7/8/19 Concept Descriptions

Data mining : ⊢ Descriptive DM
              ↳ Predictive DM

DDM - Summary of original dataset
(Clear & concise view is obtained)

PDM - Trying to predict the behavior of your data (≈ ML)

Concept - Collection of dataset

⇒ Since, datasets are huge and it is difficult to derive any information from them, summarization of data is important

Concept description
↳ Characterization
↳ Comparison (Discrimination)

representing low level concepts into higher level concepts

Generalization of data using 'age' attribute is much easier than using 'DOB' attribute.
Attributes such as cell phone no. can be completely discarded as it serves no visible useful purpose.

⇒ OLAP deals with numeric data only (generally), whereas concept description (CD) deals with complex data types including text & image.

Since, OLAP works with numeric data, aggregation functions such as sum(), avg(), count() etc are used.

OLAP is user-defined (controlled) (based on domain) where the dimensions & measures of a schema are given by the user.
CD is automated & it discovers the useful attributes by itself.
Functions such as roll-up, drill down are specified by the user in OLAP.

→ based on some threshold (on degree of generalization)

⇒ Attribute such as gender can never be generalized and attributes such as majors (CS, IT, EC etc) can be generalized to various degrees such as engineering / science based on the

- Entropy

threshold and requirement

Data generalization and summarization based characterization

Consider a dataset of sales

Sales dataset
↓
Item information
(item_id, name, brand, category, supplier, place, price)

If these lower level details can be generalized into some higher level detail like seasonal sale (say, in Christmas), then attributes such as category, supplier, place & price must be used and analyzed.

AOI (Attribute oriented induction) Approach proposed for data ~~~~ generalization & summarization in 1989 - based on characterization

AOI	DW
• Online (during runtime)	• Offline
• Before query processing (Time. slow)	• After query processing (Time. fast) precomputed data is stored (data cube)

Steps for AOI

i) Collect task relevant data (TRD)
ii) Perform generalization on TRD
iii) Apply aggregation
iv) Interactive presentation to users

if some attribute can be derived from some other attribute, it is removed as there is no point in storing both of them.

attribute removal (irrelevant attributes) which aren't related to TRD

8/8/19 The first step is also called data focusing.

On performing generalization, over generalizing and under-generalizing can pose serious problems.

<city, state, country>
On dropping state & country attributes leads to under generalization

DOB, Age
This leads to over-generalization

A large difference between training data and predicted data shows high variance, whereas if it matches to a great extent, it may lead to high biasing.

Generalization → Attribute Removal
→ Attribute generalization

Attribute can be removed if it has a large set of distinct values (for attribute there is no concept hierarchy defined). The ~~of~~ uniqueness leads to no interesting patterns and nothing much can be predicted. Ex- Names (can differ even in spellings).

Attribute generalization depends on the level of generalization required and the information which needs to be derived.
< city, state, country >
5 entries of various cities of the same state can straight away be removed/reduced by dropping 'city' attribute & keeping a count of these entries (here, 5).

How large should be the no. of distinct values?
Attribute generalization control
↑ generalization - over generalization
↓ " - under "
↓
May lead to misclassification
- preferring specialization
- Interestingness lost.

Attribute generalization threshold control
- User decides on the threshold on any attribute (user controlled)
Ex- If a column has (2-8) distinct values then generalization may not be required.

The threshold can be same for all attributes or different for every attribute

If the values in a column exceeds the threshold, attribute removal and generalization should be performed

Under-gen → Increase threshold
Over-gen → Decrease threshold
Drill down ← Roll-up

Generalized relation threshold control
No. of distinct values are counted per tuple (distinct tuples)
After generalization, how many rows should be present, is answered here
In older times, 10 to 30 distinct tuples was set as a threshold

Ex: Initial raw relevant data
=

Name	gender	major	birth-place	residence	Phone no.	GPA	(attributes)	birth-date

name	Gender	Major	Birthplace	Birth-date	residence phone #	GPA
jim	M	CS	Vancouver (Canada)	9-18-1997	331, XYZ, PQR	8.4

On application of ADT:

- 'name' is removed as it does not yield any pattern.
- 'gender' has only two value and thus cannot be generalized further.
- 'major' can be generalized under arts, science, commerce (as per requirement) Done by climbing the concept hierarchy.
- 'Birth-place' can be generalized into state or country based on req. & the threshold value isn't satisfied.
- 'Birth-date' can be converted into 'Age' and then stratified into senior, middle-aged, young etc.
- Street no. & street generate more specialization & thus, can be discarded.
- Phone no. can be discarded as it does not generate any interestingness.
- 'GPA' needs to be normalized based on the place & the grading system it uses (out of 4/8/10) and then stratified into { poor, average, good, extraordinary }

After applying all these steps, a 'count' attribute is maintained with every tuple.

