Association Analysis for an Online Education System

Behrouz Minaei,

Gerd Kortemeyer, and Bill Punch

minaeibi@cse.msu.edu

Department of Computer Science and Engineering
Michigan State University

IEEE IRI 2004, Las Vegas, Nov 10th 2004

Overview, LON-CAPA

- Latest online educational system developed at MSU, the **L**earning **O**nline **N**etwork with **C**omputer-**A**ssisted **P**ersonalized **A**pproach (LON-CAPA)
  - Three architectural layers:
    - Distributed cross-institutional content repository
    - Assembly tool for content resources
    - Full-featured course management system to readily deploy content
- Learning Content Management System
  - 3 middle school, 16 high schools, and 17 universities nationwide
  - 60,000 re-useable learning resources, including more than 18,000 sophisticated randomizing problems
- Assessment System
  - Online assessment with immediate feedback and multiple tries
  - Different students get different versions of the same problem
  - Different options, graphs, images, numbers, or formulas
- Open-Source and Free (GPL, Runs on Linux)

LON-CAPA Data

23,000 content pages
18,600 homework and exam problems
1100 simulations and animations

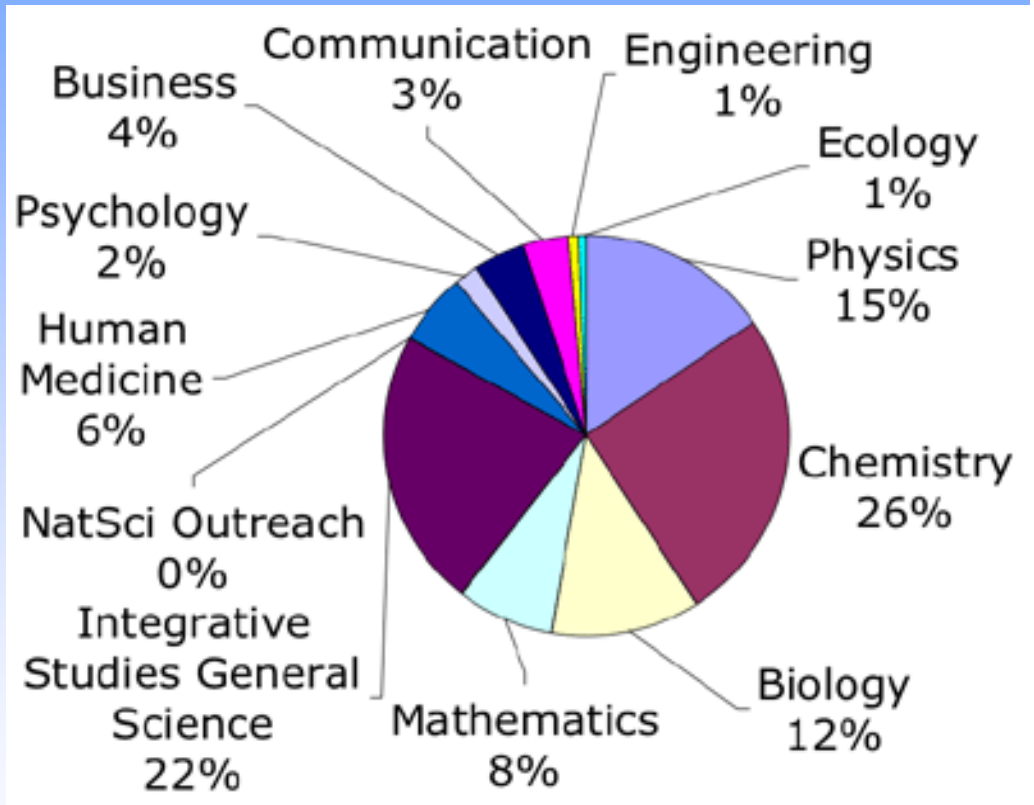
12,500 images

500 movies

- Three kinds of growing data sets:
  - Educational resources: web pages, demonstrations, simulations, individualized problems, quizzes, and examinations
  - Information about users who create, modify, assess, or use these resources.
  - Data about how students use and access the educational materials

