

Subject مجلس اول
Date ۹۹/۱۱/۲۰

داده کاوی

سرفصل مطالعه

صرفی داده کاوی و فرآیند نسق داشت

مشاهدت داده ها

پاسخی و ارتقابی نهضت داده ها

پیش پردازش داده ها

استخراج الگوهای صکر

طبقه بندی

Hosseini

"احتیاج طور احتراز است."

استخراج داشت از جم ایوه داده را داده کاوی نویسد.

داشتن = الگوی جالب \rightarrow یه هر تکوینی داشتن لعنه سی سود.

الگوی جالب چه ویژگی هایی دارد؟

• صحیر است (Valid) \rightarrow روی تعداد زیادی از داده ها صحیح باشد.

• ارتقاب ناشناخته باشد.

• در جموعه داده ها به صورت صریح تعبده باشد.

• طبقه بندی کننده داده ها را درست نماید.

Data mining = knowledge discovery = knowledge extraction

= data / pattern analysis

* سایه‌گیری که حاصل از استعلام چند سرچ بی جستجوی (با طبقه‌بندی نموده شود) = پیش‌نمایش

* سایه‌گیری که حاصل از استخراج سیستم‌های خبره هستند تر داشت صنایع بهی نموده.

جلسه‌های که در صنعت داده‌لایوی به آنها پوشیده شوند

• معناش بینایی (Scalability) ممکن است روشی برای یک تعداد

داده عملکرد درست اطلاعاتی حجم دار

عملکرد نادرست داشته باشد

• پیوستان از روش‌هایی که پردازش اطلاعات روی یک سری داده

از پیش‌نمایش و خود دارد استفاده کنیم.

• یعنی اطلاعات همچو ای پردازون داده‌هایی که پردازی آنها

کار کنیم اسنه جلسه‌های صورت در این زمینه دو صور

۱) شناسایی اطلاعات، همچو ای و ۲) جمع آوری اطلاعات، همچو ای است.

collecting

identification

• کیفیت داده‌ها پیچه دهنده باشد.

• نمود استفاده از داده به چه صور باشد.

Data mining tasks:

• Descriptive methods: find human-interpretable

→ توصیفی وصفیت صریح

Patterns that describe the data

clustering

• Predictive methods: Use some variables to

→ پیش‌بینی وصفیت آینده

Predict unknown or future values

of other variables → recommender systems

* معمولاً پست از طازه دارد اما صوران

پسون این قاتر سرّعه تغیر اول را اجرا می‌نمایم.

• معنی پاکسازی داده (Data Cleaning) داده ای که در حال حاضر به آشنا نیاره داریم

: هم است ای اس که وسائلی سالمی که در حال حاضر به آشنا نیاره داریم

را در آن فرآوری دهیم است یعنی صورت این از عجیم برای داده

داده های کلیدی و آمنابی که در آینده به آشنا نیاره داریم را

در این قسمت ذخیره می‌کنیم.

• داده های که صریط بیانی که باشد انجام دهیم هستند task-relevant Data

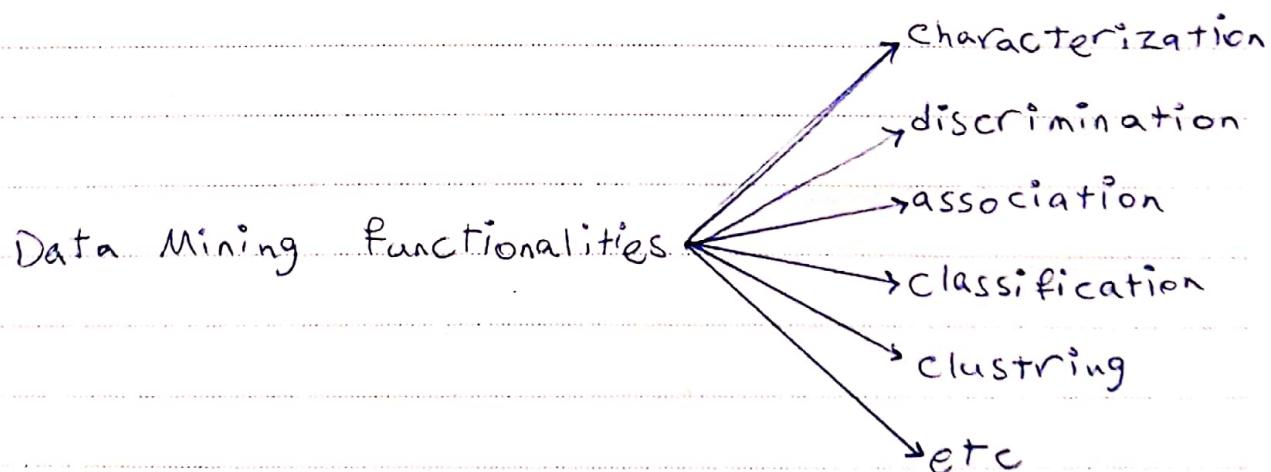
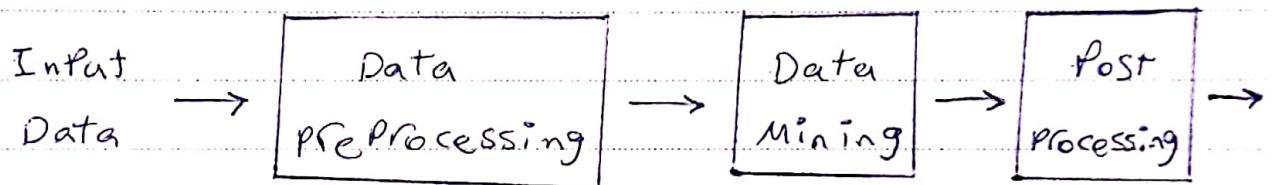
ذقاوی خاطر من می بخاطر شنیدن در پودار اسلاید ۱۵ در این است که قاتل کم

پاچه پس سانحه شد (Data cleaning) در همه صراط مملک است

مگر از شود و یه این معنی نیست که پس از هر یار ارجام این قاتل را راهی دارد

اول شروع شدم.

KDD process view from ML and statistics



آصوڑس بیسون مفتاحی

۹۹/۱۲/۱۱ جلسہ پنجم

«شناخت داده ها»

Types of Data Sets

- Record : Relational record, Data matrix, ...
- Graph and network : world wide web, social networks
- Ordered : Video data, sequential data, ...
- Spatial and multimedia : maps, Image data, ...

در این درس هدف داده یا چیزی که داده هاست.

لیست که ممکن است tuple یا record باشد مجموعه ای از Data Objects

این اساساً قیلدهای توصیف می‌کنند.

Data object از مجموعه ای از Data Set می‌باشد.

* data object also called : sample, instances, tuples, data point

اے اسے کہ وہی کہ اے data object کے data field کے : Attribute

مرکز کے انواع مختلفی دارند:

• Nominal : name, last_name, ...

• Ordinal : size{small, medium, large}, rank, ...

فرازه میں دستے ہیں ملساں ہیں۔

• Binary : Nominal attribute with only two states

■ symmetric binary & gender

 ← صفات، ← ارزش ہر دو حالت برابر است.

■ Asymmetric binary : medical test

 ← یا صفات، ← درست نہیں ہے، یعنی خاص

 ← ارزش حالت ۱ پالا تر است.

• Numeric : Quantity (integer or real-valued)

■ Interval → No true zero point

 ← نقطہ جمع و تفرقہ درانے والوں کا محتاجاً دراست. مثلاً دمای ہوا کے مقدار آئندہ اور پر برابر دھائی ایک دیناری باشندہ صریان قابل استعمالی ہیں۔

■ Ratio → Inherent zero-point

 ← ضرب و تقسیم شے درانے نوع محتاجاً دراست. مثلاً ہر دینار نصف دینار است.

* attribute also called : dimensions, features, variables, ...

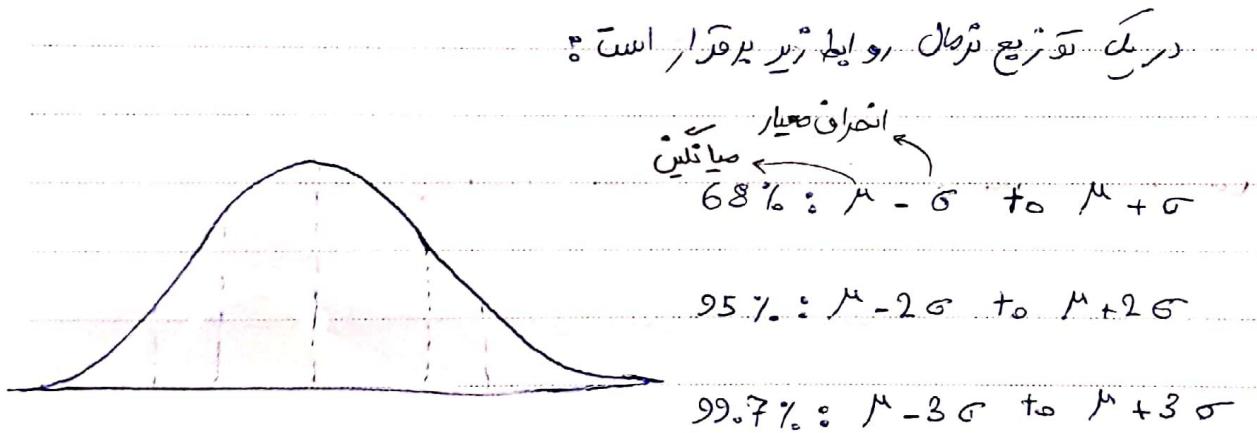
نوع دسته بندی دلیری شر برای attribute ها وجود دارد که آنها را

به دو دسته گسسته و پیوسته تقسیم می‌کند.

درین توزیع نرمال از داده ها صائبان و صایه و صدایم متنطبق هستند.

