UNIT-IV-I PART

Attribute Oriented Induction

- Data focusing: task-relevant data, including dimensions, and the result is the *initial relation*
- Attribute-removal: remove attribute *A* if there is a large set of distinct values for *A* but(1) there is no generalization operator on *A*, or (2) *A*'s higher level concepts are expressed in terms of other attributes
- Attribute-generalization: If there is a large set of distinct values for *A*, and there exists a set of generalization operators on *A*, then select an operator and generalize *A*
- Attribute-threshold control: typical 2-8,specified/default
- Generalized relation threshold control: control the final relation/rule size

How it is done

- Collect the task-relevant data (*initial relation*) using a relational database query
- Perform generalization by attribute removal or attribute generalization
- Apply aggregation by merging identical, generalized tuples and accumulating their respective counts
- Interaction with users for knowledge presentation

  Example: Describe general characteristics of graduate students in the University database

Step 1. Fetch relevant set of data using an SQL statement, e.g.,

Select * (i.e., name, gender, major, birth_place, birth_date,residence, phone#,gpa)
From student
where student_status in {"Msc", "MBA", "PhD"}

Step 2. Perform attribute-oriented induction

Step 3. Present results in generalized relation, cross-tab, or rule forms

Basic Algorithm for Attribute-Oriented Induction:

- InitialRel: Query processing of task-relevant data, deriving the *initial relation*.
- PreGen: Based on the analysis of the number of distinct values in each attribute, determine generalization plan for each attribute: removal? or how high to generalize?
- PrimeGen: Based on the PreGen plan, perform generalization to the right level to derive a "prime generalized relation", accumulating the counts.
- Presentation: User interaction: (1) adjust levels by drilling, (2) pivoting, (3) mapping into rules, cross tabs, visualization presentations.

Class Characterization: An Example

Analytical Characterization

Initial Relation

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...	...	...	...	...	...	...	...
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,..

Prime Generalized Relation

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
...	...	...	...	...	...	...

Gender \ Birth_Region	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62

1. Datacollection
   target class: graduate student
   contrasting class: undergraduatestudent
2. Analytical generalization using U_i
   attribute removal
      remove *name* and *phone#*
   attribute generalization
      generalize *major*, *birth_place*, *birth_date* and *gpa*
      accumulate counts
   candidate relation: *gender*, *major*, *birth_country*, *age_range* and *gpa*

Class Comparison Methods & Implementations:

Data Collection:

The set of associated data from the databases and data warehouses is collected by query processing and is partitioned into the target class and contrasting class.

Dimension Relevance Analysis:

When many dimensions are to be processed and is required that analytical comparison should be performed, then dimension relevance analysis should be performed on these classes, and only the highly relevant dimensions are included in the further analysis.

Synchronous Generalization:

The process of generalization is performed upon the target class to the level controlled by the user or expert specified dimension threshold, which results in a prime target class relation/cuboid.

The concepts in the contrasting class or classes are generalized to the same level as those in the prime target class relation/cuboid, forming the prime contrasting class relation/cuboid.

Presentation of the derived comparison:

The resulting class comparison description can be visualized in the form of tables, charts, and rules. This presentation usually includes a “ contrasting” measure (such as count%) that reflects the comparison between the target and contrasting classes.

(No.1 Best Selling Data Science Course On Udemy)

Example

Task - Compare graduate and undergraduate students using the discriminant rule.

for this, the DMQL query would be.

use University_Database
mine comparison as “graduate_students vs_undergraduate_students”
in relevance to name, gender, program, birth_place, birth_date, residence, phone_no, GPA
for “graduate_students”
where status in “graduate”
versus “undergraduate_students”
where status in “undergraduate”
analyze count%
from student

Now from this, we can formulate that

attributes = name, gender, program, birth_place, birth_date, residence, phone_no, and GPA.

Gen(ai) = concept hierarchies on attributes ai.

Ui = attribute analytical thresholds for attributes ai.

Ti = attribute generalization thresholds for attributes ai.

R = attribute relevance threshold.

1. Data collection -Understanding Target and Contrasting classes.

2. Attribute relevance analysis - It is used to remove attributes name, gender, program, phone_no.

3. Synchronous generalization - It is controlled by user-specified dimension thresholds, a prime target, and contrasting class(es) relations/cuboids.

Initial target class working relation (graduate student)

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...	...	...	...	...	...	...	...