9/8/19

After applying ADI:

Gender	Major	Birth-country	Age-range	Residence ...
			20..30 40..50	

GPA	count
good	
excellent	

Algorithm

1. W ← get_taskRelevantData (DMQuery, DB)
2. Prepare_for_generalization (W)
   a. Scan W → get count of distinct value
      a_i – distinct ~~attribute~~ value
   b. for every a_i:
      decide whether to generalize or not based on threshold
3. P ← generalization (W)

→ constructing ~~DB~~ Datacube on the fly for given DM Query

V/S

Predefined Data cube

ON THE FLY

- Collect TRD
  If it is very small and specialized, it is better to go for precomputed datacube.
- Roll up & drill down operations are more (because we don't drop the original

data everytime
- More response time.

Predefined
- Granularity (data cube) matches query.
- Suitable when TRD is huge / large.
- The computational overhead should not be overlooked

Presentation of derived generalization.
Deals with how the data should be displayed after the steps of generalization have been applied.
They can be represented as:
- Cross tab
- Pie chart
- Bar chart
- 3D-cube

Ex:

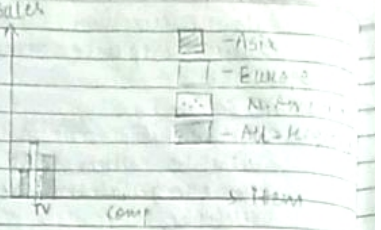
location	item	sales	count (in K)
Asia	TV	15	300
Europe	TV	12	250
N. America	TV	28	450
Asia	Comp.	120	1000
Europe	Comp.	150	1200
N. America	Comp.	200	1800

optional

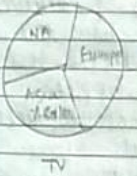
Cross Tab:

location \ item	TV		Comp.		Both_items	
	sales	count	sales	count	sales	count
Asia	15	300	120	1000	135	1300
Europe	12	250	150	1200		
N America	28	450	200	1800		
All_regions	55	1000				

Bar chart:



Pie-chart



19/8/14

t-weight (Typicality weight) (Almost similar to Support)

$$t\text{-weight} = \frac{count(q_a)}{\sum_{i=1}^{n} count(q_i)}$$

n: total no. of tuples

Range: [0-1]

$\forall x$, target-class $(X) \Rightarrow$ condition (X) $[t_1 w_1]$
$\vee \ldots$
$\vee$ condition (X) $[t_n w_n]$

Ex:

$\forall x$, item (X) = "computer" $\Rightarrow$
(location (x) = "Asia") [t: 25%]
$\vee$ (location (x) = "Europe") [t: 30%]

V[location(x) = "North America"] [t: 45%]
1800/4000

After evaluating this, we can set a threshold say 26%. Thus, the tuple with t-weight less than the threshold can be discarded.

Analytical characterization: Attribute Relevance Analysis

The attributes should follow
covariance ≈ 0 or lesser
& variance should be high.

Inherently, attribute relevance analysis tries to achieve this goal
Ex- color of a car does not matter when predicting price of the car.

By applying some statistical measure, the method aims to discard the weakly relevant / related or irrelevant attributes

For predicting salary, the date & month from DOB can be dropped, an approximate age can be calculated from the year and then divided into groups of decades like 10..20, 20..30 etc.

→ Depends on user

Ex: If target class is undergrad and person the contrasting class. (Concern dealing with these two only)

Target class
Contrasting class - tuples which aren't in target class

Methods for Attribute Relevance Analysis (ARA)

1. Information gain & entropy
Expected information

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log\left(\frac{s_i}{s}\right)$$

Entropy:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + s_{3j} + \ldots s_{mj}}{s} \quad \text{for } j$$

Information Gain:

$$\text{Gain}(A) = I(s_1, s_2, \ldots s_m) - E(A)$$

→ s - samples in training data, i.e total tuples.

