سمینار درس داده‌کاوی

ارائه‌دهندگان: علی نظری – عرفان عابدی
استاد: دکتر امیرمزلقانی

لیست مقالات

- مقاله‌ی اول:

A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining

- مقاله‌ی دوم:

LoOP: Local Outlier Probabilities

A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining

- هدف
  - الگوریتمی برای Clustering داده‌های Categorical.
- ایده
  - تبدیلی بر روی الگوریتم مشهور K-Means.

- مقدمه
  - معرفی داده‌های Categrical
  - معرفی و بررسی مشکلات k-means
  - معرفی k-modes

K-Modes

- معیار عدم شباهت برای دو بردار در فضای داده‌ها:

$$d(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j) \qquad \delta(x_j,y_j)=\begin{cases}0 & (x_j=y_j)\\ 1 & (x_j\neq y_j)\end{cases}$$

- تعمیم این معیار عدم شباهت با کمک فرکانس Categoryهای مختلف در یک دیتاست (فاصله‌ی Chi-Squared):

$$d_{\chi^2}(X,Y)=\sum_{j=1}^{m}\frac{(n_{x_j}+n_{y_j})}{n_{x_j}n_{y_j}}\delta(x_j,y_j)$$

- تعریف مد یک ست از اشیای Categorical را به صورت بردار Q و تابع زیر که برای این بردار، کمترین مقدار را دارد:

$$D(Q,X)=\sum_{i=1}^{n}d(X_i,Q)$$

Finding modes

برای پیدا کردن مد Q برای دیتاست X:

- با فرض $n_{c_{k,j}}$ برای تعداد داده‌هایی که کتگوری $c_{k,j}$ را در ویژگی A_j دارند و با فرض $f_r(A_j = c_{k,j} | X) = \frac{n_{c_{k,j}}}{n}$ به عنوان فرکانس نسبی کنگوری $c_{k,j}$ در X، می‌توان قضیه‌ی زیر را ثابت کرد:

Theorem: *The function D(Q,X) is minimised iff* $f_r(A_j = q_j | X) \geq f_r(A_j = c_{k,j} | X)$ *for* $q_j \neq c_{k,j}$ *for all* $j = 1..m$.

- که نشان می‌دهد که مد یک دیتاست لزوما یکتا نیست، به عنوان مثال مد ست {[a,b], [a,c], [c,b], [b,c]} می‌تواند [a,b] یا [a,c] باشد.

K-Modes Algorithm

- پارتیشن‌بندی X به {S1, S2, .. , Sk} که هیچ S_i تهی‌ای نداریم و تعریف {Q1, ... , Qk} به عنوان مدهای هرکدام
- حالا تابع هزینه‌ی کل را به صورت زیر تعریف می‌کنیم:

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{i,l} d(X_i, Q_l)$$

- این‌جا نیز مانند الگوریتم K-Means، هدف پیدا کردن مجموعه Qi هایی است که این تابع هزینه را مینیمم می‌کنند.

مراحل الگوریتم:

1. به ازای هر کدام از K کلاستر، یک مد پیدا می‌کنیم.
2. بقیه‌ی داده‌ها را با توجه به نزدیک‌ترین مد به آن‌ها، به کلاسترها نسبت می‌دهیم و مد هر کلاستر را بعد از هر نسبت‌دهی آپدیت می‌کنیم.
3. بعد از اتمام نسبت‌دهی‌ها، باز معیارهای عدم‌شباهت را برای داده‌ها محاسبه می‌کنیم، اگر داده‌ای وجود داشت که نزدیک‌ترین مد آن در کلاستری غیر از کلاستر فعلی آن قرار داشت، آن را به کلاستر جدید نسبت می‌دهیم.
4. مرحله‌ی ۳ را تا زمانی تکرار می‌کنیم که هیچ باز نسبت‌دهی‌ای صورت نگیرد.

K-Means (initial selection)

برای مرحله‌ی ۱ الگوریتم، دو روش داریم:

- روش اول: k داده‌ی نامشابه را به عنوان مدها انتخاب می‌کنیم.
- روش دوم:
  - فرکانس همه‌ی کتگوری‌های تمامی ویژگی‌ها را محاسبه کرده و آن‌ها را در آرایه‌ای نزولی به شکل زیر قرار می‌دهیم. در این شکل C_i,j نشان‌دهنده‌ی کتگوری i از ویژگی j است.
  - کتگوری‌های با فرکانس بالاتر را به صورت یکسان میان k مد مصنوعی اولیه تقسیم می‌کنیم، به عنوان مثال، برای شکل رو به رو و k=3 این تقسیم‌بندی را خواهیم داشت:

$$\begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} & c_{1,4} \\ c_{2,1} & c_{2,2} & c_{2,3} & c_{2,4} \\ c_{3,1} & & c_{3,3} & c_{3,4} \\ c_{4,1} & & c_{4,3} & \\ & & c_{5,3} & \end{bmatrix}$$

$$Q_1 = [q_{1,1}=c_{1,1},\ q_{1,2}=c_{2,2},\ q_{1,3}=c_{3,3},\ q_{1,4}=c_{1,4}],$$
$$Q_2 = [q_{2,1}=c_{2,1},\ q_{2,2}=c_{1,2},\ q_{2,3}=c_{4,3},\ q_{2,4}=c_{2,4}]$$
$$Q_3 = [q_{3,1}=c_{3,1},\ q_{3,2}=c_{2,2},\ q_{3,3}=c_{1,3},\ q_{3,4}=c_{3,4}].$$

  - از Q1 شروع می‌کنیم، شبیه‌ترین دیتا را پیدا کرده و آن را جای Q1 قرار می‌دهیم، سپس همین کار را با تمامی مدهای اولیه انجام می‌دهیم تا همه‌ی آن‌ها با دیتاهای واقعی جایگزین شوند. (برای جلوگیری از ایجاد کلاسترهای تهی)

LoOP: Local Outlier Probabilities

روش‌های مختلف برای تشخیص داده‌ی Outlier.

