Association Rule Mining
(Class Rare Association Rule Mining)

Cristóbal Romero
Córdoba University, Campus Universitario de Rabanales, 14071, Córdoba, Spain
cromero@uco.es

KDIS Lab
KNOWLEDGE DISCOVERY AND INTELLIGENT SYSTEMS

Definition: Association Rule

- Association Rule
  - An implication expression of the form X → Y, where X and Y are itemsets
  - Example:
    {Milk, Diaper} → {Beer}

- Rule Evaluation Metrics
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X

TID	***Items***
1	Bread, **Milk**
2	Bread, **Diaper**, **Beer**, Eggs
3	**Milk, Diaper, Beer**, Coke
4	Bread, **Milk, Diaper, Beer**
5	Bread, **Milk, Diaper**, Coke

$$\text{Example: } \{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

Mining Association Rules

- Algorithms for mining frequent rules:
  - Apriori.
  - Predictive Apriori.
  - FP-growth.
- Other related task:
  - Rare/infrequent rule mining
  - Class association rule mining
  - Numeric association rule mining
  - Multi-level or generalized rule mining
  - Constrained rule mining
  - Incremental rule mining
  - …

Effect of Support Distribution

- How to set the appropriate *minsup* threshold?
  - If *minsup* is set too high, we could miss itemsets involving interesting rare items (e.g., expensive products)
  - If *minsup* is set too low, it is computationally expensive and the number of itemsets is very large
- Using a single minimum support threshold may not be effective

Pattern Evaluation

- Association rule algorithms tend to produce too many rules
  - many of them are uninteresting or redundant
  - Redundant if $\{A,B,C\} \rightarrow \{D\}$ and $\{A,B\} \rightarrow \{D\}$ have same support & confidence

- Interestingness measures can be used to prune/rank the derived patterns

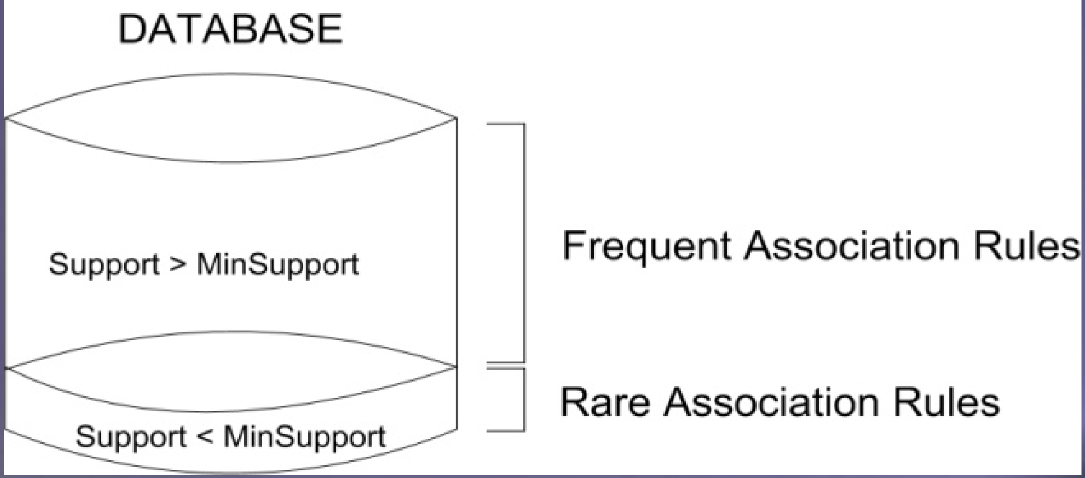
- In the original formulation of association rules, support & confidence are the only measures used

Introduction

- **Association Rule Mining (ARM)** is one of the most popular and well-known data mining methods for discovering interesting relationships between variables in data repositories.
- An **association rule** is an implication X⇒Y, where X and Y are disjoint itemsets. The intuitive meaning of such a rule is that when X appears, Y also tends to appear.
- **Rare Association Rules (RAR)** also known as non-frequent, unusual, exceptional or sporadic rules are those that only appear infrequently even though they are highly associated with very specific data.

