جامعة الطائف
TAIF UNIVERSITY

DATA MINING

Assoc. Prof. Dr. Salha Alzahrani

College of Computers and Information Technology
Taif University
Saudi Arabia

s.zahrani@tu.edu.sa

Data Mining | Topics we have covered so far…

Recap of Previous Lecture

Content of This Lecture

Summary & Checklist

Topics

- **Introduction:** Data mining and knowledge discovery.
- **Chapter (1):** Data Processing |Data Cleaning, Dealing with Missing Data,
- **Chapter (2):** Classification using Naïve Bayes Algorithm
- **Chapter (2):** Classification using Nearest Neighbour Algorithm.
- **Chapter (3):** Classification using Decision Trees.
- **Chapter (10):** Inducing Modular Rules for Classification.
- **Chapter (12):** Association Rule Mining – Part I.
- **Chapter (13):** Association Rule Mining – Part II.

Recap of Previous Lecture
Content of This Lecture
Summary & Checklist

Recap of Lecture 6

- Rule Post-pruning
- Exercise: Rule Post-pruning
- Inducing Modular Rules for Classification: The Prism Algorithm
- The Prism Algorithm: The lens24 Example
- Handout (5): Using the Prism Algorithm for Rules Induction from the lens24 dataset.
- Summary & Checklist.

Prediction | Association Rule Mining- Part: I

Recap of Previous Lecture

Content of This Lecture

Summary & Checklist

Content of Lecture 7

- Introduction to Association Rule Mining
- Measures of Rule Interestingness
- Basic measures of rules interestingness: Confidence, Support and Completeness
- Example
- Measures of Rule Interestingness: discriminability
- The Piatetsky-Shapiro Criteria and the RI Measure
- Rule Interestingness Measures Applied to the chess Dataset
- Association Rule Mining Tasks
- Summary & Checklist.

Prediction | Introduction to Association Rule Mining

- Classification rules are concerned with predicting the value of a ***categorical attribute*** that has been identified as being of particular ***importance***.

- In this chapter we go on to look at the more general problem of ***finding any rules of interest*** that can be derived from a given dataset.

- We will restrict our attention to IF . . . THEN . . . rules that have a conjunction of **'attribute = value'** terms on both their left- and right-hand sides.

- We will also assume that all attributes are ***categorical***.

Prediction | Introduction to Association Rule Mining

IF Has-Mortgage = yes AND Bank Account Status = In credit
THEN Job Status = Employed AND Age Group = Adult under 65

- Rules of this more general kind represent an ***association*** between the values of certain attributes and those of others.
- Called ***association rules***.
- The process of extracting ***association rules*** from a given dataset is called ***association rule mining*** **(ARM)**.
- The term ***generalised rule induction*** **(or GRI)** is also used, by contrast with classification rule induction.

Prediction | Measures of Rule Interestingness

- To distinguish between one rule and another we need some measures of rule quality. These are generally known as ***rule interestingness measures***.

- We will write a rule in the form
  - **If LEFT then RIGHT**

- We start by defining **four** numerical values which can be determined for any rule simply by counting:
  - N_{LEFT} Number of instances matching LEFT
  - N_{RIGHT} Number of instances matching RIGHT
  - N_{BOTH} Number of instances matching both LEFT and RIGHT
  - N_{TOTAL} Total number of instances

Prediction | Measures of Rule Interestingness

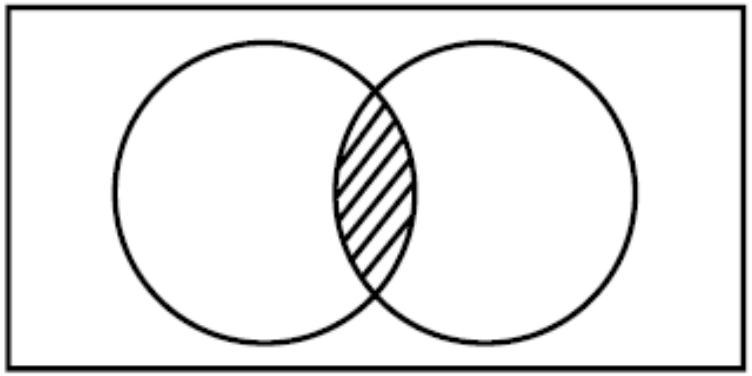

Figure 12.1 Instances matching LEFT, RIGHT and both LEFT and RIGHT

Prediction | Basic measures of rules interestingness: Confidence, Support and Completeness