MSU– Fall 2003

- 40 courses at MSU
- Total student enrollment approximately 3,067 (out of 13,400 total global student-users)
- Physics, Astronomy, Chemistry, Advertising, Biology, Biochemistry, Math, Finance, Geology, Statistics, Psychology, Civil Eng., etc.
- LON-CAPA collects data for every single access
- Logs are huge and distributed



Research Objective(s)

Help instructors predict the approaches that students will take for some types of problems

Can be used to identify those students who are at risk, especially in very large classes

- Find and Compare Association Rules amongst contrasting groups
  - Gender: Male, Female
  - Ethnicity: Caucasian, Black, Asian
  - Grades: Passed(>2), Failed(<=2)
  - A course/homework/problem in different semester

Related Work

- Bay, S. D. and Pazzani, M. J.:
  - Conjunction of attributes and values that differ meaningfully in their distribution across groups
  - STUCCO (Search and Testing for Understandable Consistent Contrast)
    - Finding Significant Contrast Sets: X^2 tests the null hypothesis that contrast-set support is equal across all groups
  - The goal in this work is to find the surprising contrasting sets, but our objective is to find the contrasting rules, introducing new measures for finding the significant differences between the groups elements.

Methodology

- Selecting data from course and students databases
- Preprocessing; cleansing unuseful data
- Feature subset extraction/selection
- Discretizing the continuous features
- Pruning the values of feature with very high support
- Select an interested contrast group
- Applying MCR algorithm given a contrasting feature
- Post-processing to identify the rule interestingness
- Select another measure or contrast group and repat the procedure

MCR (Mining Contrasting Rules) Algorithm

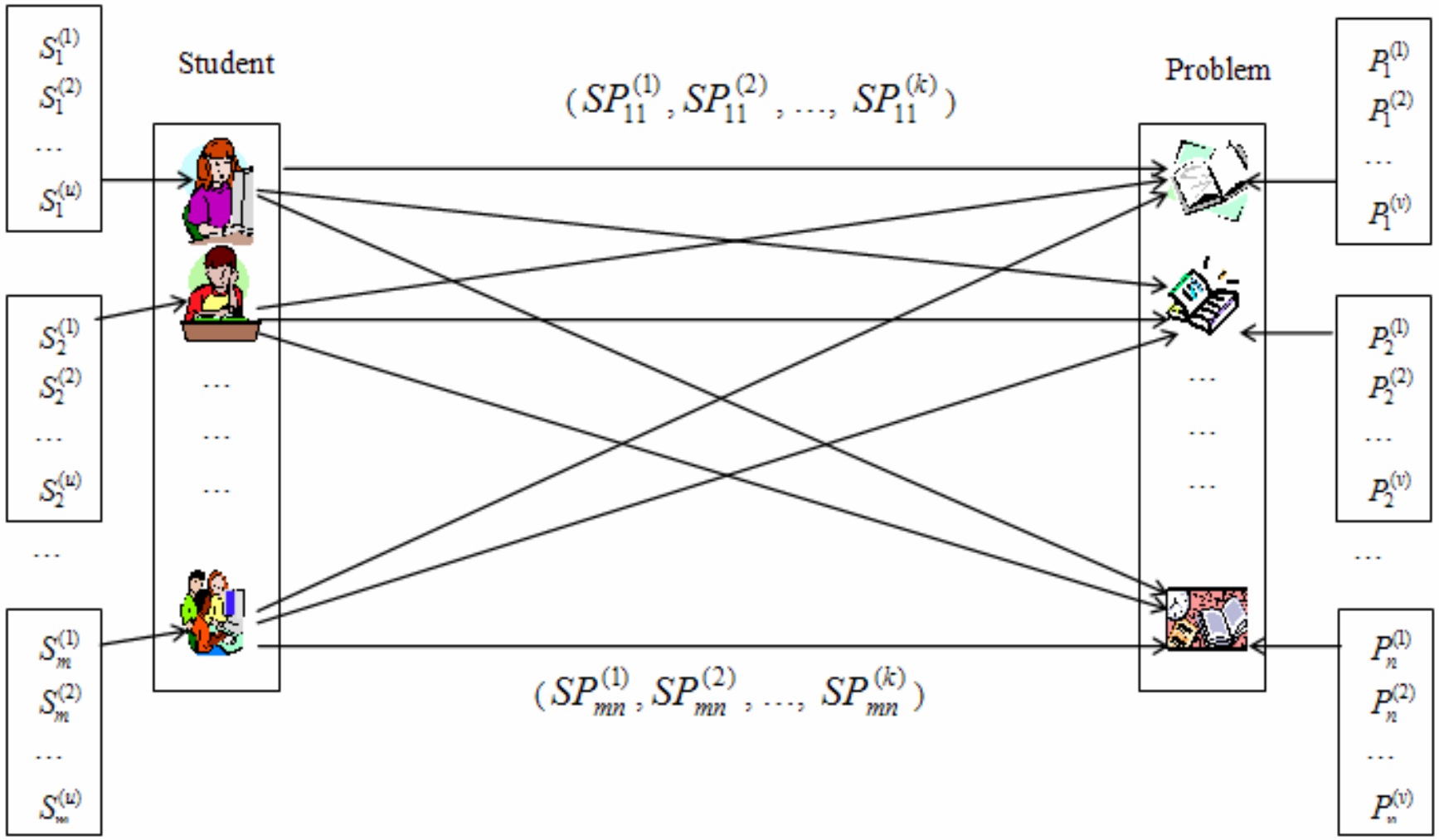
Input:

D – Input set of *N* transactions of students per problems data
A – Interested attribute includes contrast groups
σ – Minimum (very) low support
Ω – Measure for ranking the rules
k – Number of top interesting rules
M – Number of contrasting elements to be compared

- Divide data set *D* based on contrasting elements in *A* into *M* spaces
- *for* $j = 1$ to *M*
- Find the closed frequent itemsets for $D(j)$ given σ
- Generate possible rules for $D(j)$ based on the frequent itemsets
- *end*
- Find common rules among the *M* contrast groups
- Rank the common rules with respect to the Ω
- Sort the rules with respect to their rank; Select *k-top* rules;
- Validate selected rules *R* as a candidate set of interesting rules (optional)

return *R*

Data Model


Student
Problem
$(SP_{11}^{(1)}, SP_{11}^{(2)}, \ldots, SP_{11}^{(k)})$
$(SP_{mn}^{(1)}, SP_{mn}^{(2)}, \ldots, SP_{mn}^{(k)})$
$S_1^{(1)}$ $S_1^{(2)}$ ... $S_1^{(u)}$
$S_2^{(1)}$ $S_2^{(2)}$ ... $S_2^{(u)}$
$S_m^{(1)}$ $S_m^{(2)}$ ... $S_m^{(u)}$
$P_1^{(1)}$ $P_1^{(2)}$ ... $P_1^{(v)}$
$P_2^{(1)}$ $P_2^{(2)}$ ... $P_2^{(v)}$
$P_n^{(1)}$ $P_n^{(2)}$ ... $P_n^{(v)}$

Experimental Setup

Data set	Course Title	# of Students	# of Problems	Size of Activity log	# of Transactions
LBS 271	Physics I	200	174	152.1 MB	32,394
BS 111	Biological Science	382	235	239.4 MB	71,675
CEM 141	General Chemistry I	2048	114	754.8 MB	190,859