درین توزیع اریب قرموک تبریزی مقرر است:

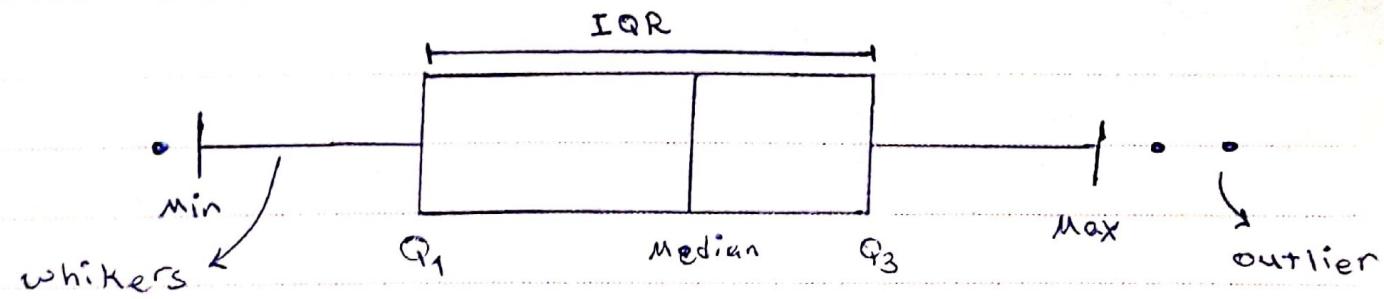
$$\text{mean} - \text{mode} \approx 3 \times (\text{mean} - \text{median})$$



(لودار جمعی ای) Box Plot

$$\text{Min} = \begin{cases} \text{Min}(x) & \text{Min}(x) > Q_1 - 1.5 \text{ IQR} \\ Q_1 - 1.5 \text{ IQR} & \text{Min}(x) \leq Q_1 - 1.5 \text{ IQR} \end{cases}$$

$$\text{Max} = \begin{cases} \text{Max}(x) & \text{Max}(x) > Q_3 + 1.5 \text{ IQR} \\ Q_3 + 1.5 \text{ IQR} & \text{Max}(x) \leq Q_3 + 1.5 \text{ IQR} \end{cases}$$



داده های بیرونی (Outlier) : داده هایی که از Min تا Max پیرامون آنها می باشد.

پس آنها وجود ندارد۔

Measuring Data Similarity and Dissimilarity

ویک صیغہ عددی اسے کہ صریح دسایت دو جلوہ، ریاضی تھا۔ Similarity

محمولاً این مقدار در پاره [0.1] اسید

1-Similarity : فاصلہ میں دو object را یاں صلی و پا اس قادہ ایسا

مطابق ملر دد.

$$\begin{array}{l}
 \text{Data Matrix} \\
 \left[\begin{array}{cccc}
 x_{11} & \dots & x_{1P} & \dots & x_{1P} \\
 \vdots & & \vdots & & \vdots \\
 x_{i1} & \dots & x_{iP} & \dots & x_{iP} \\
 \vdots & & \vdots & & \vdots \\
 x_{n1} & \dots & x_{nP} & \dots & x_{nP}
 \end{array} \right] \quad \begin{array}{l}
 \text{Disimilarity} \\
 \downarrow \\
 \text{+ singular}
 \end{array} \\
 \left[\begin{array}{cccc}
 0 & & & \\
 d(2,1) & 0 & & \\
 d(3,1) & d(3,2) & 0 & \\
 \vdots & \vdots & \vdots & \vdots \\
 d(n,1) & \dots & & 0
 \end{array} \right]
 \end{array}$$

↓
 data points
 with P dimensions

PAPCO

% attribute از ا نوع در هر یک dissimilarity, مطابق صدراً

• Nominal : $d(i,j) = \frac{p - n}{p} \rightarrow \text{number of matches}$

Simple match $\swarrow \downarrow \rightarrow \text{total number of variables}$

• Binary :

~~j~~ 1 0

1 q r

0 s t

■ symmetric: $d(i,j) = \frac{r+s}{q+r+s+t}$

■ asymmetric: $d(i,j) = \frac{r+s}{q+r+s}$

لکن زیرا ت شامل همچنین حالت ای نمی شود.

• Ordinal :

$$Z_{ij} = \frac{r_{if} - 1}{M_f - 1}$$

در این نوع طبقه بندی اعداد بین 0 و 1 می باشد که نسبت و

ادامه کار، متناسب حالت Nominal می شود.

• Mixed type:

لکن دیگر پیش ممکن است ساخت این نوع attribute ها باشد،
در این میان dissimilarity هر حالت را جدا محسوب نمی شود و
سپس یعنی نتایج حاصل میگیریں وزن دار مرتبه می شود.

Minkowski Distance Formula:

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h}$$

Special cases of Minkowski

$h=1$: Manhattan distance

$h=2$: Euclidean distance \rightarrow کمپیوتر
جوان دیستان خواندنی

$h \rightarrow \infty$: Supremum distance

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{p=1}^p |x_{ip} - x_{jp}|^h \right)^{\frac{1}{h}} = \max_p |x_{ip} - x_{jp}|$$

Cosine Similarity:

پرای بی دست آوردن شباهت دو صن اینها را تبدیل به بردار کرده و سپس حاصل

$$\cos(d_1, d_2) = \frac{\vec{d}_1 \cdot \vec{d}_2}{|\vec{d}_1| |\vec{d}_2|}$$

کسینوسی آنها را محاسبه کنیم

Data Visualization categories

Pixel-oriented
Geometric Projection

Icon-based

Hierarchical

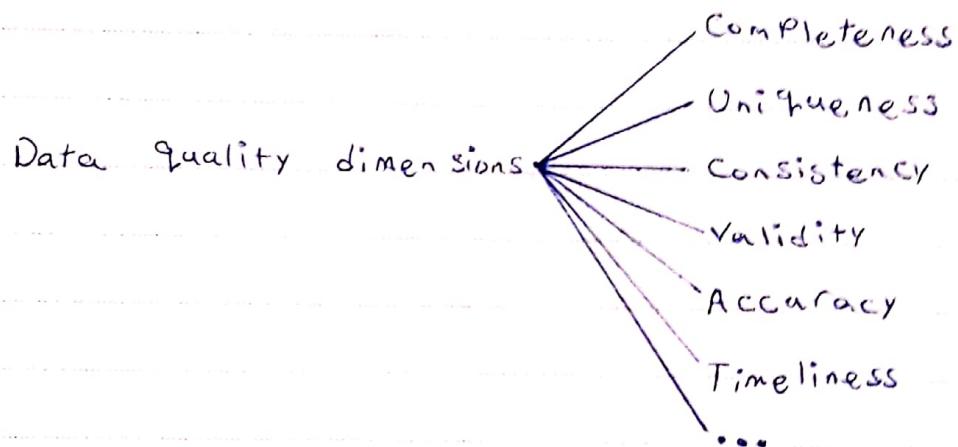
Visualizing complex data

«جیئن دادہ و یا کسائی دادہ»

کیفیت: میراث قابل استفادہ کردن (fitness to use).

data are generally considered high quality if:

"they are fit for their intend uses in operations,
decision making and planning."



Data quality types

Inherent → data quality refers
to data itself.

System dependent → data quality

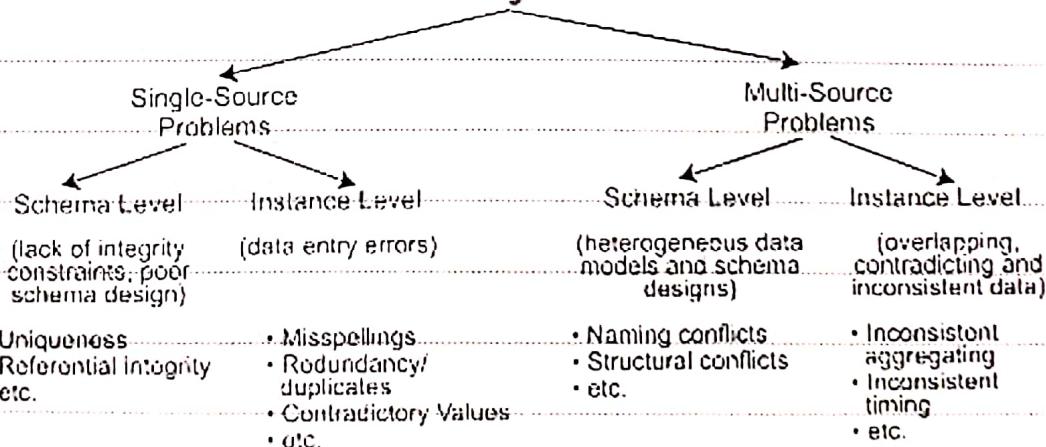
depends on the technological domain
in which data are used.

Data quality dimensions:

- Accuracy → يه صحت داده است که خود مطابق با حق است:
 - syntactic accuracy
 - semantic accuracy
- Completeness → عنی معناداده های که در این مقدار مخصوصاً است.
- Consistency → صلاحتاریخ استخراج این فرد تغیل نرتواند و بایشند.
- Validity → صلاحتاریخ، سنتیتی - مداریم.
- Currentness → صحت داده های پایه باید پرور باشند.

Data quality problems:

Data Quality Problems



Fill in missing values

Data cleaning tasks

Identify outliers and noisy data

Correct inconsistent data

How to handle missing data

- Ignore the tuple
- Fill in the missing value manually
- Fill in the missing value automatically
 - Use a global constant ; n/a
 - Use the attribute mean
 - Use the most probable value

Noisy Data:

noise: random error or variance in a measured variable

noise vs outlier: دادا نامطابق لغرض مخصوصاً، noise و outlier *

داده نویز داده ای است که بخلاف دیگر داده ها واقعی است.

* داده outlier از اطلاعات noise می باشد. (خط شکستنی در بودجه سیوسی در راه ۱-۱۰)

How to handle noisy data?

• Smooth by bin means → چالدرانی داده های یک دسته با مقدار متوسط آن دسته

• Smooth by bin boundaries → چالدرانی داده های یک دسته با مقدار متوسط آن تردیلتر هستند.

• Regression

• clustering

((پیش پردازش داده))

یه هر عملیات که مربوط به آطاکه کردن داده ها باشد اپراتور آلگوریتم داده (کاری) پیش پردازش

لعنی می شود.