- Three commonly used measures are given below. The first has more than one name in the technical literature.

Confidence (Predictive Accuracy, Reliability)

N_{BOTH} / N_{LEFT}

The proportion of right-hand sides predicted by the rule that are correctly predicted

Support

N_{BOTH}/N_{TOTAL}

The proportion of the training set correctly predicted by the rule

Completeness

N_{BOTH}/N_{RIGHT}

The proportion of the matching right-hand sides that are correctly predicted by the rule

Figure 12.2 Basic Measures of Rule Interestingness

Prediction | Measures of rules interestingness: Example

IF Has-Mortgage = yes AND Bank Account Status = In credit
THEN Job Status = Employed AND Age Group = Adult under 65

- Assume that by counting we arrive at the following values:
  - $N_{LEFT} = 65$
  - $N_{RIGHT} = 54$
  - $N_{BOTH} = 50$
  - $N_{TOTAL} = 100$

- From these we can calculate the values of the three interestingness measures
  - Confidence $= N_{BOTH}/N_{LEFT} = 50/65 = 0.77$
  - Support $= N_{BOTH}/N_{TOTAL} = 50/100 = 0.5$
  - Completeness $= N_{BOTH}/N_{RIGHT} = 50/54 = 0.93$

- The confidence is **77%**, which is not very high. However, it correctly predicts for **93%** of the instances in the dataset that match the right-hand side of the rule and the correct predictions apply to as much as **50%** of the dataset. This seems like a **valuable rule**.

Prediction | Measures of Rule Interestingness: discriminability

- Amongst the other measures of interestingness that are sometimes used is ***discriminability***. This measures how well a rule discriminates between one class and another.

$$1 - (N_{LEFT} - N_{BOTH}) / (N_{TOTAL} - N_{RIGHT})$$

- which is **1 – (number of misclassifications produced by the rule) / (number of instances with other classifications)**

- If the rule predicts perfectly, i.e. $N_{LEFT} = N_{BOTH}$, the value of discriminability is 1.
- For the example given above, the value of discriminability is
- $1 - (65 - 50) / (100 - 54) = 0.67$.

Prediction| The Piatetsky-Shapiro Criteria and the RI Measure

- The American researcher Gregory Piatetsky-Shapiro proposed three principal criteria that should be met by any rule interestingness measure.

Criterion 1
The measure should be zero if $N_{BOTH} = (N_{LEFT} \times N_{RIGHT})/N_{TOTAL}$
Interestingness should be zero if the antecedent and the consequent are statistically independent (as explained below).

Criterion 2
The measure should increase monotonically with N_{BOTH}

Criterion 3
The measure should decrease monotonically with each of N_{LEFT} and N_{RIGHT}

For criteria 2 and 3, it is assumed that all other parameters are fixed.

Figure 12.3 Piatetsky-Shapiro Criteria for Rule Interestingness Measures

Prediction | The Piatetsky-Shapiro Criteria and the RI Measure

- Piatetsky-Shapiro proposed a further rule interestingness measure called **RI**, that meets his three criteria.
- This is defined by:
  - $RI = N_{BOTH} - (N_{LEFT} \times N_{RIGHT} / N_{TOTAL})$
- RI measures the difference between:
  - actual number of matches, and
  - expected number if the left- and right-hand sides of the rule were independent.
- The value of RI:
  - generally is positive.
  - A value of zero indicate that the rule is a chance.
  - A negative value imply that the rule is less successful than chance.