RARM in Education

- The problem of discovering rare items has recently captured the interest of the data mining community.
- Rare itemsets are those that only appear together in very few transactions or some very small percentage of transactions in the database.
- They have low support and high confidence in contrast to general association rules which are determined by high support and a high confidence level.



RARM in Education

- **ARM** has been applied extensively in e-learning to discover frequent student-behavior patterns.
- However, **RARM** has been hardly applied to educational data, despite the fact that infrequent associations can be of great interest since they are related to rare but crucial cases. These rules could help the instructor to discover a minority of students who may need specific support with their learning process.
- The greatest reason for applying RARM in the field of EDM is the imbalanced nature of data in education in which some classes have many more instances than others.
- Furthermore, in applications like education, the minor parts of an attribute can be more interesting than the major parts; for example, students who fail or drop out are usually less frequent than those students who fare well.

Background

- Rare-association rules are more difficult to mine using traditional data mining algorithms, since they do not usually consider class-imbalance and tend to be overwhelmed by the major class, leaving the minor class to be ignored.
- There are several different approaches to discover rare association rules:
  - The simplest way is to directly apply the **Apriori [**Agrawal et al. 2003**]** algorithm by simply setting the minimum support threshold to a low value.
  - A different proposal, known as **Apriori-Infrequent**, involves a simple modification of the Apriori algorithm to use of the maximum support measure, instead of the usual minimum support, to generate candidate itemsets, i.e., only items with a lower support than a given threshold are considered.

Background

- A different perspective consists of developing new specific algorithms to tackle these new challenges:
  - A proposal is **Apriori-Inverse** [Koh and Rountree, 2009] that it also uses the maximum support but proposes three different kinds of additions: fixed threshold, adaptive threshold and hill climbing.
  - Another proposal is the **Apriori-Rare** [Szathmary et al. 2007] also known as Arima that is composed of two different algorithms: a naïve one, which relies on Apriori and hence enumerates all frequent itemsets; and MRG-Exp, which limits the considerations to frequent itemsets generators only.

