Mining Interesting Contrast Rules for a Web-based Educational System

Behrouz Minaei-Bidgoli, Pang-Ning Tan, and William F. Punch
Computer Science Department, Michigan State University,
East Lansing, MI, 48824, USA
{minaeibi, ptan, punch}@cse.msu.edu

Abstract

Web-based educational technologies allow educators to study how students learn (descriptive studies) and which learning strategies are most effective (causal/predictive studies). Since web-based educational systems collect vast amounts of student profile data, data mining and knowledge discovery techniques can be applied to find interesting relationships between attributes of students, assessments, and the solution strategies adopted by students. This paper focuses on the discovery of interesting contrast rules, which are sets of conjunctive rules describing interesting characteristics of different segments of a population. In the context of web-based educational systems, contrast rules help to identify attributes characterizing patterns of performance disparity between various groups of students. We propose a general formulation of contrast rules as well as a framework for finding such patterns. We apply this technique to an online educational system developed at Michigan State University called LON-CAPA.

Keywords

Data mining, association analysis, rule interestingness, contrast rules, web-based educational system

1. Introduction

Many web-based educational systems with different capabilities and approaches have been developed to deliver online education in an academic setting. In particular, Michigan State University (MSU) has pioneered systems to provide an infrastructure for online instruction. The research presented in this paper was part of the latest online educational system developed at MSU called the *Learning Online Network with Computer-Assisted Personalized Approach* (LON-CAPA) [1-2].

LON-CAPA involves three types of large data sets: 1) educational resources such as web pages, demonstrations, simulations, and individualized problems designed for use on homework assignments, quizzes, and examinations; 2) information about users who create, modify, assess, or use these resources; and 3) activity log databases which log actions taken by students in solving homework assignment and exam problems.

This paper investigates methods for finding interesting rules based on the characteristics of groups of students or assignment problems. More specifically, our research is guided and inspired by the following questions: Can we identify the different groups of students enrolled in a particular course based on their demographic data? Which attribute(s) best explain the performance disparity among students over different sets of assignment problems? Are the same disparities observed when analyzing student performance in different sections or semesters of a course?

We address the above questions using a technique called contrast rules. Contrast rules are sets of conjunctive rules describing important characteristics of different segments of a population. Consider the following toy example of 200 students who enrolled in an online course. The course provides online reading materials that cover the concepts related to assignment problems. Students may take different approaches to solve the assignment problems. Among these students, 109 students read the materials before solving the problems while the remaining 91 students directly solve the problems without reviewing the materials. In addition, 136 students eventually passed the course while 64 students failed. This information summarized in a 2×2 contingency table as shown in Table 1.

Table 1. A contingency table of student success vs. study habits for an online course

	Passed	Failed	Total
Review materials	95	14	109
Do not review	41	50	91
Total	136	64	200

The table shows that there are interesting contrasts between students who review the course materials before solving the homework problems and students who do not review the materials. The following contrast rules can be induced from the contingency table:

Review materials $\Rightarrow$ Passed, $s = 47.5\%$, $c = 87.2\%$
Review materials $\Rightarrow$ Failed, $s = 7.0\%$, $c = 12.8\%$

Figure 1. A contrast rule extracted from Table 1

where s and c are the support and confidence of the rules [3]. These rules suggest that students who review the materials are more likely to pass the course. Since there is a large difference between the support and confidence of both rules, the observed contrast is potentially interesting. Other examples of interesting contrast rules obtained

from the same contingency table are shown in Figures 2 and 3.

Passed $\Rightarrow$	Review materials,	s = 47.5%, c = 69.9%
Failed $\Rightarrow$	Review materials,	s = 7.0%, c = 15.4%

Figure 2. A contrast rule extracted from Table 1

Passed $\Rightarrow$	Review materials,	s = 47.5%, c = 69.9%
Passed $\Rightarrow$	Do not review,	s = 20.5%, c = 30.1%

Figure 3. A contrast rule extracted from Table 1

Not all contrasting rule pairs extracted from Table 1 are interesting, as the example in Figure 4 shows.

Do not review $\Rightarrow$	Passed,	s = 20.5%, c = 45.1%
Do not review $\Rightarrow$	Failed,	s = 25.0%, c = 54.9%

Figure 4. A contrast rule extracted from Table 1

The above examples illustrate some of the challenging issues concerning the task of mining contrast rules:

1) There are many measures applicable to a contingency table. Which measure(s) yield the most significant/interesting contrast rules among different groups of attributes?
2) Many rules can be extracted from a contingency table. Which pair(s) of rules should be compared to define an interesting contrast?

