Classifying Categorical Data

Modified slides by Risi Thonangi

M.S. Thesis Presentation

The Classification Problem

record	*class (or item)*
i_1 , i_2 , i_3	c_1
i_2 , i_3	c_1
i_1 , i_4	c_2
i_1 , i_3	c_1

attributes

$I = \{i_1 , i_2 , i_3 , i_4\}$

classes

$C = \{c_1,c_2\}$

The Classification Problem

record	*class*
i_1 , i_2 , i_3	c_1
i_2 , i_3	c_1
i_1 , i_4	c_2
i_1 , i_3	c_1

attributes

$I = \{i_1 , i_2 , i_3 , i_4\}$

classes

$C = \{c_1,c_2\}$

i_1 , i_2 , i_4	**?**

The Classification Problem

record	*class*
i_1 , i_2 , i_3	c_1
i_2 , i_3	c_1
i_1 , i_4	c_2
i_1 , i_3	c_1

attributes

$I = \{i_1 , i_2 , i_3 , i_4\}$

classes

$C = \{c_1,c_2\}$

i_1 , i_2 , i_4	**?**

query

Formal Problem Statement

- Given a Dataset *D*

$$D = (r_i, c_k), \quad \forall i = 1,2,\ldots,|D|$$

- Learn from this dataset to classify a potentially unseen record `` `q' `` ***[query]*** to its correct class.

- Each **record $\mathbf{r}_i$** is explained using boolean attributes $I = \{ i_1, i_2, \ldots, i_{|I|} \}$ and is labeled to one of the classes $C = \{ c_1, c_2, \ldots, c_{|C|} \}$

- $I = \{ i_1, i_2, \ldots, i_{|I|} \}$ can also be looked at as a set of items.

Preliminaries

- *itemset* A set of items – $\{ i_1 , i_2 , i_3 \}$
- *P(.)* Probability Distribution
- *frq-itemset* An *itemset* whose frequency is above a given threshold σ
- σ Support Threshold
- τ Confidence Threshold
- $\{ i_1 , i_2 \} \rightarrow \{ i_3 \}$ An Association Rule **(AR)**

$$\text{sup}[\, i_1, i_2 \rightarrow i_3 \,] = P(i_1, i_2, i_3) > \sigma$$

$$conf[\, i_1, i_2 \rightarrow i_3 \,] = \frac{P(i_1, i_2, i_3)}{P(i_1, i_2)} > \tau$$

- $\{ i_1 , i_2 \} \rightarrow c_1$ A Classification Association Rule **(CAR)**

Classification based on Associations (CBA)

- **[Bing Liu – KDD98]**
- First Classifier that used the paradigm of Association Rules

- Steps in CBA:
  - **Mine** for CARs satisfying support and confidence thresholds
  - **Sort** all CARs based on confidence
  - **Classify** using the rule that satisfies the query and has the highest confidence

Classification based on Associations (CBA)

- [Bing Liu – KDD98]
- First Classifier that used the paradigm of Association Rules

- Steps in CBA:
  - Mine for CARs satisfying support and confidence thresholds
  - Sort all CARs based on confidence
  - Classify using the rule that satisfies the query and has the highest confidence
    - With rules of the same confidence, select the rule with higher support
    - The same confidence and support, select the rule with less items
- Disadvantages:
  - Single rule based classification – Not Robust

Disadvantages with CBA: Single Rule based classification

- Let the classifier have 3 rules :
  - $i_1 \rightarrow c_1$ support: 0.3, confidence: 0.8
  - $i_2 , i_3 \rightarrow c_2$ support: 0.7, confidence: 0.7
  - $i_2 , i_4 \rightarrow c_2$ support: 0.8, confidence: 0.7

- Query $\{ i_1 , i_2 , i_3 , i_4 \}$ will be classified to the class c_1 by CBA which might be incorrect.

- CBA, being a single-rule classifier, cannot consider the effects of multiple-parameters.

