود یافته آن از معادله‌ای استفاده میکند که برای محاسبه چگالی هر نقطه بایستی تمامی نقاط در معادله گذاشته شوند و چگالی به دست آید.

در اینجا، برای غلبه بر اشکال روش بهبود یافته، به جای تمام نقاط موجود، چگالی محلی هر نقطه را با نقاط مجاور آن تعیین می‌کند، چگالی محلی هر نقطه را به صورت زیر محاسبه می‌کند

$$\rho_i = \sum_{j \in S/\{i\}, d_{ij} \le d_c} e^{-\left(\frac{d_{ij}}{d_c}\right)^2}$$

۲) تعیین نقاط نویز توسط یک گراف تصمیم جدید

نقاط نویز نمودار تصمیم تاثیر منفی قوی بر نتایج خوشه بندی دارند. بنابراین باید آنها را شناسایی کرد. در DPC، نقاطی با چگالی کمتر و معیار upward distance بیشتر به عنوان نقاط نویز در نظر گرفته می شوند، که upward distance برای هر نقطه X_i صورت زیر تعریف می‌شود:

$$\delta_i = \begin{cases} \max_j\{d_{ij}\}, & if\ \rho_i > \rho_j\ for\ \forall j \\ \min_j\{d_{ij}|\rho_j > \rho_i\}, & otherwise. \end{cases}$$

اما با مجموعه داده های نامتعادل، برخی از مراکز خوشه کوچک به اشتباه به عنوان نقاط نویز اختصاص داده می شوند. برای شناسایی صحیح نقاط نویز در یک مجموعه داده نامتعادل، یک نمودار تصمیم گیری جدید معرفی شده که از تعداد معکوس نزدیکترین همسایگان هر نقطه استفاده می‌کند. و در صورتی که سه شرط زیر برقرار بود نقطه به عنوان نویز یا داده پرت در نظر گرفته می‌شود:

$$\begin{aligned} &1) \quad \delta_i > \mu(\delta) + \sigma(\delta). \\ &2) \quad \rho_i < \mu(\rho) - \sigma(\rho). \\ &3) \quad RNN_i < \mu(RNN) - \sigma(RNN). \end{aligned}$$

که μ و δ در آن به ترتیب میانگین و انحراف معیار تمام نقاط برای RNN, densities ,upward distance هستند.

۳) طرح اولیه تولید زیر خوشه

برای جلوگیری از تخصیص تعداد خوشه‌ها به صورت دستی و فعال کردن الگوریتم برای خوشه‌بندی خودکار روی مجموعه داده‌های نامتعادل، ابتدا تعداد نسبتاً زیادی از مراکز زیرخوشه را تنظیم کرده و سپس زیرخوشه‌های اولیه را ساختیم. به طور دقیق تر پس از حذف نقاط نویز، نقاط باقیمانده که فاصله آنها بیشتر $\mu(\delta) + \sigma(\delta)$ است به عنوان مراکز زیر خوشه اولیه انتخاب می شوند. از این طریق اطمینان حاصل شود که فواصل بین مراکز مختلف به اندازه کافی بزرگ است. پس از آن، هر مرکز زیر خوشه اولیه i به یک زیر خوشه اولیه C_i مربوط می شود. تمام نقاط باقیمانده به ترتیب نزولی چگالی آنها مرتب می شوند.و به نزدیک مرکز که چگالی آن از چگالی خود نقطه بیشتر است انتصاب می‌یابند.

۴) به روزرسانی زیرخوشه

ابتدا، زیرخوشه‌های کاذب را به صورت زیر شناسایی می‌کنیم: فرض کنید که iامین زیر خوشه اولیه C_i حاوی NC_i نقطه است، و همسایگی مرکز زیر خوشه i با شعاع dc حاوی Ndc_i نقطه است. اگر C_i حاوی کمتر از نیمی از نقاط Ndc_i، یعنی $NC_i < 0:5$ Ndc_i باشد،به این معنیست که چنین خوشه‌ای یک خوشه واقعی نیست و باید حذف شود. سپس، با تخصیص نقاط در زیر خوشه‌های نادرست به نزدیک‌ترین زیرخوشه‌های همسایه، زیرخوشه‌ها به روز می‌شوند.