This paper presents a general formulation of contrast rules and proposes a new algorithm for mining interesting contrast rules. The rest of this paper is organized as follows: Section 2 provides a brief review of related work. Section 3 offers a formal definition of contrast rules. Section 4 gives our approach and methodology to discover the contrast rules. Section 5 describes the LON-CAPA data model and an overview of our experimental results.

2. Background

In order to acquaint the reader with the use of data mining in online education, we present a brief introduction of association analysis and measures for evaluating association rules. Next, we explain the history of data mining in web-based educational systems. Finally, we discuss previous work related to contrast rules.

2.1. Association analysis

Let $I = \{i_1, i_2, \ldots, i_m\}$ be the set of all items and $T = \{t_1, t_2, \ldots, t_N\}$ the set of all transactions where m is the number of items and N is the number of transactions. Each transaction t_j is a set of items such that $t_j \subseteq I$. Each transaction has a unique identifier, which is referred to as TID. An *association rule* is an implication statement of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and X and Y are disjoint, that is, $X \cap Y = \varnothing$. X is called the antecedent while Y is called the consequence of the rule [3, 4].

Support and confidence are two metrics, which are often used to evaluate the quality and interestingness of a rule. The rule $X \Rightarrow Y$ has support, s, in the transaction set, T, if s% of transactions in T contains $X \cup Y$. The rule has *confidence*, c, if c% of transactions in T that contain X also contains Y. Formally, support is defined as shown in Eq. (1),

$$s(X \Rightarrow Y) = \frac{s(X \cup Y)}{N}, \quad (1)$$

where N is the total number of transactions, and confidence is defined in Eq. (2).

$$c(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)} \quad (2)$$

Another measure that could be used to evaluate the quality of an association rule is presented in Eq. (3).

$$RuleCoverage = \frac{s(X)}{N} \quad (3)$$

This measure represents the fraction of transactions that match the left hand side of a rule.

Techniques developed for mining association rules often generate a large number of rules, many of which may not be interesting to the user. There are many measures proposed to evaluate the interestingness of association rules [17-18]. Silberschatz and Tuzhilin suggest that interestingness measures can be categorized into two classes: objective and subjective measures [6].

An objective measure is a data-driven approach for evaluating interestingness of rules based on statistics derived from the observed data. In the literature different objective measures have been proposed [5]. Examples of objective interestingness measure include support, confidence, correlation, odds ratio, and cosine.

Subjective measures evaluate rules based on the judgments of users who directly inspect the rules [6]. Different subjective measures have been addressed to discover the interestingness of a rule [6]. For example, a *rule template* [10] is a subjective technique that separates only those rules that match a given template. Another example is *neighborhood*-based interestingness [11], which defines a single rule's interestingness in terms of the supports and confidences of the group in which it is contained.

2.2. Data mining for online education systems

Recently, several researchers have worked on the application of data mining to examine or classify students' problem-solving approaches within web-based educational systems. For example, we previously developed tools for predicting the student performance with respect to average values of student attributes versus the overall problems of an online course [2]. Zaiane [12] suggested the use of web mining techniques to build an agent that recommends on-line learning activities in a web-based course. Ma et al. [13] focused on one specific task of using association rule mining to select weak

students for remedial classes. This previous work focused on finding association rules with a specific rule consequent (i.e. a student is weak or strong). Herein, we propose a general formulation of contrast rules as well as a framework for finding such patterns.

2.3. Related work

An important goal in data mining is the discovery of major differences among segments of population. Bay and Pazzani [14] introduced the notion of contrast sets as a conjunction of attributes and values that differ "meaningfully" in their distribution across groups. They used a chi-square test for testing the null hypothesis that contrast-set support is equal across all groups. They developed the STUCCO (Search and Testing for Understandable Consistent Contrast) algorithm to find significant contrast sets. Our work represents a general formulation for contrast rules using different interestingness measures. We show that alternative measures allow for different perspectives on the process of finding interesting rules.

Liu et al. [15] have also used a chi-square test of independence as a principal measure for both generating the association rules and identifying non-actionable rules. Below, we briefly discuss the chi-square test of independence and one of its shortcomings.