- روش‌های کلی (برچسب‌گذاری)
- روش‌های محلی (امتیازدهی)

آیا این تناظر الزامی‌ست؟ خیر.

مساله‌ی مورد توجه این مقاله:

- مشکل عدم تناظر رتبه‌بندی با Outlierness – مشکل یکسان نبودن ارزش این امتیاز (توزیع داده‌های اطراف)
- نیاز به متد جدید!

Local Outlier Probabilities

رویکرد:

- فرض می‌کنیم D مجموعه‌ای از n شی است و تابع d یک تابع فاصله (مثلا فاصله اقلیدسی)
- ترکیب متدهای سنتی مثل Local Outlier Factor با مفاهیم احتمالاتی به هدف خنثی‌کردن تاثیر نویز.
- معرفی فاصله احتمالاتی! از شی $o \in D$ به مجموعه‌ی $S \subseteq D$ که آن را با pdist(o, S) نشان می‌دهیم.
- خاصیت مهم:

$$\forall s \in S: \quad P[d(o,s) \leq pdist(o,S)] \geq \psi$$

- یک کره حول o با شعاع pdist که تمامی المان‌های S را با احتمال ψ پوشش می‌دهد.
- عکس pdist یک تخمین از چگالی S به ما می‌دهد.

$$\text{Pdens}(S) = \frac{1}{pdist(o,S)}$$

Local Outlier Probabilities

- با استفاده از $\lambda = \sqrt{2}\,.\,\text{erf}^{-1}(\psi)$ به جای ψ در تخمین چگالی S، می‌توان مفهوم آماری Outlierها را به عنوان اشیایی که از میانگین بیش‌تر از مقدار $\lambda.\sigma$ دور هستند را وارد تحلیل کرد.

$$\operatorname{erf} z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt.$$

- مقادیر λ و رابطه‌ی آن با ψ از قانون تجربی سه-سیگما استخراج می‌شود.
- با فرض این‌که o مرکز S است و توزیع آماری فاصله‌های آن با بقیه نقاط s نیمه‌گاوسی است، می‌توان ملاک فاصله‌ی معیار را مشابه با انحراف معیار معرفی کرد:

$$\sigma(o, S) = \sqrt{\frac{\sum_{s \in S} d(o,s)^2}{|S|}}$$

- در این‌جا فرض شده امیدریاضی فواصل با o صفر است. به این معنا که فرض نشده S توزیع نرمالی دارد، بلکه S حول o تقریبا توزیع نرمالی دارد. (با استفاده از قضیه‌ی حد میانی در احتمالات)

Local Outlier Probabilities

- پس هنگامی که موقعیت o نسبت به S حول آن از centroid نسبی آن دور باشد، تخمین ما از Outlier بودن آن بهتر می‌شود.
- در نهایت، می‌توانیم معیار Probabilistic Local Outlier Factor (PLOF) را برای یک داده‌ی o با اهمیت λ به صورت زیر تعریف می‌شود:

$$\mathrm{PLOF}_{\lambda,S}(o) := \frac{\mathrm{pdist}(\lambda, o, S(o))}{E_{s\in S(o)}[\mathrm{pdist}(\lambda, s, S(s))]} - 1.$$

$$\mathrm{pdist}(\lambda, o, S) := \lambda \cdot \sigma(o, S).$$

- که در این‌جا لاندا و بزرگی آن با توجه به چگالی S در آن منطقه محاسبه می‌شود.
- برای تجمیع شکل چندین دیتاست مختلف، مقدار میانگین آن را به صورت زیر محاسبه می‌کنیم:

$$\mathrm{nPLOF} := \lambda \cdot \sqrt{E[(\mathrm{PLOF})^2]}$$

Local Outlier Probabilities

- و در نهایت، به این دلیل که مقدار nPLOF می‌تواند به نوعی یک انحراف معیار تلقی شود، برای نرمالایز کردن آن و تبدیل آن به یک احتمال بین ۰ و ۱، فرض می‌کنیم مقادیر ما به صورت یک توزیع نرمال با میانگین ۱ و انحراف معیار nPLOF توزیع شده اند، بار دیگر تابع ارور گاوسی را دخیل می‌کنیم و ملاک Local Outlier Probability را به صورت زیر محاسبه می‌کنیم که احتمال Outlier بودن نقطه‌ی o را نشان می‌دهد:

$$\mathrm{LoOP}_S(o) := \max\left\{0, \mathrm{erf}\left(\frac{\mathrm{PLOF}_{\lambda,S}(o)}{\mathrm{nPLOF}\cdot\sqrt{2}}\right)\right\}$$