Prediction | Rule Interestingness Measures Applied to the chess Dataset

Rule	N_{LEFT}	N_{RIGHT}	N_{BOTH}	Conf	Compl	Supp	Discr	RI
1	2	613	2	1.0	0.003	0.003	1.0	0.105
2	3	34	3	1.0	0.088	0.005	1.0	2.842
3	3	34	3	1.0	0.088	0.005	1.0	2.842
4	9	613	9	1.0	0.015	0.014	1.0	0.473
5	9	613	9	1.0	0.015	0.014	1.0	0.473
6	1	34	1	1.0	0.029	0.002	1.0	0.947
7	1	613	1	1.0	0.002	0.002	1.0	0.053
8	1	613	1	1.0	0.002	0.002	1.0	0.053
9	3	34	3	1.0	0.088	0.005	1.0	2.842
10	3	34	3	1.0	0.088	0.005	1.0	2.842
11	9	613	9	1.0	0.015	0.014	1.0	0.473
12	9	613	9	1.0	0.015	0.014	1.0	0.473
13	3	34	3	1.0	0.088	0.005	1.0	2.842
14	3	613	3	1.0	0.005	0.005	1.0	0.158
15	3	613	3	1.0	0.005	0.005	1.0	0.158
16	9	34	9	1.0	0.265	0.014	1.0	8.527
17	9	34	9	1.0	0.265	0.014	1.0	8.527
18	81	613	81	1.0	0.132	0.125	1.0	4.257
19	162	613	162	1.0	0.264	0.25	1.0	8.513
20	324	613	324	1.0	0.529	0.501	1.0	17.026

$N_{TOTAL} = 647$

Figure 12.4 Rule Interestingness Values for Rules Derived from *chess* Dataset

Prediction | Association Rule Mining I

- The decision tree derived from the *chess* dataset comprises 20 rules.
- One of these (numbered rule 19) is

  IF inline = 1 AND wr bears bk = 2 THEN Class = safe

- For this rule
  - $N_{LEFT} = 162$
  - $N_{RIGHT} = 613$
  - $N_{BOTH} = 162$
  - $N_{TOTAL} = 647$
- So we can calculate the values of the various rule interestingness measures as follows:
  - Confidence = $162/162 = 1$
  - Completeness = $162/613 = 0.26$
  - Support = $162/647 = 0.25$
  - Discriminability = $1 - (162 - 162)/(647 - 613) = 1$
  - RI = $162 - (162 \times 613/647) = 8.513$

Prediction | Handout (6): Rule Interestingness Measures

Prediction | Association Rule Mining Tasks

- A common requirement is to find all rules with **confidence** and **support** **above specified threshold values**.

- An important type of association rule application for which this approach is used is known as **market basket analysis** :
  - involves analysing very large datasets collected by supermarkets, telephone companies, banks etc. about their customers' transactions (purchases, calls made, etc.) to find associations.
  - In the supermarket case, find associations between the products purchased by customers.
  - Such datasets are generally handled by:
    - **restricting attributes** to having only the values true or false (indicating the purchase or non-purchase of some product)
    - **restricting rules** to ones where every attribute included in the rule has the value true.

Recap of Previous Lecture
Content of This Lecture
Summary & Checklist

Summary & Checklist

- ☑ Introduction to Association Rule Mining
- ☑ Measures of Rule Interestingness
- ☑ Basic measures of rules interestingness: Confidence, Support and Completeness
- ☑ Example
- ☑ Measures of Rule Interestingness: discriminability
- ☑ The Piatetsky-Shapiro Criteria and the RI Measure
- ☑ Rule Interestingness Measures Applied to the chess Dataset
- ☑ Association Rule Mining Tasks.

Reminder | Next Lecture !

Next Lecture...

Association Rule Mining II (Ch. 13)

- *Be ready!*
- *Download & print the lecture notes before your class.*

Thank You !

✉ s.zahrani@tu.edu.sa