Data set	Success = YES	Female	Male	Passed	Failed	Ethnic = Caucasian
LBS 271	29,515	20,468	11696	29,412	2,752	28,552
	91.4%	63.6%	36.6%	91.4%	8.6%	88.8%
BS 111	54612	44,650	27025	37,365	34,310	57340
	76.2%	62.3%	37.7%	52.1%	47.9%	80.0%
CEM 141	176,496	106,296	84563	121,540	69,319	155,633
	92.5%	55.7%	44.3%	63.7%	36.3%	81.5%

Experiments were conducted on a 1.7 GHz Pentium 4 PC running RedHat Linux 7.3 kernel x-2.4.20-19 with 1GB RAM

Feature, Discretization

Student Attributes:

GPA

major

ethnic

Msu_Lt_Grd_Pt_Avg

Msu_Lt_Passed_Hours

Msu_Lt_Cmplt_Hours

Class_Code

Grade_Code

Hs_Gpa_Type_Code

Hs_Gpa

Birth_Date

Adr_Cnty_Code

Student*Problem Attributes:

succ

tries

time

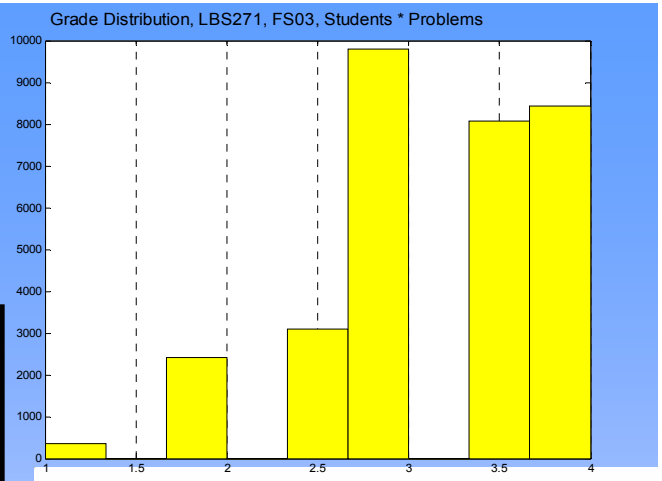
Problem Attributes

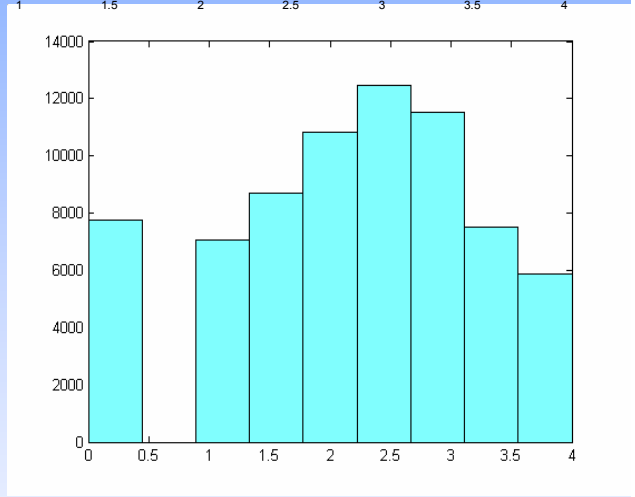
DoDiff, DoDisc, AvgTries

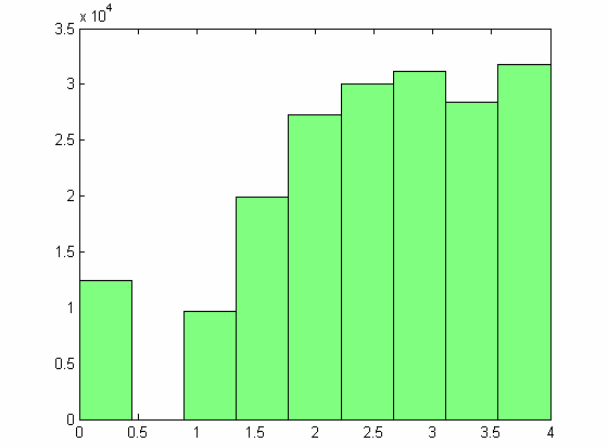
Aggregation of Attributes

2-Classes (Failed, Passed)