Chi-square testing is used as a method for verifying the independence or correlation of attributes. The chi-square test compares observed frequencies with the corresponding expected frequencies. The greater the difference between observed and expected frequencies, the greater is the power of evidence in favor of dependence and relationship. Let *CT* be a contingency table with *K* rows and *L* columns. The chi-square test for independence is shown in Eq. (5) where $1 \le i \le K$, and $1 \le j \le L$, and degree of freedom is $(K-1)(L-1)$.

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (4)$$

However, a drawback of this test is that the χ^2 value is not invariant under the *row-column scaling* property [5]. For example, consider the contingency table shown in Table 2(a). If χ^2 is higher than a specific threshold (e.g. 3.84 at the 95% significance level and degree of freedom 1), we reject the independence assumption. The chi-square value corresponding to Table 2(a) is equal to 1.82. Therefore, the null hypothesis is accepted. Nevertheless, if we multiply the values of that contingency table by 10, a new contingency table is obtained as shown in Table 2(b). The χ^2 value increases to 18.2 (>3.84). Thus, we reject the null hypothesis. We expect that the relationship between gender and success for both tables as being equal, even though the sample sizes are different. In general, this drawback shows that χ^2 is proportional to *N*.

Table 2. A contingency table proportional to table 1

(a)

	Passed	Failed	Total
Male	40	49	89
Female	60	51	111
Total	100	100	200

(b)

	Passed	Failed	Total
Male	400	490	890
Female	600	510	1110
Total	1000	1000	2000

3. Contrast Rules

In this section, we introduce the notion of contrast rules. Let *A* and *B* be two itemsets whose relationship can be summarized in a 2×2 contingency table as shown in Table 3.

Table 3. A contingency table for the binary case

	B	$\overline{B}$	Total
A	f_{11}	f_{12}	f_{1+}
$\overline{A}$	f_{21}	f_{22}	f_{2+}
Total	f_{+1}	f_{+2}	N

Let Ω be a set of all possible association rules that can be extracted from such a contingency table (Figure 5).

$$A \Rightarrow B,\ A \Rightarrow \overline{B},\ \overline{A} \Rightarrow B,\ \overline{A} \Rightarrow \overline{B}$$
$$B \Rightarrow A,\ \overline{B} \Rightarrow A,\ B \Rightarrow \overline{A},\ \overline{B} \Rightarrow \overline{A}$$

Figure 5. Set of all possible association rules for Table 3.

We assume that *B* is a target variable and *A* is a conjunction of explanatory attributes. Let μ be a set of measures that can be applied to a rule or contingency table. Examples of such measures include support, confidence, chi-square, odds ratio, correlation, cosine, Jaccard, and interest [5]. Below we provide a formal definition of "contrast rule."

Definition (General Formulation of Contrast Rules):

A contrast rule, *cr,* is a 4-tuple $<br, \upsilon(br), M, \Delta>$ where:

- $br \subset \Omega$, is the base rule,
- $\upsilon(br) \subset \Omega$ is a neighborhood to which the base rule *br* is compared,
- $M=<m_{base}, m_{neighbor}>$ is an ordered pair of measures where $m_{base}, m_{neighbor} \in \mu$, and m_{base} measures the rules in *br* and $m_{neighbor}$ measures the rules in $\upsilon(br)$,
- $\Delta(m_{base}(br), m_{neighbor}(\upsilon(br)))$ is a comparison function between $m_{base}(r)$ and $m_{neighbor}(\upsilon(br))$.

A contrast rule, *cr,* is interesting if and only if $\Delta(m_{base}(br), m_{neighbor}(\upsilon(br))) \ge \sigma$, where σ is a user defined threshold, which implies that there is a large difference between *br* and its neighborhood with respect to *M*.

Figure 6. Formal definition of a contrast rule

As shown in Figure 6, the contrast rule definition is based on a paired set of rules, base rule *br* and its neighborhood $\upsilon(br)$. The base rule is a set of association rules with which a user is interested in finding contrasting

association rules. Below are some examples that illustrate the definition.