→ $\{s_1, s_2, \ldots s_m\}$
↓
tuples belonging to target class c_i

→ Probability of arbitrary sample from s belonging to class $i = \frac{s_i}{s}$

Entropy analyzes every attribute available against the target class.

Expected info identifies the no of tuples which can uniquely describe the target class.

$$0 < \frac{s_i}{s} < 1$$

$$\log\left(\frac{s_i}{s}\right) < 0$$

To compensate: $I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} \left(\frac{s_i}{s}\right) \log_2\left(\frac{s_i}{s}\right)$

a) Undergraduate (contrasting class)

gender	major	birth-country	age-range	gpa	count
M	Science	Foreign	<=20	very-good	18
F	Business	Canada	<=20	fair	20
M	Business	Canada	<=20	fair	22
F	Science	Canada	21..25	fair	24
M	Engg.	Foreign	21..25	very-good	22
F	Engg.	Canada	<=20	excellent	24
					130

Graduate (Target class)

Gender	major	birth-country	age-range	gpa	count
M	Science	Canada	21..25	very-good	16
F	Science	foreign	26..30	excellent	22
M	Engg.	foreign	26..30	excellent	18
F	Science	foreign	26..30	excellent	25
M	Science	Canada	21..25	excellent	21
F	Engg.	Canada	21..25	excellent	18
					120

Expected Info:

$$I(S_1, S_2) = -\sum_{i=1}^{2} \frac{S_i}{S} \log\left(\frac{S_i}{S}\right)$$

$$= -\frac{S_1}{S} \log\left(\frac{S_1}{S}\right) - \frac{S_2}{S} \log\left(\frac{S_2}{S}\right)$$

$$= -\frac{120}{250} \log\left(\frac{120}{250}\right) - \frac{130}{250} \log\left(\frac{130}{250}\right)$$

$$= 0.9988$$

22/8/19 To check whether 'major' is a relevant attribute or not.

Here, $v = 3$ (Science, Eng., Business)

Expected info. for Science:

$S_{11} = 84$ (for grad)

$S_{21} = 42$ (for undergrad.)

$$I(S_{11}, S_{21}) = -\frac{84}{126} \log\left(\frac{84}{126}\right) - \frac{42}{126} \log\left(\frac{42}{126}\right)$$

$$= 0.9183$$

For Engg.

$S_{12} = 36$ (grad.)

$S_{22} = 46$ (undergrad.)

$$I(S_{12}, S_{22}) = -\frac{36}{82} \log\left(\frac{36}{82}\right) - \frac{46}{82} \log\left(\frac{46}{82}\right)$$

$$= 0.9892$$

For Business:

$S_{13} = 0$ (grad)

$S_{23} = 42$ (undergrad)

$$I(S_{13}, S_{23}) = -\frac{42}{42}\log(1) = 0$$

$$E(\text{"major"}) = \frac{126}{250} I(S_{11}, S_{21}) + \frac{82}{250} I(S_{12}, S_{22}) + \frac{42}{250} I(S_{13}, S_{23})$$

$$= 0.7873$$

$$\text{Gain(major)} = 0.9988 - 0.7873 = 0.2115$$

If threshold has been set to, let say, 0.1 then, this attribute is considered and not discarded.

Attribute relevant analysis steps

1) Data collection.
   - Determine target and contrasting class

2) Preliminary Relevance Analysis using conservative AOI

3) Remove irrelevant or weakly relevant attributes using selected relevance analysis

4) Generalize concept description using AOI.

Ex for linear regression:

$(x-2005)$

x (Year)	t	Y (Sales)	t*Y	t^2
2005	0	12	0	0
2006	1	19	19	1
2007	2	29	58	4
2008	3	37	111	9
2009	4	45	180	16

$\Sigma t = 10 \quad \Sigma y = 142 \quad \Sigma xy = 368 \quad \Sigma x^2 = 30$

using $y = at + b$

Now, a (slope) $= \dfrac{n\Sigma ty - \Sigma t \Sigma y}{n\Sigma t^2 - (\Sigma t)^2}$

$= \dfrac{5(368) - (10)(142)}{5(30) - (10)^2}$

$= 8.4$

$b = \left(\frac{1}{n}\right)(\Sigma y - a\Sigma t)$

$= 11.6$

$\therefore y = 8.4t + 11.6$ can be used to predict any future value.