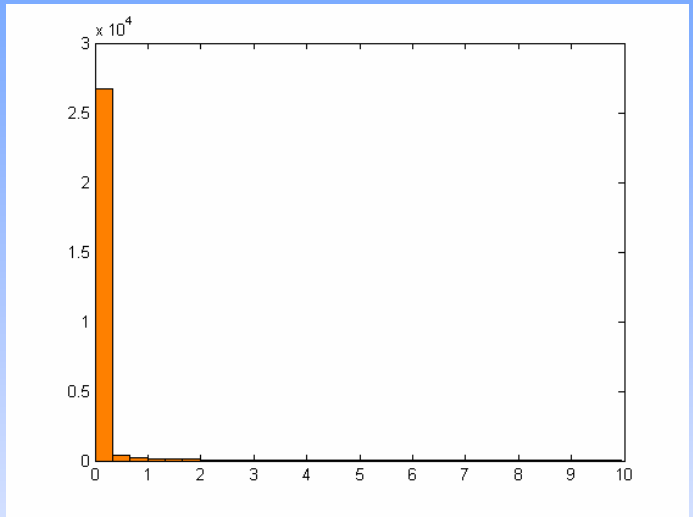
3-Classes (Low, Middle, High)

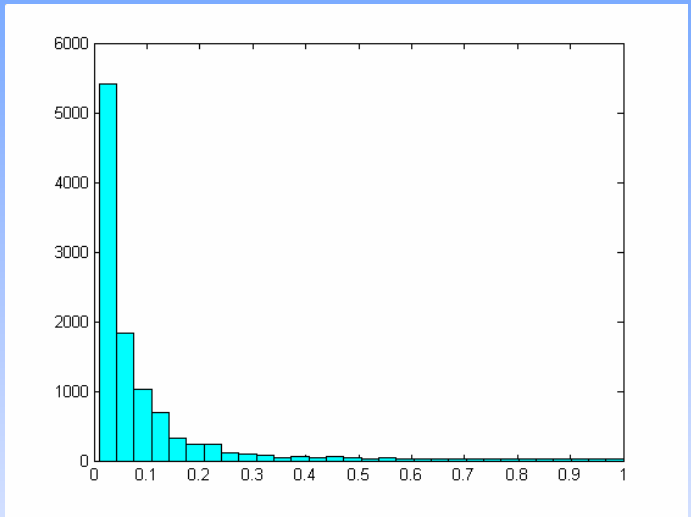


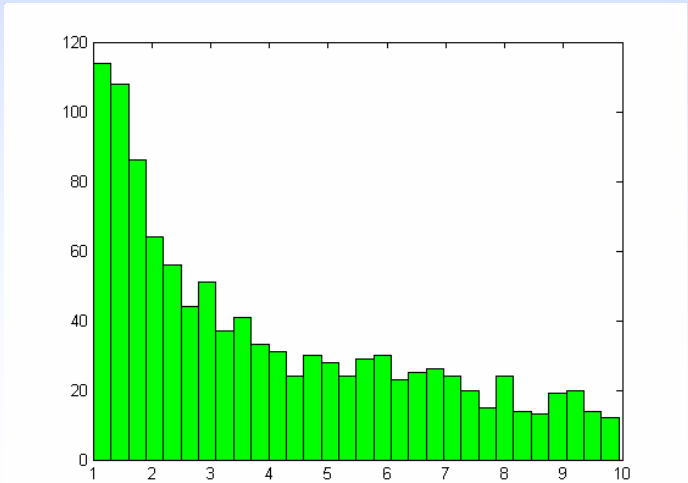


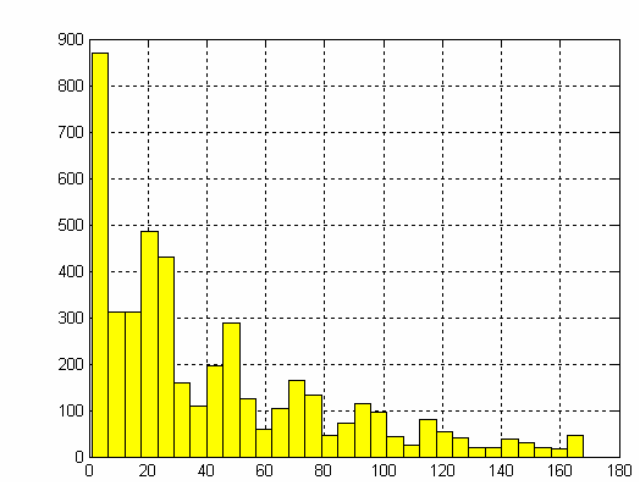


Discretizating “time”