***Example 1**: cr_1 (Difference of confidence)*

The first type of contrast rules examines the difference between rules $A \Rightarrow B$ and $A \Rightarrow \overline{B}$. An example of this type of contrast was shown in Figure 1. Let confidence be the selected measure for both rules. Let absolute difference be the comparison function. We can summarize this type of contrast as follows:

- *br*: $\{A \Rightarrow B\}$
- $\upsilon(r)$: $\{A \Rightarrow \overline{B}\}$
- *M*: <confidence, confidence>
- Δ: absolute difference

The evaluation criterion for this example is shown in Eq. 5. This criterion can be used for ranking different pairs of contrast rules.

$$\begin{aligned}\Delta &= |\, c(r) - c(\upsilon(r)) \,| \\ &= |\, c(A \Rightarrow B) - c(A \Rightarrow \overline{B}) \,| \\ &= \left| \frac{f_{11}}{f_{1+}} - \frac{f_{12}}{f_{1+}} \right| = \left| \frac{f_{11} - f_{12}}{f_{1+}} \right|, \end{aligned} \tag{5}$$

where f_{ij} corresponds to the values in the i-th row and j-th column of Table 3. Since $c(A \Rightarrow B) + c(A \Rightarrow \overline{B}) = 1$, therefore,

$$\begin{aligned}\Delta &= |\, c(A \Rightarrow B) - c(A \Rightarrow \overline{B}) \,| \\ &= |\, 2c(A \Rightarrow B) - 1 \,| \\ &\propto c(A \Rightarrow B). \end{aligned}$$

Thus, the standard confidence measure is sufficient to detect an interesting contrast of this type.

***Example 2**: cr_2 (Difference of proportion)*

An interesting contrast could be considered between rules $B \Rightarrow A$ and $\overline{B} \Rightarrow A$. An example of this contrast was shown in Figure 2. Once again, let confidence be the selected measure for both rules. Let absolute difference be the comparison function. We can summarize this type of contrast as follows:

- *br*: $\{B \Rightarrow A\}$
- $\upsilon(br)$: $\{\overline{B} \Rightarrow A\}$
- *M*: <confidence, confidence>
- Δ: absolute difference

The evaluation criterion for this example is shown in Eq. 6, where Δ is defined as follows:

$$\begin{aligned}\Delta &= |\, c(r) - c(\upsilon(r)) \,| \\ &= |\, c(B \Rightarrow A) - c(\overline{B} \Rightarrow A) \,| \\ &= \left| \frac{f_{11}}{f_{+1}} - \frac{f_{12}}{f_{+2}} \right| = \left| \rho(A \Rightarrow B) - \rho(A \Rightarrow \overline{B}) \right| \end{aligned} \tag{6}$$

where ρ, is the rule proportion [19] and is defined in Eq. 7.

$$\rho(A \Rightarrow B) = \frac{P(AB)}{P(B)} = c(B \Rightarrow A) \tag{7}$$

***Example 3**: cr_3 (Correlation and Chi-Square)*

Correlation is a broadly used statistical measure for analyzing the relationship between two variables. The correlation between A and B in Table 3 is measured as follows:

$$corr = \frac{f_{11} f_{22} - f_{12} f_{21}}{\sqrt{f_{1+} f_{+1} f_{2+} f_{+2}}} \tag{8}$$

The correlation measure compares the contrast between the following set of base rules and their neighborhood rules:

- *br* is $\{A \Rightarrow B, B \Rightarrow A, \overline{A} \Rightarrow \overline{B}, \overline{B} \Rightarrow \overline{A}\}$
- $\upsilon(br)$ is $\{A \Rightarrow \overline{B}, \overline{B} \Rightarrow A, \overline{A} \Rightarrow B, B \Rightarrow \overline{A}\}$
- *M*: <confidence, confidence>,
- Δ: The difference in the square root of confidence products (see Eq. 9).

$$\Delta = \sqrt{c_1 c_2 c_3 c_4} - \sqrt{c_5 c_6 c_7 c_8} \tag{9}$$

where c_1, c_2, c_3, c_4, c_5, c_6, c_7, and c_8 correspond to $c(A \Rightarrow B)$, $c(B \Rightarrow A)$, $c(\overline{A} \Rightarrow \overline{B})$, $c(\overline{B} \Rightarrow \overline{A})$, $c(A \Rightarrow \overline{B})$, $c(\overline{B} \Rightarrow A)$, $c(\overline{A} \Rightarrow B)$, and $c(B \Rightarrow \overline{A})$ respectively. Eq. 10 is obtained by expanding Eq. 9.