Major Tasks in Preprocessing

Data Cleaning

Data Integration

Data Reduction

Data Transformation

؛ بمحض ترکیب کردن داده ها از چند طبقه داخل یک مجموعه ملیاً به است

مشکلات عدم وجود Data Integration

- Entity identification Problem →
یافتن صادر مترکی من در نوع و ترتیب اطلاعات با جاگذشت روپردازی.
- Redundancy and correlation analysis
مقدار مکانیست قابل name در موضع مطابق دو صور مختلف است، لذا.
- Tuple duplication
- Data value conflict detection and resolution

Entity identification problems

؛ معنی و نتیجہ دو مجموعه داریم کسی کو اینم قابل های آشنا باشد

تصویری نظریه روشی Schema integration
برای metadata match یا range match می باشد.

؛ object matching بس از این آشنا باشد

object مجموعه دلیل است که نه صور داریم.

Redundancy and correlation analysis:

صلانی است بین قابلیتی در صیغه وجود داشته Redundancy لـ correlation و Redundancy

باشد Redundancy عوید پر دو دلیل صلانی است از دهد.

attribute نامی صلانی است
دارای تمام صفاتی در صیغه باشد

attribute پتوان ارزیک باشد
در صیغه دارای صفت شود.

دو روش پرای تحقیقی Redundancy attribute که دارند صیغه دارد.

nominal بولی دادهای ← correlation analysis. ①

که ارتقایت های موجود در Chi-square test (ست). این سنت عددی را پرداخت

عدد که حرفی این عدد پر از احتمال آن دو صفر به هم واپسی باشد

پیشتر است. پیش از دلیل correlation بین دو مقدار با عدد χ^2 سنجیده می شون

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

correlation بجهت میانی است.

* درجه آزادی باز پراپریتی $m \times n$ کلیه میول (degree of freedom) است.

$$\text{degree of freedom} = (n-1)(m-1)$$

Numeric داده های برای correlation analysis ②

دلیل از روش های موجود در این حالت است Pearson's correlation coefficient.

این روش رابطه آماری بین دو متغیر پیوسته را اندازه لیری می کند که به عنوان

سیترین روش پرای اند اند لیری بین صفات های دلخواه ساخته شده است.

روش کوواریانس است.

Pearson's correlation coefficient روش و ترتیب های روشی هست.

محدوده مقادیر آن بین -1 (بی معنی همیشه قوی مخلوس و

+ (بی معنی همیشه قوی متسق) است.

متغیر از واحد اند اند لیری attribute است.

متقارن است.

| Negative | | | No correlation | Positive | | |
|----------|----------|------|----------------|----------|----------|--------|
| Strong | Moderate | Weak | | Weak | Moderate | Strong |
| -1 | -0.8 | -0.5 | -0.1 | +0.1 | 0.5 | 0.8 |

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{(n-1) \sigma_A \sigma_B} = \frac{\sum_{i=1}^n (a_i b_i) - n \bar{A} \bar{B}}{(n-1) \sigma_A \sigma_B}$$

n = number of tuples

\bar{A} and \bar{B} : means of A and B

σ_A and σ_B : standard deviation of A and B

$\sum (a_i b_i)$: sum of the AB cross-product

numerics \rightarrow $\text{cov} \leftarrow \text{Covariance}$ ③

$$\text{cov}(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

positive correlation \rightarrow If $\text{cov} > 0$

negative correlation \rightarrow If $\text{cov} < 0$

* در صورتی که مسئله پاسخ ندارد، $\text{cov} = 0$ (اسناد ملخص این قضیه حمیشہ درست)

مسئله پاسخ ندارد، $\text{cov} = 0$ (از آنها پس از اسناد ملخص این دو مقدار نیست)

Tuple duplication:

چنانچه موجود صدای که در دو مجموعه مختلف کلید یکسانی برای ادثام قادر نبود و وجود

داداشت پذیرم

Data value conflict detection and resolution:

نحوه ارزیابی از خلاف

- Differences in representation, scaling or encoding

وجود واحدهای اندازه‌گیری
 مختلف (دقیق و تأثیرگذار) ← ■ Difference on the abstraction level

miss detection یعنی ناهم حجم داده در حالتی که فقط صفت سعد و صفت

داداشت پذیرم سه اسکرینی برای این که وجود در داده

- Dimensionality reduction

از متش صفت‌ها و تشدید از صفت‌های صفت

- Numerosity reduction → کاهش تعداد کور رها

- Data compression

Dimensionality Reduction:

هفت این نویس، که هشت حجم داده استصراف صفات صم، ۷ هشت و نیز، امر اسن

سرعت بردارش داده و یکی از را صفر است که حود میتواند پر

۴ تکنیک اسن

و تکنیک موجکاری موجداری Wavelet transforms.

مسوی سی پایه ای حسین اسناده

ارتعاشات صفات و صفری موجداری علیش.

و تحلیل کا صفت PCA.

نمایش صور ۳ بعدی به دو بعدی و حذف بعدی اهمیت را

در حال تکیه n بعدی به m بعدی

* (عن روش نقطه برای داده های numeric) عالی اسناده است.

و در این روش برای Attribute Subset Selection.

که هشت حجم داده ها و که هشت ابعاد، برای مجموعه ای از کل صفات

که در این صفات داده های صفت داشتند، انتخاب کنیم. کرام صفت ها را

نادیده می نماییم؟

Redundant attributes.

Irrelevant attributes.

P4PCO

اگر می‌دانیم صفت‌های مختلف ارائه شده‌اند، ۲ ترکیب مختلف ارائه شده‌اند که یکی می‌تواند حاوی این صفات باشد،

آنرا که روش‌های مختلفی برای استخراج یک مجموعه مناسب وجود دارد، معمولاً

حریمانه (greedy) نامیدند.

Typical heuristic methods:

- Step-wise forward selection
- Step-wise backward elimination
- Combination of forward selection and backward elimination
- Decision tree induction

• (Attribute creation) Feature generation •

در این روش پاترنس می‌توان صفت‌یابی صفت‌های جدید پژوهشی ایجاد کرد.

مثلث: جایگزینی مساحت به طی طول و عرض در یک پایه داده

Numerosity Reduction:

کاهش داده در این روش صیغه یک متد کلی است:

- Parametric methods (e.g. regression)

در این روش مثلاً پایه‌ی مقدارهای داده‌ها سما پارامترهای آن مقدارهای را تفسیر می‌کند.

PAPCO صیغه و انتزاعی قوی داده‌ها بین تبارهای مشویم

■ Regression Analysis

تحصیلی از داده ها بر اساس داده های موجود رُدْه صَحِّه شود و نک best fit

از داده ها به دست آید و با راصت های این مُطْبَق تفسیری صَحِّه شود.

• Non-Parametric

در این روش ها عیچ سدلی و معود مدارد یا که رُدْه صَحِّه شده (از داده ها استحاب صَحِّه شود).

■ Histogram Analysis:

داده ها به چند دسته تقسیم شده و صَلَیق هر دسته تفسیری صَحِّه شود.

أَوْاعِيَّ تَقْسِيمٍ بَيْنَ داده ها

equal bucket range → equal width •

equal frequency ← equal depth •

■ clustering:

داده ها به چند صُونَسْ تَقْسِيم شده و صَرُبْه حَوْسَه به عنوان مُعايِّدَه آن صُونَسْ

مُطْبَق تفسیری صَحِّه شود.

■ Sampling

نمونه از داده ها به صورت تصادفی کل داده ها محسوب نموده.

♦ Simple random sampling

with replacement without replacement

♦ Stratified sampling

نمونه از داده های جدید و قدیم باستفاده از تابع $f(x)$: Data Transformation

→ A function that maps the entire set of values

of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values.

⇒ data transformations / بروش

ویرایش داده ها

با طبقه بندی و ساخت اصلیتی : Attribute construction

اجماع اطلاعاتی : Aggregation

range پردازشی داده ها : Normalization

متقطع سازی داده ها : Discretization

Normalization:

• Min-Max normalization

$$v' = \frac{v - \min_A}{\max_A - \min_A} \cdot (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

مثال: ١٣,٩٠٠ درجة حرارة [٢٠, ١٠] إلى [٩٨,٠٠] إلى ١٢,٠٠٠ درج حرارة: $v' = ?$

$$v' = \frac{13,900 - 12,000}{98,000 - 12,000} (1-0) + 0 = 0, 114$$

• Z-score normalization

$$v' = \frac{v - \mu_A}{\sigma_A}$$

mean
standard deviation

مثال: اعداد قبل را پرای این فضول استفاده کنیم:

$$v' = \frac{13,900 - 12,000}{19,000} = 1, 492$$

• normalization by decimal scaling

$$v' = \frac{v}{10^j} \rightarrow j \text{ is the smallest integer such that } \max(|v'|) < 1$$

Aggregations

Data cube construction

The diagram illustrates the process of creating a data cube from a detailed sales table. On the left, a large table is shown with columns for Year, Quarter, and Sales. The rows are grouped by Year (2008, 2009, 2010) and further subdivided by Quarter (Q1, Q2, Q3, Q4). An arrow points from this detailed table to a smaller summary table on the right, which contains only the total sales for each year.

| Year 2010 | | |
|-----------|-----------|--|
| Quarter | Sales | |
| Q1 | \$224,000 | |
| Q2 | \$408,000 | |
| Q3 | \$350,000 | |
| Q4 | \$586,000 | |

| Year | Sales |
|------|-------------|
| 2008 | \$1,568,000 |
| 2009 | \$2,356,000 |
| 2010 | \$3,594,000 |

Discretization:

- Binning → Top down split, unsupervised
- Histogram Analysis → Top down splits unsupervised
- Clustering Analysis → Top down split or bottom up merging, unsupervised
- Decision-tree Analysis → Top down split, supervised
- Correlation Analysis → bottom up merge, unsupervised

Concept Hierarchy Generations

کے مبنیہ صفاتی رُنگاہیں را پڑائی طے ایجاد کر کر کے این مقامیں حفاظت کرنے والے

حستہ، این روشنی یہ صورت یا رُنگیں صیم را دادہ را پوسیلے جاتی ہیں مگر یہ مکانیں سطح پر ہیں