Initial contrasting class working relation (graduate student)

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Bob Schumann	M	Chem	Calagary, Alt, Canada	10-1-78	2642 Halifax St, Burnaby	294-4291	2.96
Ammy. Eau	F	Bio	Golden, BC, Canada	30-3-76	463 Sunset Cres, Vancouer	681-5417	3.52
...	...	...	...	...	...	...	...

4. Drill down, roll up and other OLAP operations on target and contrasting classes to adjust levels of abstractions of resulting description.

Prime generalized relation for the target class: Graduate students

Major	Age_range	Gpa	Count%
Science	20-25	Good	5.53%
Science	26-30	Good	2.32%
Science	Over_30	Very_good	5.86%
...	...	...	...
Business	Over_30	Excellent	4.68%

Prime generalized relation for the contrasting class: Undergraduate students

Major	Age_range	Gpa	Count%
Science	15-20	Fair	5.53%
Science	15-20	Good	4.53%
...	...	...	...
Science	26-30	Good	5.02%
...	...	...	...
Business	Over_30	Excellent	0.68%

5. The presentation- Data is presented as generalized relations, cross tabs, bar charts, pie charts, or rules,
contrasting measures to reflect a comparison between target and contrasting classes.
e.g. count%

Class Description: Presentation of both Characterization and Comparison:

location	item	sales (in million dollars)	count (in thousands)
Asia	TV	15	300
Europe	TV	12	250
North_America	TV	28	450
Asia	computer	120	1000
Europe	computer	150	1200
North_America	computer	200	1800

Table 5.3: A generalized relation for the sales in 1997.

Cross Tab

location \ item	TV		computer		*both_items*	
	sales	count	sales	count	sales	count
Asia	15	300	120	1000	135	1300
Europe	12	250	150	1200	162	1450
North_America	28	450	200	1800	228	2250
all_regions	45	1000	470	4000	525	5000

Table 5.4: A crosstab for the sales in 1997.

Quantitative Discriminant Rules

To find out the discriminate features of target and contrasting classes can be described as a discriminate rule.
It associates an interesting measure d-weight with each tuple.
Cj - target class
Qa - a generalized tuple covers some tuples of class, but can also cover some tuples of contrasting class
d-weight - range: [0, 1]
d-weight = count(Qa)/summation(count(Qa))

Example

Status	Birth_country	Age_range	Gpa	Count
Graduate	Canada	25-30	Good	90
Undergraduate	Canada	25-30	Good	210

In the above example, suppose that the count distribution for major ='science' and age_range = '20..25" and GPA ='good' is shown in the tables.

The d_weight would be 90/(90+210) = 30% w.r.t to target class and the d_weight would be 210/(90+210) = 70% w.r.t to contrasting class. i.e.

The student majoring in science is 21 to 25 years old and has a good GPA then based on the data, there is a probability that she is a graduate student versus a 70% probability that she is an undergraduate student. Similarly, the d-weights for other tuples also can be derived.

Mining Class Comparison

Comparison: Comparing two or more classes

- Method:
  - Partition the set of relevant data into the target class and the contrasting class(es)
  - Generalize both classes to the same high level concepts
  - Compare tuples with the same high level descriptions
  - Present for every tuple its description and two measures
    - support - distribution within single class
    - comparison - distribution between classes
  - Highlight the tuples with strong discriminant features
- Relevance Analysis:
  - Find attributes (features) which best distinguish different classes

Presentation of Generalized Results

- Generalized relation:
  - Relations where some or all attributes are generalized, with counts or other aggregation values accumulated.
- Cross tabulation:
  - Mapping results into cross tabulation form (similar to contingency tables).
  - Visualization techniques:
  - Pie charts, bar charts, curves, cubes, and other visual forms.
- Quantitative characteristic rules:
  - Mapping generalized result into characteristic rules with quantitative information associated with it, e.g.,
- t-weight:
  - Interesting measure that describes the typicality of
    - each dis junct in the rule
    - each tuple in the corresponding generalized relation
    - n – number of tuples for target class for generalized relation
    - $q_i \ldots q_n$ – tuples for target class in generalized relation
    - q_a is in $q_i \ldots q_n$

$$t\_weight = count(q_a) / \sum_{i=1}^{n} count(q_i)$$

grad(x) Λ male(x) ⇒ birth_region(x) = "Canadd[t:53%] ∨ birth_region(x) = "foreign[t:47%]

$$\forall x \in attributes(X)[<x,l,u> \in X \wedge <x,l',u'> \in \hat{X} \Rightarrow l' \le l \le u \le u']$$