$$\Delta = \sqrt{\frac{P(AB)}{P(A)} \frac{P(AB)}{P(B)} \frac{P(\overline{A}\overline{B})}{P(\overline{A})} \frac{P(\overline{A}\overline{B})}{P(\overline{B})}} - \sqrt{\frac{P(A\overline{B})}{P(A)} \frac{P(A\overline{B})}{P(\overline{B})} \frac{P(\overline{A}B)}{P(\overline{A})} \frac{P(\overline{A}B)}{P(B)}} \tag{10}$$

$$\Delta = \frac{P(AB)P(\overline{A}\overline{B}) - P(A\overline{B})P(\overline{A}B)}{\sqrt{P(A)P(B)P(\overline{A})P(\overline{B})}} \tag{11}$$

Eq. 11 is the correlation between A and B as shown in Eq. 8. Chi-square measure is related to correlation in the following way:

$$corr = \sqrt{\chi^2 / N} \tag{12}$$

Therefore, both measures are essentially comparing the same type of contrast.

Contrast rules and interestingness measures

Different measures have different perspectives on finding interesting rules. Specifically, each measure defines a base rule and a neighborhood of rules from which interesting contrast rules can be detected. In our

proposed approach a user can choose a measure and detect the corresponding contrast rules. In addition, the user has flexibility to choose a base rule/attribute according to what he or she is interested in. Then he or she selects the neighborhood rules as well as the measures to detect the base rule and its neighborhood. This is similar to rule template approaches (see 2.1). We implemented examples 1-3 for LON-CAPA data sets, which will be explained in section 5.

4. Algorithm

In this section we propose an algorithm to find surprising and interesting rules based on the characteristics of different segments of students/problems. The difficulty with algorithms such as Apriori is that when the minimum-support is high, we miss many interesting, but infrequent patterns. On the other hand if we choose a minimum-support that is too low the Apriori algorithm will discover so many rules that finding interesting ones becomes difficult.

Herein, we propose an automatic rule miner to discover hidden patterns amongst the contrast elements, even those with low support. We call this the Mining Contrast Rules (MCR) algorithm.

Mining Contrast Rules (MCR) Algorithm:

Input: D – Input set of N transactions
B – Target variable, the basis of interesting contrasts
σ – Minimum (very) low support
m – A measure for ranking the rules
k – Number of the most interesting rules
Divide data set D based on the values of the target variable
foreach j in B
  Select $D(j)$, a subset of transactions including j
  Find the set of closed frequent itemsets, $L(j)$ within $D(j)$
  foreach $\ell \in L(j)$
    Generate rule $\ell \Rightarrow j$
    Compute measure $m(\ell \Rightarrow j)$
  end
end
Find common rules among the different groups of rules
foreach br and $\upsilon(br)$ pair compute difference in measures, Δ
Sort the rules with respect to Δ
Select top k rules
return R

Figure 7. Mining Contrast Rules (MCR) algorithm for discovering interesting candidate rules

In order to employ the MCR algorithm, several steps must be taken. During the preprocessing phase, we remove items whose support is too high. For example, if 95% of students pass the course, this attribute will be removed from the itemsets so that it does not overwhelm other, more subtle rules. Then we must also select the target variable of the rules to be compared. This allows the user to focus the search space on subjectively interesting rules. If the target variable has C distinct values, we divide the data set, D, into C disjoint subsets based on the elements of the target variable, as shown in Figure 7. For example, in the case where gender is the target variable, we divide the transactions into male and female subsets to permit examination of rule coverage.

Using Borgelt's implementation[1] of the Apriori algorithm (version 4.21), we can find closed itemsets employing a simple filtering approach on the prefix tree [16]. A closed itemset is a set of items for which none of its supersets have exactly the same support as itself. The advantage of using closed frequent itemsets for our purposes is that we can focus on a smaller number of rules for analysis, and larger frequent itemsets, by discarding the redundant supersets.

We choose a very low minimum support to obtain as many frequent itemsets as is possible. Using perl scripts, we find the common rules between two contrast subsets. Finally, we rank the common rules with all of the previously explained measures, and then the top k rules of the sorted ranked-rules are chosen as a candidate set of interesting rules. Therefore an important parameter for this algorithm is minimum support, σ; the lower the σ, the larger the number of common rules. If the user selects a specific ranking measure, m, then the algorithm will rank the rules with respect to that measure.