کا حصہ صدھر، این مبنیہ صفاتی کا راصح برائی پر راصح توسط صاحبصنان آئندہ رُنگوں

یا طرح اسے data warehouse مخصوص کر دے

Example of numeric data: age → youth, adult, senior

Example of nominal data: address → street < city < province < country

* بعضی از مبنیہ صفاتی ہمارا صیم تو ان پر اساس تحریری و تحلیل ممکن صفاتیں ہر وہیں ہی طور

جود کار، جود نکرد، مثلاً صفت حادی کے صفاتیں صفاتیں جیسا کہیں داری درستھے پاہیں تری

اڑ مبنیہ صفاتی فرگر صیم لیں گذل استثنائے روز خای ہندی، طاہ، سال، وہیں وہیں

| | | |
|----------|---------|-----------------|
| Country | 15 | distinct values |
| ↓ | | |
| Province | 365 | " " |
| ↓ | | |
| City | 9567 | " " |
| ↓ | | |
| Street | 674,339 | " " |
| PAPCO | | |

پیش بوداریش یا پایان‌نیون

جلسه دوازدهم - ۰۰/۱/۳۰

«استخراج الّوّهای طریق»

نه مختص این است که باید الّوّهی در پیش داده‌ها دائمًا حالت تکرار است و

ساخته این داده‌ها برای ما می‌باشد، پر اهمیت است زیرا می‌توانیم پر اساس

این الّوّهها توصیم کنیم. این الّوّهی تواند صنایع ای را آفّق‌ها باید زیررساند.

باید ساخته، بیش از ۵۰۰۰ باشد.

از جمله کاربردهای این روش catalog-design basket data analysis

و click stream analysis.

ایده اصلی این که از تحلیل سبد خرید مستقریان بک روشهای سروغ شده است.

متاتابع (ولی):

Dataset: $I = \{I_1, I_2, \dots, I_m\}$

Transaction set: $D = \{T_1, T_2, \dots, T_n\} \rightarrow$ مجموعه خرید

K-itemset: $X = \{x_1, \dots, x_K\} \rightarrow 2\text{-itemset} = \{I_1, I_2\}$

Support count: Frequency or occurrence of an itemset

Support: relative support: the fraction of transaction that contains X

minsup ای support count است اگر (frequent) می داده صفر است

بررگردانی باشد در حال کلی:

An itemset X is frequent if: X's support \geq minsup

Confidence

حداری است که شاند دیده بیرون از خود آیتم A باعث خود آیتم B شده است.

* support is the percentage of transactions in D
that contains $A \cup B$

$$\text{support } A \rightarrow B = P(A \cup B) = \text{support_count}(A \cup B) / n$$

* confidence is the percentage of transactions in D
containing A that also contain B

$$\text{confidence } A \rightarrow B = P(B|A)$$

$$\text{confidence } A \rightarrow B = \text{support_count}(A \cup B) / \text{support_count}(A)$$

Association Rule

اگر طبقاً بهمین این association rule را استخراج نمی دو صراحتاً کار پایه انجام دهن

itemset بزرگتر

• تولید همان اینضیبی قوی ای روی می بزرگار

→ rules must satisfy minimum support
and minimum confidence

اگر صرطیه اول را تجاه ندهیم علاوه بر افراد حجم محاسبات بیان مولید نقدایی

فایل invalid می شود.

تعریف دلخواهی سبق همان انتصافی به شکل زیر است:

Find all the rules $X \rightarrow Y$ with minimum support and confidence

مثال: پایی الگویی $\text{minSup} = 50\%$, $\text{minConf} = 50\%$ صورت پرداز dataset

| T_id | Items bought |
|------|----------------------------------|
| 10 | Beer, Nuts, Diaper |
| 20 | Beer, Coffee, Diaper |
| 30 | Beer, Diaper, Eggs |
| 40 | Nuts, Eggs, Milk |
| 50 | Nuts, Coffee, Diaper, Eggs, Milk |

برکار و فوائین انتصافی را باید

Frequent Patterns:

{Beer}: 3, {Nuts}: 3, {Diaper}: 3,

{Eggs}: 3, {Beer, Diaper}: 3

Association Rules:

$$\text{Beer} \rightarrow \text{Diaper} = \frac{3}{3} = 100\% \rightarrow (60\%, 100\%)$$

$$\text{Diaper} \rightarrow \text{Beer} = \frac{3}{4} = 75\% \rightarrow (60\%, 75\%)$$

نکته در حدود چهل درصد از Beer, Diaper را مربوط به support count

این rule قوی است.

روش پیاپی شده در جلسه قبل یهایی حجم داده های یالا رجبار، اشکال حواهد شد و

حجم محاسبات فوق العاده زیادی را پیهدمراه حواهد داشت. صنعاً اگر ط

در یک قروشله میباشد ۱۰۰ قلم کالا داشته باشیم تعداد sub-Pattern هایی که

پایه دررسی کنیم در حدود $1.027 * 10^{30}$ است که صد اربیلیون بیاری است در صورتی

که اگر این قروشله ها صد اربیلیون بیشتر از ۵۰٪ اصل کالا را درست نمی‌هی همین دلیل

هزار به استفاده از یک روش جدید برای یافتن الکوهای پرکلار خسیم.

Closed and Maximal Frequent Itemsets

- An itemset X is closed if

- X is frequent

- no super-Pattern $Y (X \subset Y)$ with the same support as X

- An itemset X is maximal if

- X is frequent

- no frequent super-Pattern $Y (X \subset Y)$

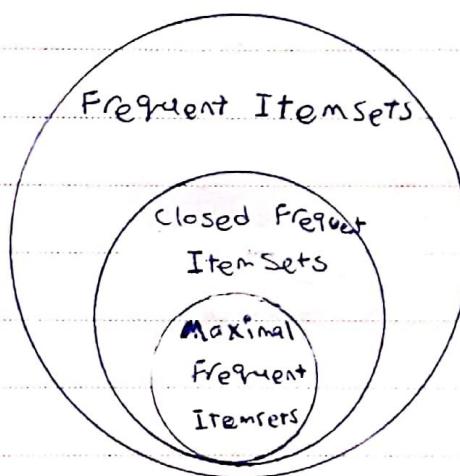
* پس صریح ها and وجوه حاره.

يُعرف $Y \subset X$ بـ Y مماثلة لـ X if Y super-pattern يُعطى support مقدار +

• $\text{sup}(X)$

يعني X هو مجموعة Y super-pattern *

Relationship between frequent itemset representations:



Frequent Itemset mining methods:

- APriori (Candidate generation)
- FP Growth (Projection-based)
- ECLAT (Vertical data format)

• Apriori

ویرگی سیستم پودن (downward closure) : صفر زم صحیحه از نک آیم سنت

ملکر، حتی ملکر است = ملک

if {beer, diapers} is frequent, so is {beer, diaper}

پایه ب ویرگی فوچ هنلیا که آیم سنت عای اعضا که ملکر را پیدا کنیم،

پایه ب ا کردن الکوهای طریق ۲ عضوی و ۰۰۰ ملک عیش لالکوهای اعضا

را با هم بررسی کنیم و میار به بررسی دلیل عضو حا میشیم.

الگوریتم Apriori سر با استفاده از همین ویرگی تعداد scan ها را کاهش

می دهد و می تواند به جای scan م ترالکس ها در هر صفحه نافیس از میں

الکوهای که در صفحه قبلاً استخراج شده الکوهای A+K کادی بیدا کنیم.

* بر عکس ویرگی سیستم پودن لذت بر عکس ایست همین اگر نک آیم سنت ملکر بپاس هر

مجموعه دارای این زم صحیحه سر ملکر نیست

ویود دسته پاسخ نه itemset کل اگر : Apriori Pruning Principle

باشد، superset، آن بیانیه قویید / پرسنی شود.

teriori algoritmi Pseudo code:

$L_1 = \{$ frequent items $\}$

for (K=1 ; L_K != Ø ; K++)

C_{K+1} = candidates generated from L_K

for each transaction t in database

increment the count of all candidates in C_{k+1} that are contained in t .

L_{K+1} = candidates in C_{K+1} with min_support

return $U_K L_K$

بَوْلَيْدَ قَوْسَةَ الْحُمَّةِ :

تعداد قوامی ب پرایل صحیح سه K-itemset ای میتواند باشد

است۔ پیاراں پر توجہ پر صدھن سدھن frequent itemset ہے ایک حاصل ہے۔

اگر maximal حاکم آنها باشد، جدای از آنها و همچنان رابطه‌ای آنها با سایریم.

مسئلہ اول اوریم : Aprori

• تعداد transaction-set کی scan ریکارڈ (سے) *

• تعداد candidate ہائے ایجاد شدہ ریکارڈ اسے *

• نصوہ سے، شرکتی support کا مطلبی (سے) *

روشنی اوریم : Aprori

Hash-based *

Transaction Reduction *

Partitioning *

Sampling *

Dynamic itemset counting *

• اینہ اپنے روشن کا k-itemset کا عدد candidate ہائے ایجاد شدہ ریکارڈ اسے *

درست روشن بے اسقادہ اریکی کاچھ جو میں صورتیں، transaction set کا hash

نلاسٹ صیغہ * سب سے متداول ہے اس کا frequent itemset یعنی 2 کا یعنی صرف میں

ایسا ہے کہ ہمارے کو hash کا کچھ دھرمی را درجول دھرمیں دیں۔

درست روشن کے صورتیں از کلیا، کلرا، سودھریان، backet count، افراش صدھیم

PAPCO

Scan این روش کا هش تعداد ترالس های Transaction Reduction

ای transaction طای بعدی است. این روش می تواند دسته در iteration طای بعدی است.