In the year 2012, Sales = ?

Here, $t = 2012 - 2005 = 7$

$\therefore y = 8.4(7) + 11.6$

$= 70.4$

PAGE NO.
DATE / /

26/6/19

Mining class comparison

Sometimes the target class is of not much importance to the user rather the attributes and their relation are may be more relevant.

Computer Science class & Physics class are not related but still we comparable how students are performing in their respective domains

Here, the target & contrasting class share some similar dimension or attribute.

How class comparison is performed?

i) Data collection

ii) Dimension relevance analysis
- The dimensions which are highly relevant are kept & others discard (Done using information gain)

iii) Synchronous generalization
- Generalize the target class by using the user specified threshold.

If target class is generalized to country level then, contrasting class should also be generalized to the same level. This is called synchronous generalization

iv) Presentation of derived comparison - cross tables, pie-charts and other representations can be used.

→ d-weight

$$\text{d-weight} = \frac{\text{count}(q_a \in C_j)}{\sum_{i=1}^{m} \text{count}(q_a \in C_i)}$$

m - total target + contrasting classes

d-weight for q_a ⇒ ratio of the no. tuples in target class to total no. of tuples in target as well as contrasting classes

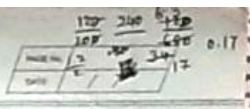
Range: [0% - 100%] OR [0, 1]

If the value of d-weight is ↑ it can be said that the tuple is derived from the target class. If d-weight is ↓, it can be said that the particular tuple is derived from the contrasting class.

$\forall x$, target-class(X) ⇔ condition(X) [d: d-weight]

Status	Major	age range	GPA	count
graduate	Science	21..25	good	90
undergrad	Science	21..25	good	210

$\forall x$, status(X) = "graduate" ⇐ major(X) = "Science" ∧ age-range(X) = "21..25" ∧ gpa(X) = "good" [d: 30%]



$$d = \frac{90}{90+210} = \frac{90}{300} = 0.3 \quad (30\%)$$

Ex:

location/item	TV	computer	both-item
Europe	80	240	320
North America	120	560	680
Both region	200	800	1000

⇒

location/item	TV count	tw	dw	computer count	tw	dw	Both count	tw	dw
Europe	80	25%		240	40%		320		
NA	120	17.65%		560	60%		680		
Both region	200			800			1000		

$\frac{80}{320} \times 100$ $\qquad$ $\frac{80}{200} \times 100$

t-weight:

$\forall x$, target$(x) \Rightarrow$ condition$_1(x)$ $[t: w_1]$ $\vee \ldots$
$\vee$ condition$_m(x)$ $[t: w_m]$

d-weight:

$\forall x$, target$(x) \Leftarrow$ condition$_1(x)$. $[d: w_1] \vee \ldots$
$\vee$ condition$_m(x)$ $[d: w_m]$

Combining:

$\forall x$, target$(x) \Leftrightarrow$ condition$_1(x)$ $[t: w_1, d: w_1']$
$\vee \ldots \vee$
condition$_m(x)$ $[t: w_m, d: w_m']$

1999

⇒ $\forall x$, location(x) = "Europe" $\Leftrightarrow$
(item(x) = "TV") $[t: 25\%, d: 40\%] \vee$
(item(x) = "computer") $[t: 75\%, d = 30\%]$

Probability of the item being TV & coming from Europe is 25%.

&

~~The~~ If the item is TV, the probability that it was sold in Europe is 40%, (And since here we have only 1 discriminating class (North America) 60% of the TV were sold in NA)

Measuring the dispersion of data

Method:	Degree of spread of particular data (how it varies along the axis)
i) Standard deviation	
ii) Inter-quartile range	

- Quartiles, outliers & box-plots

Q_1 = 25 percentile
Q_2 = 50 percentile (median)
Q_3 = 75 percentile
Q_4 = 100 percentile

IQR - Inter quartile range
$= Q_3 - Q_1$
(to observe the behaviour of the data in the middle of the distribution)

⇒ Particularly useful for skewed data

Using IQR, we cannot say anything

about the data residing at extremes thus, 5 point number summary is used.

Steps:

		Order
i)	Observe Q_1	(2)
ii)	" Median	(3)
iii)	" Q_3	(4)
iv)	" minimum	(1)
v)	" maximum	(5)

Box-plot (Representation of 5 number summary)