5. Experiments

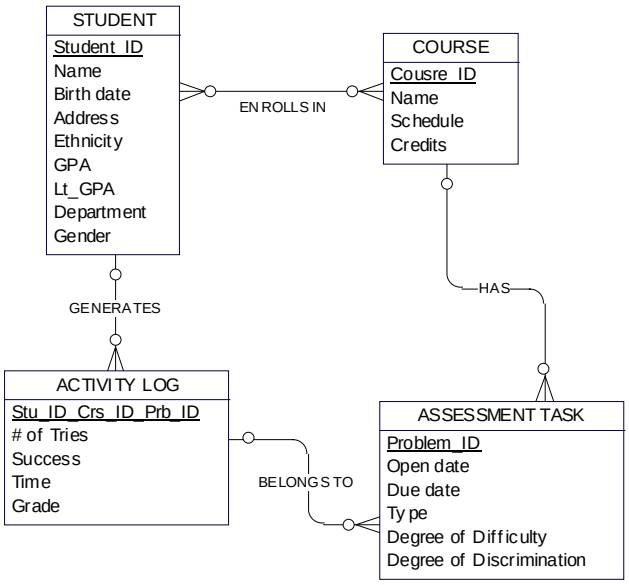

Figure 8. Entity Relationship Diagram for a LON-CAPA course

In this section we first provide a general model for data attributes, data sets and their selected attributes, and

[1] The code for this program is available at http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html.

then explain how we handle continuous attributes. Finally, we discuss our results and experimental issues.

5.1. Data model and attributes

In order to better understand the interactions between students and the online education system, a model is required to analyze the data. Ideally, this model would be both descriptive and predictive in nature. The model is framed around the interactions of the two main sources of interpretable data: students and assessment tasks (problems). Figure 8 shows the actual data model, which is frequently called an entity relationship diagram (ERD) since it depicts categories of data in terms of entities and relationships.

The attributes selected for association analysis are divided into four groups within the LON-CAPA system:

a) Student attributes: which are fixed for any student. Attributes such as Ethnicity, Major, and Age were not included in the data out of necessity – the focus of this work is primarily on the LON-CAPA system itself, so the demographics of students is less relevant. As a result, the following three attributes are included:

- *GPA*: is a continuous variable that is discretized into eight intervals between zero and four with a 0.5 distance.
- *Gender*: is a binary attribute with values Female and Male.
- *LtGPA* (Level Transferred (i.e. High School) GPA)**:** measured the same as GPA

b) Problem attributes: which are fixed for any problem. Among several attributes for the problems we selected the four following attributes:

- *DoDiff* (degree of difficulty)**:** This is a useful factor for an instructor to determine whether a problem has an appropriate level of difficulty. DoDiff is computed by the total number of students' submissions divided by the number of students who solved the problem correctly. Thus, DoDiff is a continuous variable in the interval [0,1] which is discretized into terciles of roughly equal frequency: easy, medium, and hard.
- *DoDisc* (degree of discrimination)**:** A second measure of a problem's usefulness in assessing performance is its discrimination index. It is derived by comparing how students whose performance places them in the top quartile of the class score on that problem compared to those in the bottom quartile. The possible values for DoDisc vary from –1 to +1. A negative value means that students in the lower quartile scored better on that problem than those in the upper. A value close to +1 indicates the higher achieving students (overall) performed better on the problem. We discretize this continuous value into terciles of roughly equal frequency: negatively-discriminating, non-discriminating, and positively-discriminating.
- *AvgTries* (average number of tries)**:** This is a continuous variable which is discretized into terciles of roughly equal frequency: low, medium, and high.

c) Student/Problem interaction attributes: We have extracted the following attributes per student per problem from the activity log:

- *Succ*: Success on the problem (YES, NO)
- *Tries*: Total number of attempts before final answer.
- *Time*: Total time from first attempt until the final answer is derived.

d) Student/Course interaction attributes: We have extracted the following attributes per student per course from the LON-CAPA system.

- *Grade*: Student's Grade, the nine possible labels for grade (a 4.0 scale with 0.5 increments). An aggregation of "grade" attributes is added to the total attribute list.
- *Pass-Fail*: Categorize students with one of two class labels: "Pass" for grades above 2.0, and "Fail" for grades less than or equal to 2.0.

5.2. Data sets

For this paper we selected three data sets from the LON-CAPA courses as shown in Table 4. For example the second row of the table shows that BS111 (Biological Science: Cells and Molecules) integrated 235 online homework problems, and 382 students used LON-CAPA for this course. BS111 had an activity log with approximately 239 MB of data. Though BS111 is a larger course than LBS271 (first row of the table), a physics course, it is much smaller than CEM141 (third row), general chemistry I. This course had 2048 students enrolled and its activity log exceeds 750MB, corresponding to more than 190k transactions of students attempting to solve homework problems.