که در هر مرحله ای پر تکرار در آن وجود ندارد در مرحله انتها

scan ای کا $(K+1)$ -itemset

ایدی این روش کا هش تعداد Scan طای است. طبق این روش طای

کار را به صورت صفاتی انجام داده و

استخراج می کنیم. در این روش ابتدا database بین قسم شده و

سین به صورت موازی داخل این یعنی k طیار می کردیم و پس

از آن الگوهای یافت شده، پلیپار دیلم یا کل داده ها پرسی کردیم که

scan کا $(k+1)$ -itemset کل عملیات با ۲

ب انجام می رسد.

در این روش k -هاش تعداد scan هاست Sampling for frequent Patterns.

در این روش الگوهای مکرر را برای یک نمونه کوچک از داده ها به دست آورده

و سپس این جمجمه itemset را در کل داده ها حضور یافته و مجموع آن را

این لکوچ در کل داده ها می‌نامند. در این روش فقط maximal

عنوانی نمونه داده ها را بررسی می‌کنند یعنی اگر نمونه داده ها abcd بوده

پس از بررسی ab و ac و ... و abcd را بررسی می‌کنند.

در این روش k -هاش تعداد scan و ایده این روش Dynamic Itemset counting است.

در این روش طبق شلی پویند اقدام به سراسر itemset ها می‌کنیم. در الگوریتم اصلی

پس از k -هاش frequent یا پیوستی itemset این داشت و استناده ای

از تعداد تکرار آن نمی‌کردیم. در این روش ما به صفحه اینکه از که در حال

بررسی آن می‌ستیم تعداد تکرار آن از minsup پیشتر نبود عملیات scan را برای

آن بپایان رسانده و سرانجام itemset بعدی را بررسی می‌کنیم.

• Vertical data format (ECLAT)

در روش این، دستگاه که هر کدام را می‌شود تعدادی transaction که بعد از APriori

آیتم دارد که این روش‌ها خاصیتی می‌سازند.

روش دیگری که پروسه می‌شود است که ساختار داده

در آن از اتفاقی به عمودی تبدیل می‌شود پسین صورت که طبق دیگر ناچیز تعداد

transaction‌ها مستقیماً در این صورت بجا آنکه مستحبم کنید در هر transaction

چه آیتم‌هایی موجود است و مستحبم می‌شوند هر آیتم در چه transaction‌ها

آمده است.

در این روش ابتدا یک را پایه گذاری می‌کنند که دارند که این را scan می‌کنند.

transaction set کو support count در این روش vertical

می‌شود. سپس در ادامه می‌باشد که روش تبلیغاتی K-Itemset می‌باشد که می‌توانند

1-itemset می‌باشد که از پیدا کردن 1-itemset می‌باشد که این 2-itemset می‌باشد.

که می‌باشد که این 2-itemset می‌باشد که این 3-itemset می‌باشد.

بروز رسانی های وسیع

- از مهان روشن candidat‌های پرایوری تولید (ستاده صفر)
- پرایوری پیدا کردن K+1-itemset دویا، به پایانه داده نماید.

صفحه وسیع

transaction set که طبقه بندی خواهد شد

بررسی فواید اینچی که این اسلوب مخصوص کافون استقرار چشیده از نظر

داشتن صراحت و min_support و min_conf کافون درست

بیست. این اتفاق رفته رفع درجه که آیتمها با یکدیگر وابستگی داشته باشند می‌دهند

ظهور علاوه بر صفات support و confidence مفهوم

دیگری با عنوان را بنویسیم که این بررسی تأثیرات این وابستگی روی آیتم را محسوس نمایم.

$$\text{lift } A, B = P(A \cup B) / P(A) \cdot P(B)$$

- $\text{lift} > 1 \rightarrow \text{positive correlation}$
- $\text{lift} = 1 \rightarrow \text{independency}$
- $\text{lift} < 1 \rightarrow \text{negative correlation}$

پر اسٹوڈنٹ این صڑاں آئے lift < 1 یا شیش یعنی حُریٰ بیل مخصوص نہ سما یا عتھِ حُرید

یک مخصوص دلیل نہ شدہ بلکہ یا صت عدم حُرید آئے مخصوص نہ شدہ است۔ حُریٰ جیسے

lift = 1 پر معنی عدم کائسِ داری آئیم ہا بِ رحم است۔

((یادِ دلیل) پدون ناظر و پاناظر)

در لَدَسَری بِ ناظر دادہ ہا دارای اکٹا مخصوص و دادہ ہا پر اسٹوڈنٹ training set

label classifier میں شود در صورت کہ در لَدَلِیل پدون ناظر دادہ ہا پدون اکٹا

محض و عدف قرار دادہ دادہ ہای صباہ در کلاس یا صونہ محض است۔

اکٹا دادہ کا کا ہا عمرہ آئیں مخصوص کا کا ہا نہیں نہیں بیک مخصوص یا بیک منبع حارجی

کوئی 5 label ہا ریجیسٹر کند۔

روش ہای supervised learning میں اکٹا این اکٹا را پہ طبق دھند کے آئندہ ای

کہ صورت label مدارد را label پریس۔

* معم روش ہای unsupervised (رُنوج frequent pattern) *

« طبقه بندی »

روش های طبقه بندی پر ۳ روش عالی (رات حسنه)

Classification rule

Decision trees

Mathematical formula

classifier : یعنی یک مجموعه داده آموزشی درین که برای یادگیری

استفاده شود که داده های دارای label حسنه

طبقه بندی طوری که داده های جو نام روند است

Model construction

→ training test job

Model usage

classifier درصد داده های ایست نه به درستی و نه خطأ Accuracy rate

طبقه بندی نسبه اند

* داده های آموزشی و test باید از یک مسئله باشند.

: Classification vs Prediction

رُهانی که داده های یا لسته (nominal یا کategorical) یا اتال (label)

داده های لسته برای label یعنی داده های جدید استاده ممکن است اینها را

لسته می شود اما آن چیزی که قرار است با پیش پیش نیم می صدر از classification

پیوسته (مثل مبلغ دلار درصد آینده) یا آنها می توانند اینها را numeric prediction می نامند.

با اینکه می شود اینها را prediction نامند اینها را انجام داده ایم.

* برعکس prediction و classification داده های ترسیم ممکن است.

معماری روشی روشی طبقه بندی:

دقّت ← Accuracy .

؛ توأمی classifier در پیش پیش لسته

؛ صیران محدودی مقدار پیش پیش شده با واقعی predictor .

سريع - شامل دو بخش سرعان یا دیری و سرعت طبقه بندی Speed .

missing values noise handle ← Robustness .

کارآمدی در با حجم داده های پرگ Scalability .

صیزان قابل در یوندن برای ما Interpreability .

: Decision Tree

درخت تعلم متشکل از یکن حتی ریز است.

پایه‌ترین دره، بیان لز و شرطی ها داده هاست
که بیشترین خاصیت صفاتی کشید را دارد.

گره‌های درونی، ویرگ طی داده ها

نهادها، شامل مقادیر ویرگ ها

labelها ← Leaf nodes .

هم الوریق عای صویغ برای ساخت درخت تعلم صفاتی حفظ آنچهی کر

آنرا از یکدیگر صفاتی که معناری است که با آن درخت را صفتی می‌سازد.

درخت از بالا به پایین و با روکرد تقسیم و حل ساخته شود.

شرطی توقف الوریق ساخت درخت تعلم:

• همه داده هایی با هماننده صریط به یک لاس باشند

• ویرگ اسی برای بررسی باقی نهاده باشد.

• داده ای برای بررسی باقی نهاده باشد.

در حالت ایده‌آل هدف طبقه‌بندی است که با درخت تخصیم به پایه‌تیس‌ها) خالص برسیم

محیط‌رهای اندیشه‌لری صیران حوب بودن یک ریتtribute

بر اساس مفهومی به نام entropy و information gain .

در الگوریتم‌های ID3 و C4.5 استفاده می‌شود

بر اساس اصل ناهمogenity یعنی Gini Index .

الگوریتم CART صورت استفاده حرار می‌لیرد.

لین روش یک روش بیوپاوت از Information gain است.

پر اساس کاهش واریانس محول نموده و در درخت‌های Variance Reduction .

بررسیون صورت استفاده حرار می‌لیرد.

entropy می‌خواهی است که عدم تعطیت را برای یک صیران تصادفی سؤال می‌دهد

و هدف آن شناساندن صیران به تطمیع است که با استفاده از فرمول زیر

پر جست می‌آید:

$$H(x) = \sum_{i=1}^m p_i \log_2(p_i)$$

entropy of x

* در درخت‌تخصیم طبقه‌بندی صفتی هستیم که ییشتن کاهش بی تطمیع یا ییشتن IG

P4PCO

با داشت پاسخ

FP

Calculate Information Gain:

- Step 1: compute expected information (entropy) of the current partition $\text{Info}(D)$

$$\text{Info}(D) = - \sum_{i=1}^m P_i \log_2 (P_i) \quad \rightarrow \text{number of classes}$$

- Step 2: compute $\text{Info}_A(D)$

\rightarrow the amount of information would we still need to arrive at an exact classification after partitioning using attribute A

$$\text{Info}_A(D) = \sum_{j=1}^r \frac{|D_j|}{|D|} \times \text{Info}(D_j) \quad \rightarrow \text{number of partitions}$$

- Step 3: compute IG of attribute A

$$G_{\text{ain}}(A) = \text{Info}(D) - \text{Info}_A(D)$$

اُن دوست تخصیص قرار ہے سردازیا ہیچھا، Primary key کی جئے *

و یوں دیار دے جائیں

• Gain Ratio for Attribute Selection (C4.5)

کلیه از مسئله‌ای که در انتخاب attribute پر این اسناده از gain وجود دارد

این اسناد که اگر ویرایش انتخاب شده دارای مقادیر متفاوت زیادی باشد مقدار $Info(D)$

فرمول به صورت می‌شود که باعث افزایش (A) gain و قواعد نیز در صورتی که انتخاب

این ویرایش باعث ایجاد بیشترین تفاوت می‌شود. برای حل این مشکل از gain ratio

استفاده می‌شود که باعث هماهنگی داده‌ها می‌گردد.

$$Gain\ Ratio(A) = \frac{Gain(A)}{Split\ Info(A)}$$

$$Split\ Info_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|}$$

* ویرایشی که دارای بیشترین gainratio و پسندیده عوامل ویرایش انتخاب می‌شود.