Experimentation and Results

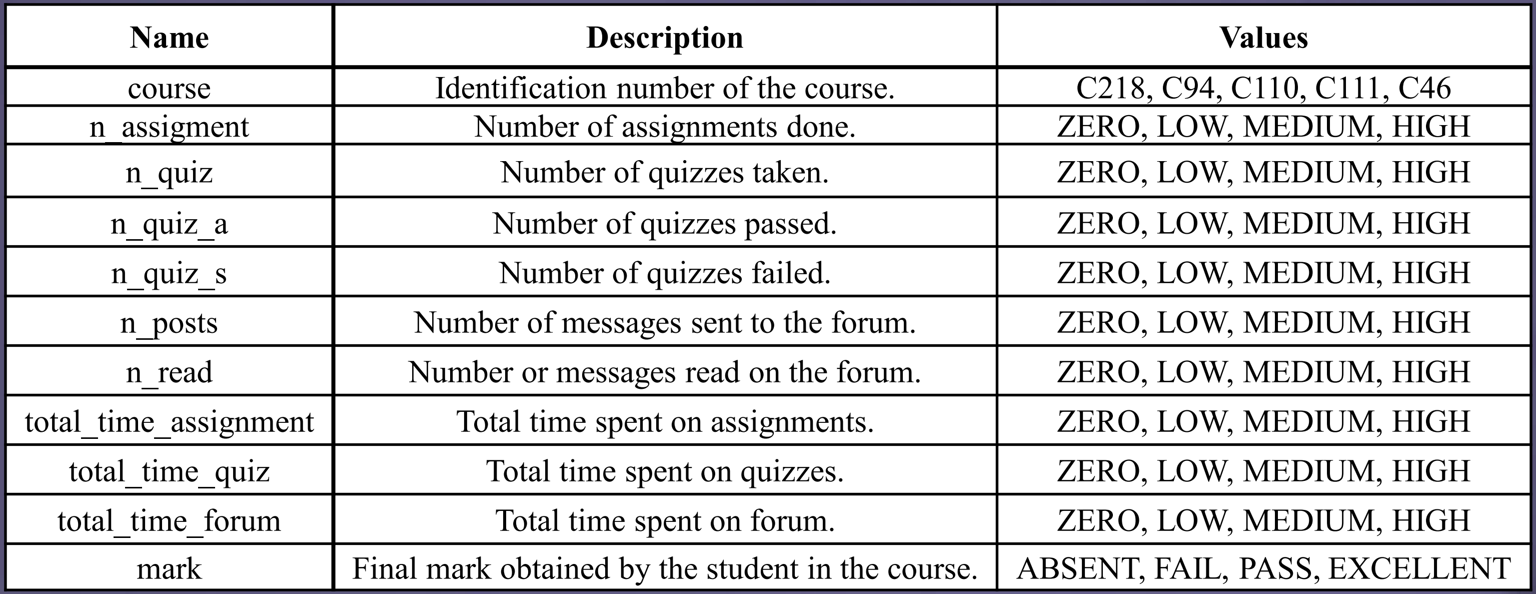
Data

- In order to test the performance and usefulness of applying RARM to e-learning data, we have used student data gathered from the Moodle system.
- These data are from 230 students in 5 Moodle courses on computer science at the University of Córdoba about all activities that students perform on-line (e.g., assignments, forums and quizzes).
- This student usage data has been preprocessed in order to be transformed into a suitable format to be used by our data mining algorithms.

Experimentation and Results

Summary Table

- We have created a summary table which integrates the most important information about the on-line activities and the final marks obtained by students in the courses.