Table 4. Characteristics of three MSU courses which used LON-CAPA in fall semester 2003

Data set	Course Title	# of Students	# of Prob.	Size of Activ. log	# of Trans.
LBS 271	Physics_I	200	174	152.1 MB	32,394
BS 111	BiologicalScience	382	235	239.4 MB	71,675
CEM141	Chemistry_I	2048	114	754.8 MB	190,859

For this paper we focus on two target variables, gender and pass-fail grades, in order to find the contrast rules involving these attributes. A constant difficulty in using any of the association rule mining algorithms is that they can only operate on binary data sets. Thus, in order to analyze quantitative or categorical attributes, some modifications are required – binarization – to partition the values of continuous attributes into discrete intervals and substitute a binary item for each discretized item. In this

paper, we mainly use equal-frequency binning for discretizing the attributes.

5.3. Results

This section presents some examples of the interesting contrast rules obtained from the LON-CAPA data sets. Since our approach is an unsupervised case, it requires some practical methods to validate the process. The interestingness of a rule can be subjectively measured in terms of its actionability (usefulness) or its unexpectedness [6-9].

One of the techniques for mining interesting association rules based on unexpectedness. Therefore, we divide the set of discovered rules into three categories:

1. *Expected and previously known*: This type of rule confirms user beliefs, and can be used to validate our approach. Though perhaps already known, many of these rules are still useful for the user as a form of empirical verification of expectations. For our specific situation (education) this approach provides opportunity for rigorous justification of many long-held beliefs.
2. *Unexpected*: This type of rule contradicts user beliefs. This group of unanticipated correlations can supply interesting rules, yet their interestingness and possible actionability still requires further investigation.
3. *Unknown*: This type of rule does not clearly belong to any category, and should be categorized by domain-specific experts. For our situations, classifying these complicated rules would involve consultation with not only the course instructors and coordinators, but also educational researchers and psychologists.

The following rule tables present five examples of the extracted contrast rules obtained using our approach. Each table shows the coded contrast rule and the "support" and "confidence" of that rule. Abbreviations are used in the rule code, and are summarized as follows: *Succ* stands for success per student per problem, *LtGPA* stands for transfer GPA, *DoDiff* stands for degree of difficulty of a particular problem, and *DoDisc* stands for degree of discrimination of a problem. In our experiments, we used three measures to rank the contrast rules:

5.3.1 Difference of confidences

The focus of this measure is on comparing the confidences of the contrast rules ($A \Rightarrow B$ and $A \Rightarrow \overline{B}$). Therefore, top rules found by this measure have a high value of confidence ratio (c_1/c_2). Contrast rules in Table 5 suggest that students in LBS 271 who are successful in homework problems are more likely to pass the course, and this comes with a confidence ratio c_1/c_2=12.7.

Table 5. LBS_271 data set, difference of confidences measure

Contrast Rules	Support & Confidence
(Succ=YES) ==> Passed	(s=86.1%, c=92.7%)
(Succ=YES) ==> Failed	(s=6.8%, c=7.3%)

This rule implies a strong correlation among the student's success in homework problems and his/her final grade. Therefore, this rule belongs to the first category; it is a known, expected rule that validates our approach.

Table 6. CEM_141 data set, difference of confidences measure

Contrast Rules	Support & Confidence
(Lt_GPA=[1.5,2)) ==> Passed	(s=0.6%, c=7.7%)
(Lt_GPA=[1.5,2)) ==> Failed	(s=7.1%, c=92.3%)

Contrast rules in Table 6 could belong to the first category as well; students with low transfer GPAs are more likely to fail CEM 141 (c_2/c_1=12). This rule has the advantage of actionability; so, when students with low transfer GPAs enroll for the course, the system could be designed to provide them with additional help.

5.3.2 Difference of Proportions

The focus of this measure is on comparing the rules ($B \Rightarrow A$ and $\overline{B} \Rightarrow A$). Contrast rules in Table 7 suggest that historically strong students that take long periods of time between their first (incorrect) solution attempt and subsequent attempts tend to be female. This rule could belong to the second category. We found this interesting contrast rules using the difference of confidences to discover the top significant rules for BS 111. Though the support of the rules is low, that is the result would be of an interesting rule with low-support.