* آن ویرایش انتخاب شده یک ویرایش بیوهمه پاسد در بوده با آن یعنی از راه‌لر تقسیم‌بندی

کردن آن به بازه‌ها مختلف و سپس استفاده از بازه به عنوان مقدار آن ویرایش است

که در حلقه انتخاب صریح بازه برای بدآوردن بیشترین صریح‌الرتبه یکی از روش‌ها

این است که مقدارینه هر دو مقادیر سوالی را ($a_i + a_{i+1}/2$) به عنوان ضریب امکان

آزاد و سپس تعداد داده ها را مقایسه نماییم. این روش را تا پیدا کردن بین مقادیر

جدولی اداسه صی درسته.

* آن ویرگی اختیاری سده دارای مقادیر لحسته پاسد اما طبقاً پرسی ایجاد

لسم صی باست که تبر صحنه ای لز مقادیر را در مرد، قرار دفعیم مثلاً در تعیین پندی با

رنگ رنگ، قصر و آب را یک مجموعه در نظر می نمیم و داریم.

color ∈ {red, green}

Yes / No

g Gini Index for Attribute Selection (CART)

این معنای ریاضی پر که هشتم نصی است که با استفاده از فرمول زیر بدست صی آیده

$$gini(D) = 1 - \sum_{j=1}^n p_j^2 \quad \rightarrow \text{number of classes}$$

$$gini_A(D) = \frac{|D_1|}{D} gini(D_1) + \frac{|D_2|}{D} gini(D_2)$$

* ویرگی ای که لحسته (D) را داشته باشد به عنوان ویرگی خدا کشیده انتظاری می شود.
P4PCO

است بعنی در این روش Classification and Regression Tree مخفف CART

که دو درخت Classification (برای کلاسیفیکیشن) و Regression (برای پیش‌بینی شده که دارد)

با آن تعلق دارد) و Regression (برای پیش‌بینی شده که صادر

صدید است) دریم که با اینکه سایرها دارد اما صلاحت روی انطباق جای

که باید حد اشود صفاور دعست.

مسئلۀ الوریتم ID3:

- الوریتم رویکرد صریح‌تر دارد و راه حل یعنی را تخمین نمی‌کند.

راه حل \leftarrow اسقاده از backtracking

- الوریتم صیغه‌ای برداهه های آصره overfit نمود.

- راه حل \leftarrow اسقاده از درخت تصمیم کوچکتر و محدود کردن عمق درخت

- الوریتم برای اسقاده در داده های پیوسته مشکل است.

- راه حل \leftarrow اسقاده از بینرین صداره جدا کننده

- محاطه IG در ویرگن حاوی که مقدار صفاور زیادی دارد پایاس می‌شود.

- راه حل \leftarrow اسقاده از gain ratio

مسئل الگوریتم C4.5 تابیل به ساخت روش های مصنوعی که یک پرستیز آن ایس ار

پترنتر لر دلیری صیبا شد است.

مسئل الگوریتم CART بایسشن سیت به وثیکیتی که دارای مقادیر متفاوت ریاضی

همست و اینجذب مسئل در توانی که تعداد کلاس ها یا لا است صیبا شد.

(Tree Pruning) عرس درخت تصمیم

برای جلوگیری overfitting از هزار به عرس درخت خاریم که یکی ایک روشن

وجود دارد.

در این روش در عمل مساحت درخت با عرس برخی مساحت ها

ماضی ادامه پیدا کردن درخت در آن صیسویم. در اینجا

لک threshold + طبق عمق درخت یا صیزان دهن نعیم صیشوود

که در صورت رسیدن به آن روند متوقف می شود.

در این روش درخت ناصل مساحت سده و سینی به اینداده Past Pruning.

از لک جمیع داده test اقدام به عرس درخت صیسویم.

% Classification in Large Databases

در راست تخصیص در بحث در حجم زیاد داده سُست که برای حل این efficient

مثال دو راهلار ارائه می شود:

برای AVC-list در این روش یک RainForest.

و سپس باعث می یابیم برای هر دو فرآیند AVC-set آنرا می سازیم.

AVC-set تبدیل ساده تعداد رکوردهای است که در ای ای لیبلها میباشد.

همشون و چون خود رکوردهای تبلید اریثی شوند در کار برای را راههای

برگ مثالی لجیا دهنی شود.

برای اینکه از کشش آثاری bootstrapping عمل می کند BOAT

این کشش ویرانی های یک دیگر را در دیگر دیگر دیگر دیگر

تکراری نموده و سپس ب انجام عملیات پروسا آن دیگر در برابر دیگر دیگر

کلی تخصیص هایی صرف نمی کند، در این روش ابتدا درخت های کمیم کوتاه

ساخته و تعیین با مرتب آنها به یک درخت سیار سازی ب درخت اصلی

می رسانیم.

و ترتیب می شود BOAT

• فقط ب دو scan می بازد.

• از RainForest سریعتر است.

• در کار با داده های که صریحاً به هر چیزی متناسب نیست.

Bayesian Classification

روشی صیغه‌بر تئوری بیز است که اصل عضویت در یک کلاس پیش‌بینی می‌کند.

دو روش طبقه‌بندی بر اساس آن تئوری (راهنمای شود):

Naïve Bayesian Classifiers •

با فرض مستقل چندین داده‌ها

عمل می‌کند.

Bayesian Belief Network •

با اینکه صد٪ را با این ارتباط

ین و مرتبه را سر داده‌اند

(تئوری بیز) Baye's Theorem

اگر X data tuple باشد (در این تئوری یک evidence) در تصریح کردند

می‌شود) و H یک فرضیه پرای صیران تعلق X به کلاس صحسن X باشد

آن‌ها بار احتمال H به شرط X داریم:

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)}$$

فرقه می کسیم در مجموع n کلاس که با C_1, C_2, \dots, C_m ملخصه شده است

داریم \bullet پیش یستی مشهود رکورد X به کلاس C_i تعلق دارد آنرا فقط اگر:

$$P(C_i | X) > P(C_j | X) \text{ for } 1 \leq j \leq m, j \neq i$$

پایان پیدا کردن پیشترین احتمال پایه (برای) یعنی $P(C_i | X)$.

$$P(C_i | X) = \frac{P(X | C_i) P(C_i)}{P(X)}$$

جزئی (برای) همه کلاس ها تابع اس سه پارامتر صورت را یعنی $P(X | C_i)$.

$$P(C_i) = \frac{|C_i|}{|D|}$$

$$P(X | C_i) = \prod_{k=1}^n P(x_k | C_i) = P(x_1 | C_i) \times \dots \times P(x_n | C_i)$$

Native Bayesian Classifier

در این روش با فرض مستقل بودن کلاس ها از میدان X و با استفاده از تابع پیش

احتمال را حساب کرده و کلاس که X پی از آن احتمال پیشتر داشته باشد

ب عنوان label پیش یستی سده استخاب شود.

مسئلی کہ درستگاہ اور وجود درد این است کہ Naïve Bayesian Classifier

در دنیا واقعی و تئوری دعا از یکدیگر متفق نیستند. پرای حل این مشکل از BBN استاده صبحی معلم.

متلک دیلری که در اسقاده از لین الگوریتم و یواد هارد این است که همان اس-

اصیال، هدایت و تری بدل عدم وجود حسن (ورود) صفر سود و ارزیابی از

اھيالان صقر سود حاصل اھيال لى سُت صقر صى سود كە چىخ اھناب للاس علماء

صیل سورہ براۓ جلوگاری اے حداد ابن حشیل مک صغار کم را بے دینا سست

حالات کے وجود خدا رہ اپنا تمھاری سیمیوں

وہی مصروفیت NBC

سیاہ دساری آنہاں

۷- نتیجه حوب را پل عبود در اثر موارد

: NBC روشن معاشر

فرنگی کردن صسفک بودن داده ها

Rule-Based Classification

در این روش طبقه بینی بر اساس تعدادی قانون If - Then (الجیام هی پذیرد) :

نمایه کار کل این مجموعه به صورت زیر می باشد :

IF condition THEN conclusion

IF age = youth AND student = yes THEN buy = yes

* The IF Part is known as the rule antecedent

→ consists on one or more attribute

* The THEN Part is known as rule consequent

→ contains a class prediction

* آن شرط پس از IF برای شود صلویم شرط شده و satisfied

قانون رکورد را cover می کند

* آن قانون R توسط triggered شود صلویم آن قانون شده و آن قاعده

لیکن قانون توسط X شده باشد صلویم آن قانون fired شده است

مسئلات

• صحیح قانونی توسط X، Satisfied نشود.

راه حل سے تعریف یہ قانون پیشمرض (Default Rule) کہ درصورت رخ دادن

چیز حالتی جاگیر ہے قانون نہیں با مدعیت:

condition ≠ φ

و پس از یہ حاصل triggered شود.

در اسٹا پس از بروزی ہے

قوامیں ایجاد میں شود.

Conflict Resolution ← راه حل

Conflict Resolution

برای حل مسئل Conflict دو روکردن وجود دارد.

• براساس تعداد field حی Size ordering approach.

موہود در conditions (صیغہ سنتی شرطی) conditions در این حالت ایں

قوامیں triggered شدہ اہم قانونی کے سنتراہی لئے پا شد fire میں شود.