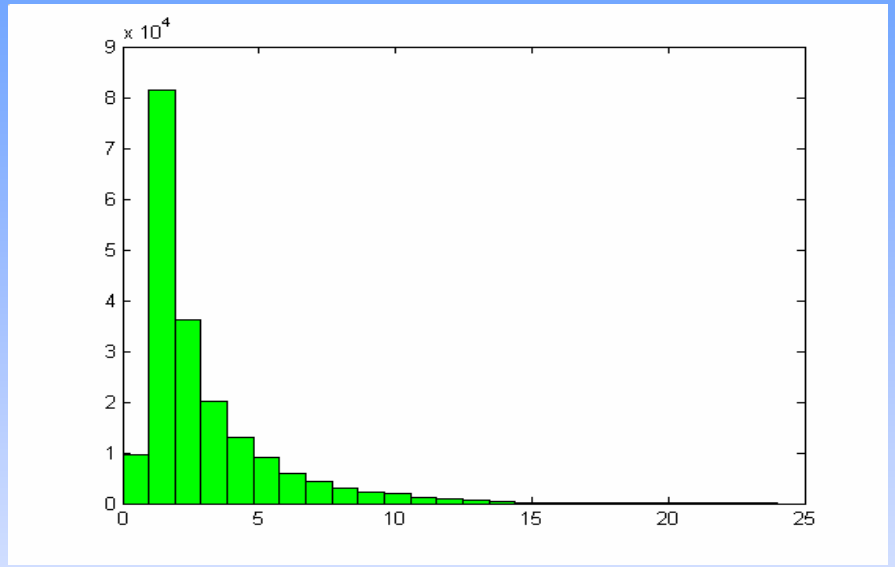


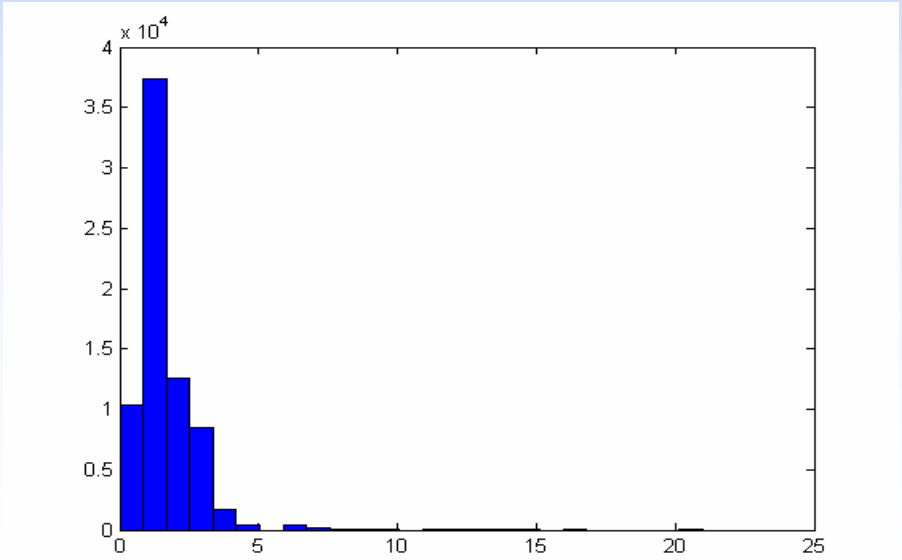


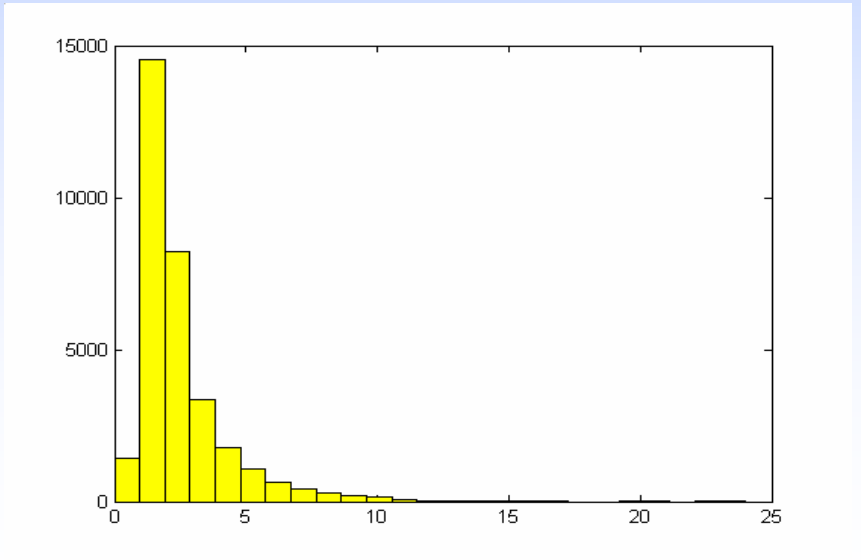


Discretizating “tries”

Data set	Try = 0	Try = 1	Try = 2
LBS 271	1,430 4.5%	14,530 45.2%	8,220 25.6%
BS 111	10,400 14.5%	37,378 52.2%	12,540 17.5%
CEM 141	9,558 5.0%	81,521 42.7%	36,259 19.0%







Experimental Evaluation

- Difficult to evaluate the success of the method
  - This is an unsupervised evaluation
  - Present to the expertise, subjective validation
  - Compare the results with some related algorithm (STUCCO) on a common data set (sensus.data)
  - The baseline method can be minimum threshold of statistical difference for contrasting rules with respect to a ranking measure

Experimental Results

- Example (LBS271, Gender)

(Age=20 & GPA=[3.5,4] & Tries=1) ==> Male [934 (8.0)%] (s=2.9%, c=20.7%)

(Age=20 & GPA=[3.5,4] & Tries=1) ==> Female [3586 (17.5)%] (s=11.1%, c=79.3%)

- Example (CEM141, 2-Classes)

(GPA=[2,2.5) & Sex=Female) ==> Passed [1648 (1.4)%] (s=0.9%, c=12.4%)

(GPA=[2,2.5) & Sex=Female) ==> Failed [11639 (16.8)%] (s=6.1%, c=87.6%)

- How can we find the most surprising rules?
  - Order the rules
  - Need some measurement for ranking the rules

Criteria for Rule Ranking

1. | n1/N1 – n2/N2 |
2. | n1/n2 – N1/N2 |
3. Odds Ratio = (p/(1-p))/(q/(1-q)) where p = n1/N1 and q = n2/N2
4. Log Odds Ratio = |log ((p/(1-p))/(q/(1-q)))|
5. Chi-Square value = $\sum_i \sum_j \frac{(Oij - Eij)^2}{Eij}$
6. Entropy
7. Gini