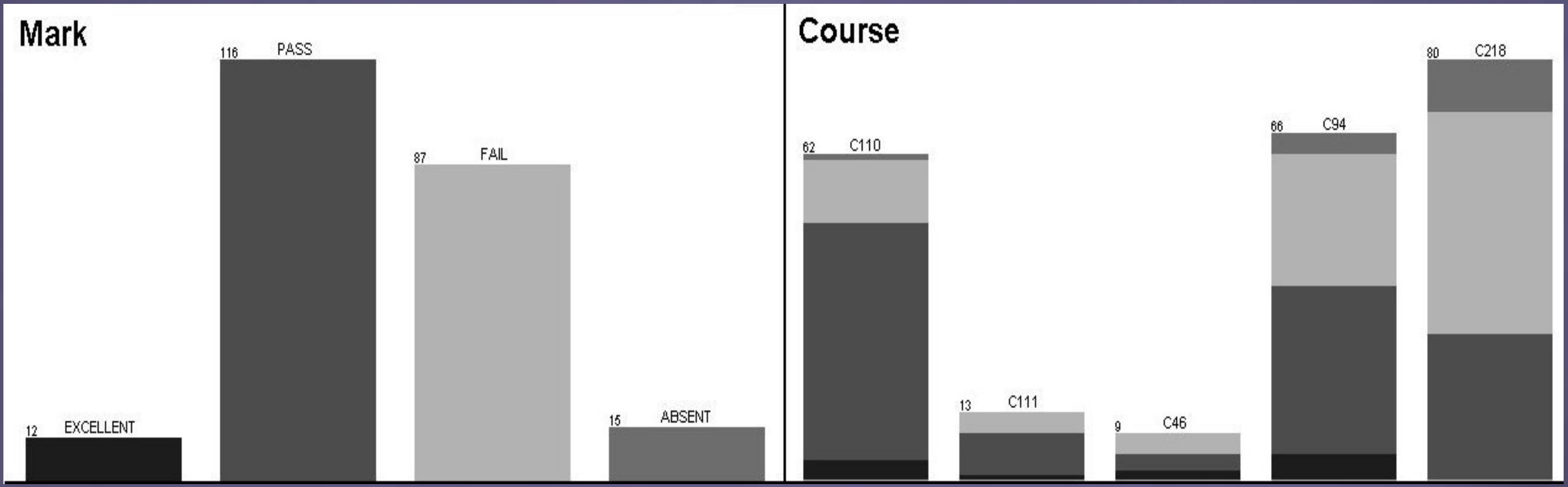


Name	Description	Values
course	Identification number of the course.	C218, C94, C110, C111, C46
n_assigment	Number of assignments done.	ZERO, LOW, MEDIUM, HIGH
n_quiz	Number of quizzes taken.	ZERO, LOW, MEDIUM, HIGH
n_quiz_a	Number of quizzes passed.	ZERO, LOW, MEDIUM, HIGH
n_quiz_s	Number of quizzes failed.	ZERO, LOW, MEDIUM, HIGH
n_posts	Number of messages sent to the forum.	ZERO, LOW, MEDIUM, HIGH
n_read	Number or messages read on the forum.	ZERO, LOW, MEDIUM, HIGH
total_time_assignment	Total time spent on assignments.	ZERO, LOW, MEDIUM, HIGH
total_time_quiz	Total time spent on quizzes.	ZERO, LOW, MEDIUM, HIGH
total_time_forum	Total time spent on forum.	ZERO, LOW, MEDIUM, HIGH
mark	Final mark obtained by the student in the course.	ABSENT, FAIL, PASS, EXCELLENT

Experimentation and Results

Imbalanced Attributes

- Due to the way their values are distributed, the course and mark attributes are clearly imbalanced, i.e., they have one or more values with a very low percentage of appearance.



Experimentation and Results

Rare Class Association Rules