• خود بستی پر دو حالات است:

• براساس کلاس Class-based ordering

دادہ حا را دانسته پا شد fire میں شود.

rule-based ordering
• براساس اولویت قانون کر

این اولویتیں صیغہ براساس عبارتی مختلف و یا حصی یہ مخصوص

PAPCO تھیں شود قانونی کم fire شود را اختیار کریں

آخر میں سسٹم ہائی طبقہ پر براہ راست Rule-based ہے *

عمل چوں۔

Rule Assessment

صریح اور مغول ہائی accuracy & coverage

زیر ہی دست میں کوئی

$$\text{Coverage}(R) = \frac{n_{\text{covers}}}{101} \rightarrow \begin{array}{l} \text{number of tuples} \\ \text{covered by } R \end{array}$$

$$\text{accuracy}(R) = \frac{n_{\text{correct}}}{101} \rightarrow \begin{array}{l} \text{number of tuples correctly} \\ \text{classified by } R \end{array}$$

→ training data set

Rule Extraction from a Decision Tree

قوانین استخراج اور درست تصمیم قابل قبول تر ہیں۔ ہر قانون یک قانون دارد کہ

در آن ہر مقدار بیش پیش نہ شدہ براہ کلا من رائیگاری میں لکھ دیتی ہا عبارت از:

• Mutually exclusive

• No rule conflict

• Exhaustive

• One rule for each attribute-value combination

• The set of rules does not require a default rule

Rule Induction & Sequential Covering Method

این (لوریتم) به استخراج مسئلم داده های آموزشی (train data) مجهز است.

در این روش توانیم به درست ترنسی lear^n هم سویند و ابتدا مراحل لاین الترنس

مجموعه ای از سواین استخراج‌های سوتند. گام‌ها عبارتند از:

- Rules are learned one at a time
 - Each time a rule is learned, the tuples covered by the rules are removed
 - Repeat the process on the remaining tuples until termination condition

در این روش بی ارای صادر مصنوعی میکنند و نتیجه آن ساخته ای باشد که

نعداد داده صای پیشتری را در هر جایی دهد به عنوان ساخته اهلی صرطه استحباب

سُو و صَدِيدَ عَمَلَيَّانَ رَأَيْرَوَى آنْ تَعَافَهَ اَدَامَهَ صَيْرَهِيمَهُ مَسِينَ بَىْ رَسِيدِيَّنَ بَىْ سَرَطَ

پایان صریح backtracks کرده و عملیات را برای مقدار دلخواه کلاس کلراص سینم.

پس از استمرار جنگ و میان رساندن این قوانین با یکدیگر conflict داشته باشد که

برای رفع آن از موضع (conflict Resolution) استفاده می‌کنیم.

: Lazy Learner

روش‌های طبقه‌بندی که کالرمن یاد می‌کنیم بیک صرطه ساخت صد و بیک صرطه استفاده از

صد داشتند. اما روش‌ها که lazy صد های هستند که در مار model construction

خطیان جاوه انتخاب می‌دهند بله علتی این که داده جدید پرای تعیین کلasse به آنها داده

صرفه شود آنرا با داده‌های پیشنهادی تردد و پر اساس آن تشخیص را درآمده می‌داند.

سرویلر د راجع در lazy learner در عبارت از

K - Nearest Neighbor (KNN)

locally weighted regression

case-based reasoning

8. K-Nearest Neighbor

در این رویی همه داده‌ها تبدیل به بیک نظر در مقایسه نیزی می‌شوند. سپس

با توجه به صفت‌هایی که کshortest عامله اقلیدسی را با داده‌ها دارند کلasse

مورد تصریح پرای آن تعیین می‌شوند.

Target function in KNN could be:

- Discrete classification \rightarrow Yes/No - True - False
- Real valued Prediction

لکی از روش‌های دلگیری که صیغه بر KNN ارائه می‌شود
Distance-weighted

صیغه تولید حیسابی‌های نزدیک‌ترین ورنه بیشتری دارند.

$$w = \frac{1}{d(x_0, x_i)^2}$$

که ورن

* لکی از خوبی‌های این روش در برای noise data چاکارهای صادر می‌کنند.

* لکی دلگیر از خوبی‌های این روش محدود است در برای missing data اس.

ارجاست وقتی با بیشترین مقادیر در صورده داده‌ها طبلرین صلیلد.

Model Evaluation and Selection

مدلهای موجود برای ارزیابی دقیقی عبارتند از:

Holdout method •

Cross validation •

Bootstrap •

Confusion Matrix

پاک رسی است که به کل آن میتوانیم صریح از محاسبه کنیم.

| Actual class \ Predicted class | C_1 | $\neg C_1$ |
|--------------------------------|----------------------|----------------------|
| C_1 | True Positives (TP) | False Negatives (FN) |
| $\neg C_1$ | False Positives (FP) | True Negatives (TN) |

در این روش پس از طبقه بندی ۴ حالت عوّق بوجود می آید که با استفاده از آن

محاسبه این موارد را محاسبه کنیم

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Error Rate} = 1 - \text{Accuracy} = \frac{FP + FN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP} \rightarrow \begin{array}{l} \text{نمایشی در مدل} \\ \text{داده های مطابق} \end{array} \rightarrow \begin{array}{l} \text{نمایشی در مدل} \\ \text{داده های مطابق} \end{array}$$

$$\text{Recall} = \frac{TP}{TP + FN} \rightarrow \begin{array}{l} \text{نمایشی در مدل} \\ \text{داده های مطابق} \end{array} \rightarrow \begin{array}{l} \text{نمایشی در مدل} \\ \text{داده های مطابق} \end{array}$$

Perfect score is 1.0

* مدل تو انسنچندر مدل خوب را پوشش دهد.

• F measure (F1 or F-score)

رُهاني که صفحه اهیم با در تصریح ترفندها و recall هم

را استخراج کنیم از میان هارویک آنها استفاده صورتیم

$$F_1 = \frac{2 \times \text{Precision} \times \text{recall}}{\text{Precision} + \text{recall}}$$

$$F_\beta = \frac{(1 + \beta^2) \times \text{Precision} \times \text{recall}}{\text{Precision} + \text{recall}}$$

• Holdout method

در این روش یکتی از داده ها (مجموعاً $\frac{1}{3}$) برای test نهاده شده و مدل را با

داده های جیغ عیناً train کنیم. سپس داده های test را به دیگر داده

و عملیات فوق را K بار تکرار کنیم. صنعتی accuracy های پرسیده برای

مدل صنعتی accuracy

: Cross - Validation

لین روش ایجاده استفاده از داده ها، برای train پی طی دهد. در این روش

داده ها را به K قسمت تقسیم کرده و هر یکی از آنها را کلاسیفیکرده و از آن به عنوان

استفاده کرده و مدل خود را با K-1 قسمت باقیمانده train و test

: Bootstrapping

در این روش با Sampling همراه با جایگزینی داده های دیگر سطح داده ها را

با نمونه هایی با جایگزینی انجام می کرد. در میان حیزی حدود ۹۳٪

داده ها در صیغه داده های train و test باقیمانده داده های

عنصر accuracy وجود دارد که عبارت از

Interpretability •

efficiency in disk-resident databases → Scalability •

missing values, noise data → Robustness •

Speed •

حدت زمان ساخت مدل

مدت زمان استفاده از مدل

«خوشه پندی»

: Cluster Analysis

کلاسستر یک مجموعه از داده ها هست که میتوان ساخته را به یکدیگر و مترین

ساخته را به کلاسستر های دیگر درآورد است.

هدف از خوشه پندی تحلیل ساخته های داده هایی است که در یک مجموعه هم از هم متفاوتند.

دو کاربرد اصلی خوشه پندی عبارتند از:

• اسناده بیانگران یک اپاره Stand-alone برای درک ترتیب داده ط

• اسناده یعنوان نام پیش بودارن برای یافتن الگوریتم ها

خوشه پندی به تمام های دلیری سه صد ازده میتواند عبارت از:

Automatic classification

Data segmentation

Learning by observation

چه چیزهای خوشه پندی میتوانند?

supervised classification

simple grouping

result of query

Clustering as a Pre Processing Tools

- Summarization / Sampling → for regression, classification, etc
- Compression → for Image Processing
- Finding K-Nearset Neighbor
- Outlier detection

Quality of Clustering

برای مستحبه اینکه یک عمل حوت پندی چقدر خوب بوده از دو عضدی، اسقاده چاکیست:

- high intra-cluster similarity
- low inter-cluster similarity

الجیع محیط رحمای دلیر کی نیز وجود دارد که مسئله از مسائله بیرونی حوت پندی صی بردازد:

این پارسیس بندی یعنی صورت Single level Partitioning criteria است

بصورت هرگذرا

این کلاسترها با یکدیگر میتوانند جدا شوند ← Separation of clusters.

که این از این میتواند مبنی بر فاصله است یا مبنی بر جگالی

که این از این میتواند مبنی بر فاصله است یا این از این میتواند مبنی بر جگالی

برای کدام بخش از داده ها؟

چالش‌های موجود در خوشه‌بندی:

Scalability:

Ability to deal with different types of attributes.

محدودیت‌های ایجاد شده برای خوشه‌بندی \leftarrow constraint-based clustering.

Interpretability and usability:

Major clustering approaches:

روش‌های خسته کننده به شکل مصنوعی عمل نموده و \rightarrow داده‌ها را در پایتون‌ها که مستلزم کم باشم همچوشاگان شاره و قرار می‌دهد.

روش‌های خسته کننده پیاساوه، در حقیقت داده‌ها ایجاد می‌کنند

بر اساس تراکم خوشه‌بندی انجام می‌دهند \rightarrow بر اساس تراکم خوشه‌بندی انجام می‌دهند.

بر اساس یک ساختار، سلسه مرتبی چندین چندین خوشه‌بندی را انجام می‌دهند.

• Model-based approach

• Frequent pattern-based approach

• User-guided or constraint-based

• Link-based

2 Partitioning Clustering

برتیشن پڑی کے dataset ا عمومی ہے K لائسٹر یعنی کے مجموع صدیوں کا (E)

کھٹک سوڑ

$$E = \sum_{i=1}^K \sum_{p \in C_i} (p - c_i)^2$$

centroiod or medoid of cluster C_i

لیکن (نحوه) الگوریتم های معروف این قریع حسنه پندی الگوریتم K-means است.

يمكننا تطبيق K-means على مجموعات متساوية الحجم، حيث يتم تعيين كميات متساوية في كل مجموع.

تعداد کلانترها و عدد iteration ها است.