۵) استراتژی ادغام خوشه‌های فرعی

ابتدا نقاط مرزی هر زیر خوشه مشخص می شود. نقاط مرزی به عنوان نقاطی تعریف می شوند که چگالی آنها کمتر از چگالی متوسط زیرخوشه داده شده‌است.

برای ادغام زیرخوشه‌های به روز شده، اگر S مجموعه داده را نشان دهد، $No(X)$ مجموعه‌ای از نقاط نویز را در S و $In(X)$ نشانگر زیرمجموعه ای از X به استثنای تمام نقاط مرزی X باشد. شعاع ادغام r با معادله زیر تعریف می شود:

$$r = \begin{cases} d_c, & if\ No(S) = \emptyset \\ \max_{x_i \in S/No(S)}\{D_i\}, & otherwise. \end{cases}$$

و با استفاده از این شعاع و شروط زیر برای ادغام به خوشه بندی نهایی میرسد :

برای ادغام زیر خوشه های Cm و Cn، باید نشان دهیم $d(C_m, C_n) = \min\{d_{ij} \mid x_i \in C_m, x_j \in C_n\}$ و دو مورد باید در نظر گرفته شود:

۱) اگر $d(C_m, C_n) > r$; سپس Cm Cn و Cn دور هستند و ادغام نمی شوند.

۲) در غیر این صورت

i) اگر $d(C_m, C_n) = \min\{d_{ij} \mid x_i \in C_m, x_j \in C_n\}$، $d(C_m, C_n) = \min\{d_{ij} \mid x_i \in In(C_m), x_j \in In(C_n)\}$، سپس ~~Cn~~$Cm$ و ~~Cm~~Cm بسیار نزدیک هستند و می توانند مستقیماً ادغام شوند. ~~همانطور که در شکل ۸a نشان داده شده است، زیر خوشه های ۱ و ۲ شرایط فوق را برآورده می کنند و ادغام می شوند. به همین ترتیب، زیر خوشه های ۳ و ۴ نیز شرایط فوق را برآورده می کنند و ادغام می شوند.~~

ii) در غیر این صورت،

الف) $(x_t, x_s) = \text{argmin}_{x_i \in C_m, x_j \in C_n}(d_{ij})$ نشان می دهد. اگر مجموع چگالی x_t و x_s از میانگین چگالی مراکز زیر خوشه آنها بزرگتر باشد، یعنی اگر $\rho_t + \rho_s > \rho_{C_{mn}}$ سپس C_m و C_n ادغام می‌شوند.

ب) در غیر این صورت، دو خوشه فرعی با هم ادغام نمی شوند.

این فرآیند ادغام تا زمانی تکرار می شود که هیچ جفتی از زیر خوشه‌ها شرایط ادغام را برآورده نکنند. پس از فرآیند ادغام، اگر نقاط نویز باقی نماند، خوشه بندی کامل می شود. در غیر این صورت، هر نقطه نویز به نزدیکترین خوشه خود اختصاص داده می‌شود.

در نهایت با پرداختن به جزئیات بیان می‌کند که حل این مسئله با پیچیدگی زمانی $o(n^2)$ ممکن است.

در نتیجه به عنوان مزایا میتوان به این موارد اشاره کرد که :

۱) به هیچ پارامتر ورودی نیاز ندارد.

۲) می تواند به طور خودکار مراکز خوشه و تعداد خوشه ها را تعیین کند.

۳) برای مجموعه داده ها و مجموعه داده های نامتعادل با اشکال و توزیع های دلخواه مناسب است