Classification based on Multiple ARs (CMAR)

- [WenminLi-ICDM01]
- Uses multiple CARs in the **classification step**

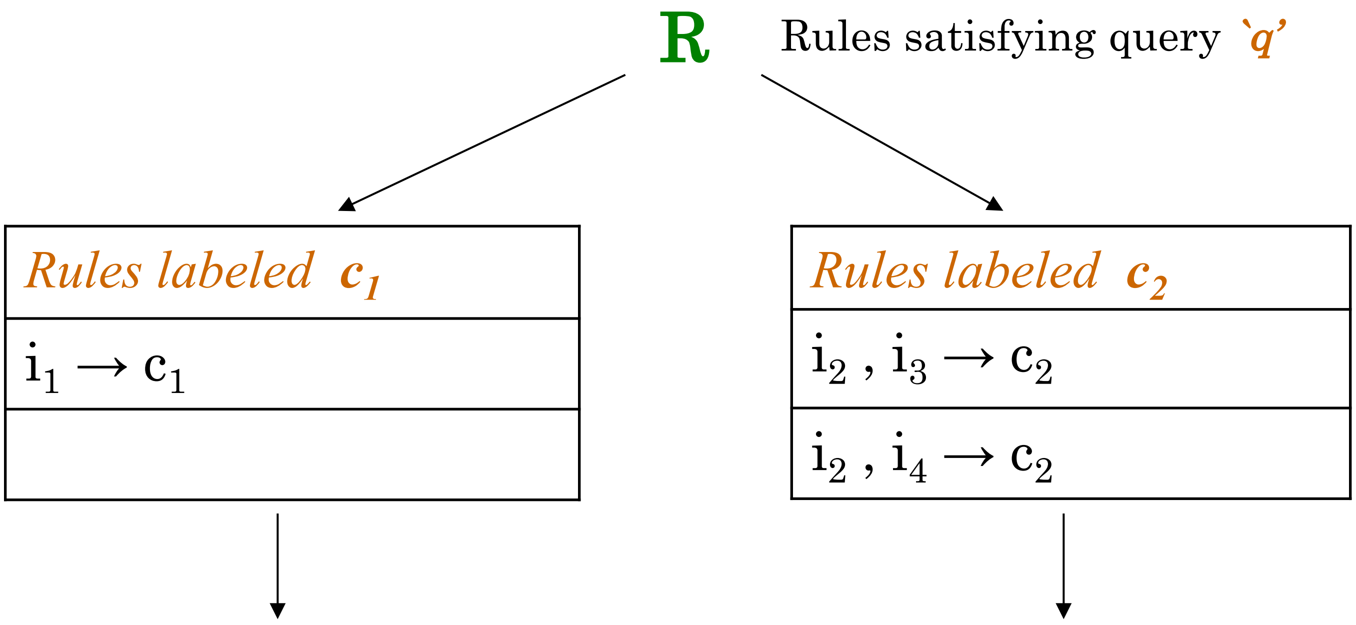
- Steps in CMAR:
  - **Mine** for CARs satisfying support and confidence thresholds
  - **Sort** all CARs based on confidence
  - **Find** all CARs which satisfy the given query
  - **Group** them based on their class label
  - **Classify** the query to the class whose group of CARs has the maximum ***weight***

Classification based on Multiple ARs (CMAR)

- [WenminLi-ICDM01]
- Uses multiple CARs in the **classification step**

- Steps in CMAR:
  - **Mine** for CARs satisfying support and confidence thresholds
  - **Sort** all CARs based on confidence
  - **Find** all CARs which satisfy the given query
  - **Group** them based on their class label
  - **Classify** the query to the class whose group of CARs has the maximum ***weight***

CMAR contd.



R Rules satisfying query *`q'*

Rules labeled ***c_1***

$i_1 \longrightarrow c_1$

Rules labeled ***c_2***

$i_2 , i_3 \longrightarrow c_2$

$i_2 , i_4 \longrightarrow c_2$

Output the class with the highest sum of weighted chi squares of all rules in each class

* https://cgi.csc.liv.ac.uk/~frans/KDD/Software/CMAR/cmar.html

CMAR contd.

- Outperforms C4.5 and CBA on accuracy
- Less storage requirements compared to CBA
- Lower running time compared to CBA
- Accuracy does not depend too much on confidence and coverage threshold