اللوريم إيم سواليه كي ميلز بابا حيابد local optimal معمولاً دريـك K-means

• K-means الگوریتم کلuster

۹. مُسْتَهْدِفَ بِرَأْيِ دادِهِ هَمِيَّةٌ لِـهُ حَرِيكَ فَهَمَّاً، مُوسَمَةٌ لِـهُ بَعْدِيَّهُمْ رَدِيكَ قَابِلَ اجْرَاسَهُ

• نیاں جے تعین ک دار د

و سنتیتی نے outliers و noise data حساس است۔

• فقط حسنه های دایره ای تقلیل بولید صراحت

اللَّوْرِسِيَّ كَمِيرَايِ رُفْعَ مُسْلِلَاتِ K-means K-medoids ارَانَه صُورَه تَامَ دَارَد كَه بِطَيَّ

اسْقَادَه ازْ مَنْطِقَه دَارَه دَعَى يَكِنْ لَلاسْتَنَه عَوَانْ جَاهِدَه لَلاسْتَرَه، دَادَه اَيَّه در صَرَفِ لَلاسْتَرَه

قَدَارَ تَرْفَهَ لَسَتَ رَاهِي عَوَانْ جَاهِدَه لَلاسْتَرَه صَحْرَفِي صَيَّدَه.

الْسَّيَّه اَوْنَ روَيَنْ سُرُّ حَمْ شَارِبَه حَمْسَنْ كَه حَدَرَ وَحَمْ يَرُ روَيِ دَادَه هَاهِي يُورُدَه قَاهِيلَه اَهِيرَاهِيَسَتِ روَيِ

بَعْدَ اَدَدَه دَعْفَهَنْ اَنْ تَنَادَه اَسَهَه (Scalability) (مُدَارَه)

٤. Hierarchical Clustering

اَزْ صَارِيَسَنْ فَاصِله بِعَوَانْ صَحَيَه، حَوْهَمْ بَيْسَيَه اَسْقَادَه صَلَه وَصَرِيَّه آنْ عَدَمْ شَارِبَه حَمْسَنْ

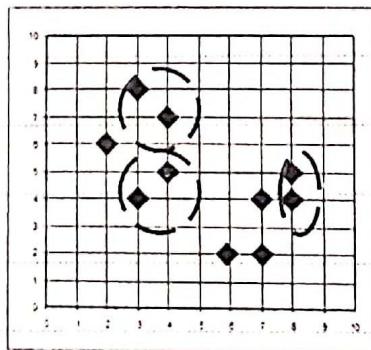
K اَسَهَه بِاَنْ دَوَ روَيَكِرد bottom-up وَ top-down حلَصَلَه.

در روَيَكِرد bottom-up کَه روَيَكِرد agglomeration لَسَتَه هَاهِي لَوَكِلَه توَلِيدَه وَ

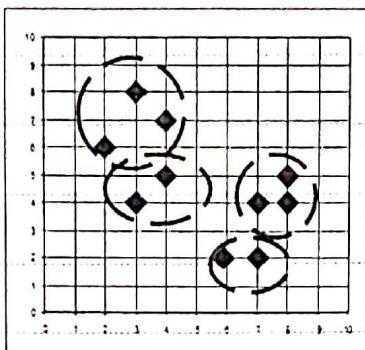
وَسَيَّسَنْ تَيَابَهَسَتَه اَنْهَا حَوَشَهَه بِاَنْ لَيَلِيلَه اَهْفَزَ صَيَّهَه (اسْقَادَه اَزْ صَدَرَه Single-link)

در روَيَكِرد top-down کَه روَيَكِرد divisive database اَسَهَه مَهْدَه در رَيَلَه حَوَشَه

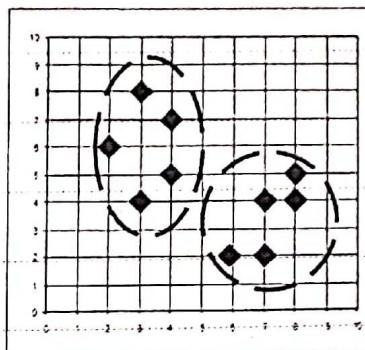
هَهَرَه دَادَه وَسَيَّسَنْ آنْ رَاهِي حَوَشَهَه هَاهِي لَوَكِلَه تَقْسِيمَه صَيَّهَه



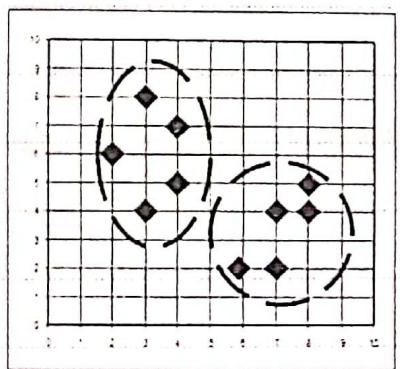
→



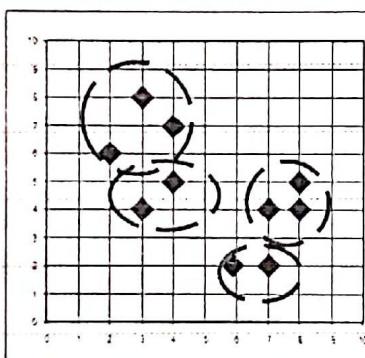
→



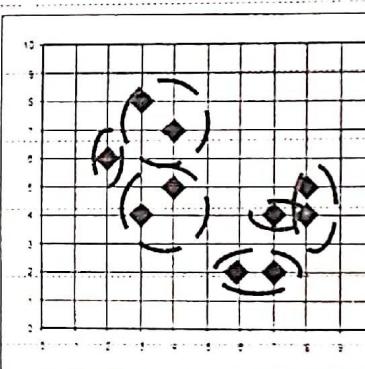
نحوه عملکرد روش bottom-up



→

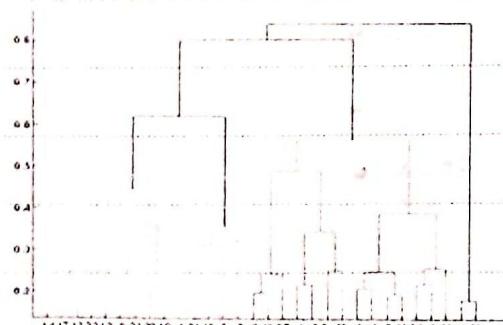


→



نحوه عملکرد روش top-down

مودار dendrogram



این مودار نحوه merge سین کلاس ها را نشان می کند

صورتگر در آن حور عودی سیان دهنده صیزان

dissimilarity پیش فو کلاس را نشان می دهد

همولا پیش فو کلاس را نشان می دهد dissimilarity کلاس های بینی صیان می شود

Distance between clusters :

- Single link $\rightarrow \text{dist}(K_i, K_j) = \min(t_{ip}, t_{iq})$
- Complete link $\rightarrow \text{dist}(K_i, K_j) = \max(t_{ip}, t_{iq})$
- Average $\rightarrow \text{dist}(K_i, K_j) = \text{avg}(t_{ip}, t_{iq})$
- Centroid $\rightarrow \text{dist}(K_i, K_j) = \text{dist}(c_i, c_j)$
- Medoid $\rightarrow \text{dist}(K_i, K_j) = \text{dist}(m_i, m_j)$

Density-based clustering

در این روش خوشه‌بندی بر اساس چنگ داده‌ها انجام می‌شود. ویرایش کاری اصلی آن

عیا، سد از

Discover clusters of arbitrary shape.

Handle noise.

One scan.

متکلم اصلی این روش تزار بی تعیین دقیق با اصرهای آن پرای پیش دادن به الگوریتم است.

الگوریتم های این سه پرای این روش عیا، سد از

DBSCAN.

OPTICS.

DENCLUE.

CLIQUE.

پر اصرارهای موجود در روش Density-based

نیز نیز نشانه صفت رسمی حساسیتی هر node است.

حداقل تعداد نقاط موجود در حساسیتی هر node برای اینکه MinPts

نقاط تعلقی که خوب شدیدهند.

* اگر تعداد داده های موجود در نقاط حساسیتی کمتر از MinPts باشند آنها Outlier هستند.

الگوریتم های Density-based میباشند، سیستم پیشنهادی واردی حساسیت

Grid-based Clustering

لک سه قطعه، Grid resolution یا multi-resolution داریم و در این روش فضای

مسکن بی اساس یک سری صفحات متشکل بی خوب شدن مختلف تقسیم بندی صورت میگیرد.

در لایه اول هر خوب شده داریم که در لایه بعدی هر کدام به خوب شدن دلیل تقسیم صورت میگیرد

لایه رویند باید از این ادامه صورت گیرد.

تفصیل این روش پیشنهادی در آن است که در آن روش صیغه ما

بود اما صیغه در این روش پر اصرارهای آماری است.

الگوریتم های ارائه شده برای این روش عبارتند از:

STING •

Wave Cluster •

CLIQUE •

هزینه روشنایی Grid-based این است که این اطلاعات دارد که به شکل صارتی

انجام شود به عین دلیل مخصوص پرایم اینجا می‌شود.

مشکل این روشنایی است که صریح‌تر هر سلول فقط به صورت (فعی و عمومی) می‌تواند وضمناً باشد
بصورت صوراً بآسانی.

بعضی از این روشنایی‌ها می‌توانند پذیری:

| Method | General Characteristics |
|-----------------------|--|
| Partitioning methods | <ul style="list-style-type: none">Find mutually exclusive clusters of spherical shapeDistance-basedMay use mean or medoid (etc.) to represent cluster centerEffective for small- to medium-size data sets |
| Hierarchical methods | <ul style="list-style-type: none">Clustering is a hierarchical decomposition (i.e., multiple levels)Cannot correct erroneous merges or splitsMay incorporate other techniques like microclustering or consider object "linkages" |
| Density-based methods | <ul style="list-style-type: none">Can find arbitrarily shaped clustersClusters are dense regions of objects in space that are separated by low-density regionsCluster density: Each point must have a minimum number of points within its "neighborhood"May filter out outliers |
| Grid-based methods | <ul style="list-style-type: none">Use a multiresolution grid data structureFast processing time (typically independent of the number of data objects, yet dependent on grid size) |