- We performed a comparison between ARM and different RARM algorithms to discover **Rare Class Association Rules**.
- A **Class Association Rule** is a special subset of association rules with the consequent of the rule limited to a target class label (only one predefined item in our case Mark attribute).

Item1* ∩ *item2* ∩ … ∩ *Itemn* → *Class

- In our specific context, these rules are very useful for educational purposes, since they show any existing relationships between the activities that students perform using Moodle and their final exam marks.
- To obtain Class Association Rules we have modified ARM and RARM algorithms in order to obtain only those rules that have a single attribute (in our case, the mark attribute) in their consequent.

Experimentation and Results

Parameters

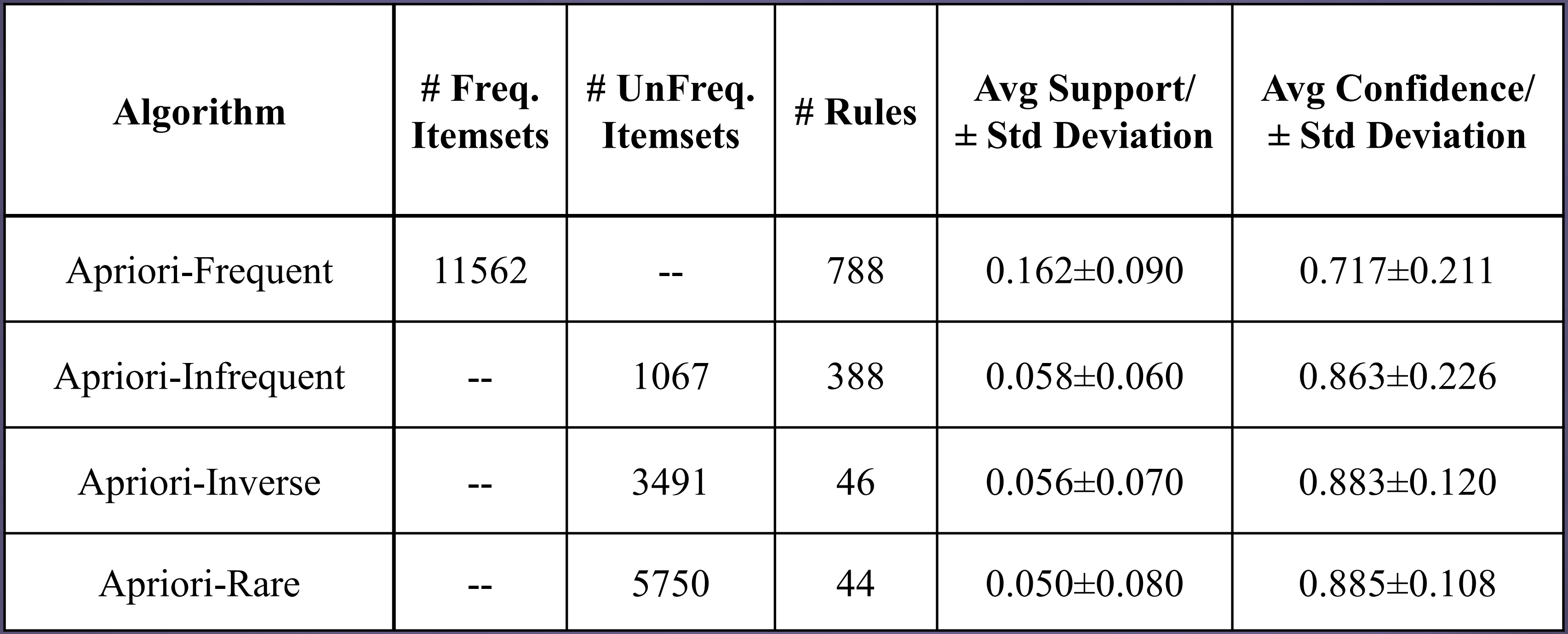
- We evaluated the four different Apriori proposals with the following configuration parameters:
  - **Apriori-Frequent**, setting the minimum support threshold at a very low value (0.05).
  - **Apriori-Infrequent**, **Apriori-Inverse** and **Apriori-Rare** setting the maximum support at 0.1.