Discussion

CEM 141_FS03

First Measure

(Lt_GPA=[3,3.5)) ==> Passed [44187 (36.4)%] (s=23.2%, c=87.6%)
(Lt_GPA=[3,3.5)) ==> Failed [6283 (9.1)%] (s=3.3%, c=12.4%)

Second Measure

(Age=19 & & Lt_GPA=[3,3.5) & Major=MECH_EGR & Sex=Male) ==> Passed [2163 (1.8)%] (s=1.1%, c=95.5%)
(Age=19 & Lt_GPA=[3,3.5) & Major=MECH_EGR & Sex=Male) ==> Failed [103 (0.1)%] (s=0.1%, c=4.5%)

Odds-Ratio

(Lt_GPA=[2,2.5) & Sex=Female & Time<30_second & Tries=2) ==> Passed [123 (0.1)%] (s=0.1%, c=14.9%)
(Lt_GPA=[2,2.5) & Sex=Female & Time<30_second & Tries=2) ==> Failed [705 (1.0)%] (s=0.4%, c=85.1%)

Logs Odds-Ratio

(GPA=[3,3.5) & Lt_GPA=[3,3.5) & Sex=Male & Time=1_20_hours) ==> Passed [1156 (1.0)%] (s=0.6%, c=92.2%)
(GPA=[3,3.5) & Lt_GPA=[3,3.5) & Sex=Male & Time=1_20_hours) ==> Failed [98 (0.1)%] (s=0.1%, c=7.8%)

Chi- Square

(Major=PREDENTAL & Time=1_5_minutes & Tries=2) ==> Passed [122 (0.1)%] (s=0.1%, c=63.5%)
(Major=PREDENTAL & Time=1_5_minutes & Tries=2) ==> Failed [70 (0.1)%] (s=0.0%, c=36.5%)

Gini/Entropy

(Lt_GPA=[1.5,2)) ==> Passed [1133 (0.9)%] (s=0.6%, c=7.7%)
(Lt_GPA=[1.5,2)) ==> Failed [13493 (19.5)%] (s=7.1%, c=92.3%)

Conclusions

- L-C servers are tracking students' activities in large logs
- Developed an algorithm to discover a set of surprising contrasting rules
- This help both instructors and students:
  - Instructor: to design the course more effectively, detect anomaly
  - Students: use the resources more efficiently
- Future Work
  - Include the fixed attributes of the problems, (clustering, Bloom Taxonomy, etc.)
  - More Measurements tend toward discover higher coverage rules
  - Extend to contrasting groups with many elements
  - Build a tool to do all the phases in a package; pass the data through a magic box to find some obscure patterns
  - Tools to recommend tasks, automatically adapt course materials
  - Tools can be personalized, manually or automatically

References

- Bay, S. D. and Pazzani, M. J., “Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, 2001.
- Bay, S. D. “Multivariate Discretization for Set Mining”. *Knowledge and Information Systems,* 2001.
- Bay, S. D. and Pazzani, M. J. “Discovering and Describing Category Differences: What makes a discovered difference insightful?”, *Proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society*, 2000.
- Bay, S. D. and Pazzani, M. J. “Detecting Change in Categorical Data: Mining Contrast Sets”, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 302-306*, 1999*
- Minaei-Bidgoli, B., Punch, W.F., “Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System”, *Proc. of the Genetic and Evolutionary Computation Conference GECCO* 2003, pp. 2252-2263
- Tan, P.N., Steinbach M., and Kumar V., *Introduction to Data Mining*, to be appear as a book, 2004
- Agrawal, R., Srikant, R. "Fast Algorithms for Mining Association Rules", *Proceeding of the 20th International Conference on Very Large Databases*, Santiago, Chile, September 1994