We also assigned the value 0.7 as the confidence threshold for all the algorithms.

Experimentation and Results

Summary of Results

- Comparison Table of ARM and RARM proposals:



Algorithm	# Freq. Itemsets	# UnFreq. Itemsets	# Rules	Avg Support/ ± Std Deviation	Avg Confidence/ ± Std Deviation
Apriori-Frequent	11562	--	788	0.162±0.090	0.717±0.211
Apriori-Infrequent	--	1067	388	0.058±0.060	0.863±0.226
Apriori-Inverse	--	3491	46	0.056±0.070	0.883±0.120
Apriori-Rare	--	5750	44	0.050±0.080	0.885±0.108

Experimentation and Results

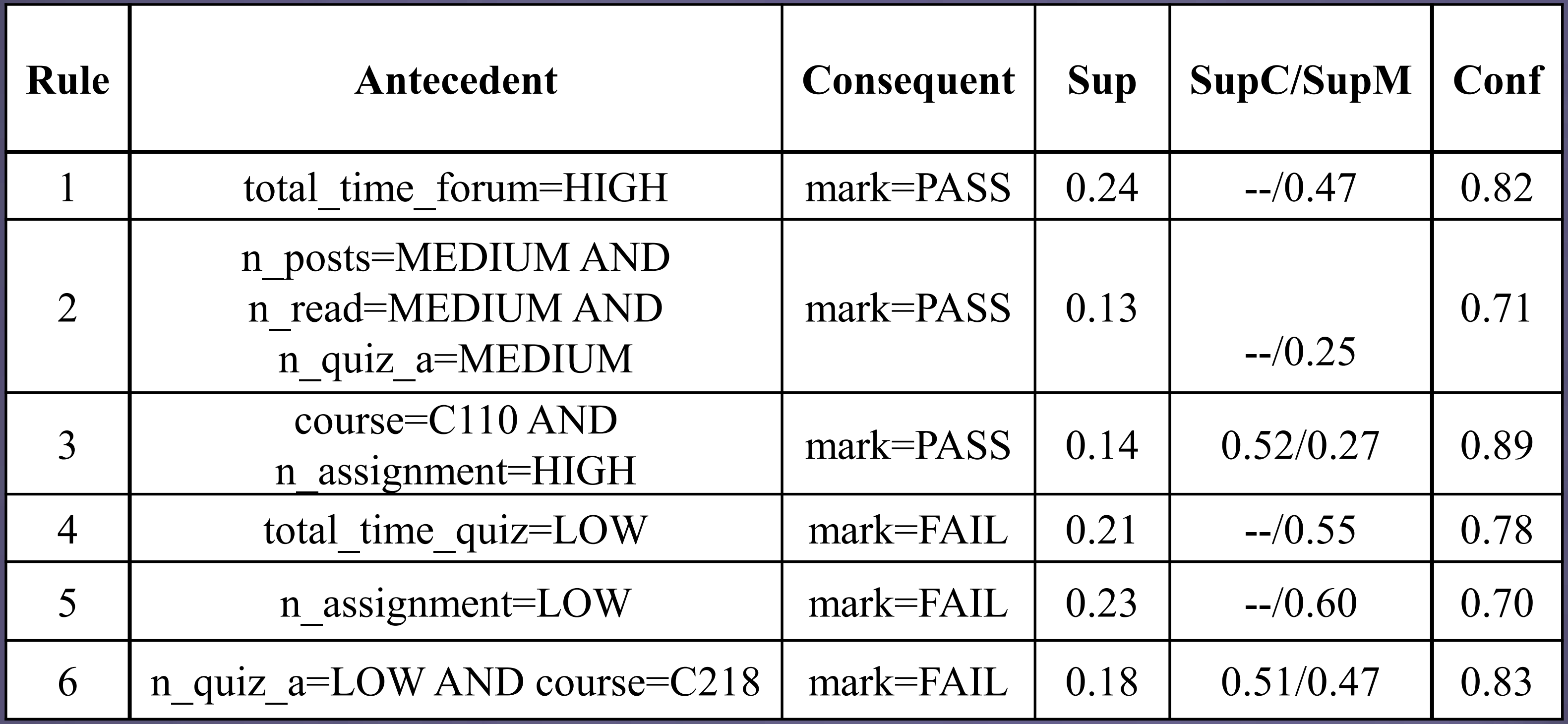
Examples of discovered rules

- Next, we show some examples of rules that were obtained using A) the ARM (Apriori) and B) RARM (Apriori-Rare) algorithms.
- For each rule, we show the antecedent and the consequent constructed, as well as some evaluation rule measures such as the support, the confidence and two different versions of the conditional support.

Experimentation and Results

Examples of discovered rules

- Rules extracted using the Apriori-Frequent algorithm.



Rule	Antecedent	Consequent	Sup	SupC/SupM	Conf
1	total_time_forum=HIGH	mark=PASS	0.24	--/0.47	0.82
2	n_posts=MEDIUM AND n_read=MEDIUM AND n_quiz_a=MEDIUM	mark=PASS	0.13	--/0.25	0.71
3	course=C110 AND n_assignment=HIGH	mark=PASS	0.14	0.52/0.27	0.89
4	total_time_quiz=LOW	mark=FAIL	0.21	--/0.55	0.78
5	n_assignment=LOW	mark=FAIL	0.23	--/0.60	0.70
6	n_quiz_a=LOW AND course=C218	mark=FAIL	0.18	0.51/0.47	0.83

Experimentation and Results

Examples of discovered rules

- Rules extracted using the Apriori-Rare algorithm.

Rule	Antecedent	Consequent	Sup	SupC/SupM	Conf
1	n_quiz=HIGH AND n_quiz_a=HIGH	mark=EXCELLENT	0.045	--/0.69	0.86
2	total_time_assignment=HIGH	mark=EXCELLENT	0.045	--/0.69	0.86
3	n_posts=HIGH AND course=C46	mark=EXCELLENT	0.045	1.00/0.69	1.00
4	total_time_assignment=ZERO AND total_time_forum=ZERO AND total_time_quiz=ZERO]	mark=ABSENT	0.050	--/0.76	0.78
5	n_posts=ZERO AND n_read=ZERO	mark=ABSENT	0.050	--/0.76	0.78
6	n_quiz=ZERO AND course=C111	mark=ABSENT	0.050	0.88/0.76	1.00

The end

Thank you for your interest.

Questions? Comments?