Table 7. BS_111 data set, difference of proportion measure

Contrast Rules	Support & Confidence
Male ==> (Lt_GPA=[3.5,4] & Time>20_hours)	(s=0.1%, c=26.3%)
Female ==> (Lt_GPA=[3.5,4] & Time>20_hours)	(s=0.6%, c=89.7%)

5.3.3 Chi-square

It is a well-known condition in chi-square testing for contingency tables that cell expected values need to be above 5 to guarantee the veracity of the significance levels obtained [16]. We do pruning if this limitation is violated in some cases, and this usually happens when the expected support corresponding to f_{11} or f_{12} in Table 3 is low.

Table 8. CEM_141 data set, chi-square measure

Contrast Rules	Support & Confidence
(Lt_GPA=[3,3.5) & Sex=Male & Tries=1) ==> Passed	(s=4.4%, c=82.7%)
(Lt_GPA=[3,3.5) & Sex=Male & Tries=1) ==> Failed	(s=0.9%, c=17.3%)

Contrast rules in Tables 8 suggest that students with transfer GPAs in the range of 3.0 to 3.5 that were male

and answered homework problems on the first try were more likely to pass the class than to fail it. (c_1/c_2=4.8). This rule could belong to the second category. We found this rule using the chi-square measure for CEM 141.

Table 9. LBS_271 data set, difference of confidences measure

Contrast Rules	Support & Confidence
(DoDiff=medium & DoDisc=non_discriminating & Succ=YES & Tries=1) ==> Passed	(s=28.9%, c=94.1%)
(DoDiff=medium & DoDisc=non_discriminating & Succ=YES & Tries=1) ==> Failed	(s=1.8%, c=5.9%)

Contrast rules in Table 9 show more complicated rules for LBS 271 using difference of proportion (c_1/c_2=15.9); these rules belong to the third (unknown) category and further consultation with educational experts is necessary to determine whether or not they are interesting.

6. Conclusion

LON-CAPA servers are recording students' activities in large logs. We proposed a general formulation of interesting contrast rules and developed an algorithm to discover a set of contrast rules investigating three different statistical measures. This tool can help instructors to design courses more effectively, detect anomalies, inspire and direct further research, and help students use resources more efficiently. An advantage of this developing approach is its broad functionality in many data mining application domains. Specifically, it allows for contrast rule discovery with very low minimum support, therefore permitting the mining of possibly interesting rules that otherwise would go unnoticed.

More measurements tend to permit discovery of higher coverage rules. A combination of measurements should be employed to find out whether this approach for finding more interesting rules can be improved. In this vein, we plan to extend our work to analysis of other possible kinds of contrast rules.

7. Acknowledgment

This work was partially supported by the National Science Foundation under ITR 0085921.

8. References

[1] Kortemeyer, G., Bauer, W., Kashy, D.A., Kashy, E., and Speier, C., "The LearningOnline Network with CAPA Initiative", *Proceedings of IEEE conference on Frontiers in Education*, vol. 31,(2001) p. 1003. See also http://www.loncapa.org .

[2] Minaei-Bidgoli, B., Kashy, D.A., Kortemeyer, G., Punch, W.F., "Predicting Student Performance: An Application of Data Mining Methods with an educational Web-based System", *IEEE conference on Frontier In Education FIE 2003*, Nov. 2003 Boulder.

[3] Agrawal, R., Imielinski, T.; Swami A., "Mining Associations between Sets of Items in Massive Databases", *Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data*, Washington D.C., May 1993.

[4] Agrawal, R., Srikant, R. "Fast Algorithms for Mining Association Rules", *Proceeding of the 20th International Conference on Very Large Databases*, Santiago, Chile, September 1994.

[5] Tan, P.N., Kumar V., and Srivastava, J. "Selecting the Right Objective Measure for Association Analysis", *Information Systems*, 29(4), 293-313 (2004).

[6] Silberschatz A. and Tuzhilin, A., "On subjective measures of interestingness in knowledge discovery". *Proceeding of KDD*, 275-281, 1995.

[7] Piatetsky-Shapiro G. and Matheus. C. J., "The interestingness of deviations". In *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases*, pp. 25-36, 1994.

[8] Liu, B., Hsu, W., Mun, L.F. and Lee, H., "Finding Interesting Patterns Using User Expectations", *IEEE Transactions on Knowledge and Data Engineering*, Vol 11(6), pp. 817-832, 1999.

[9] Silberschatz A. and Tuzhilin, A. "What makes patterns interesting in Knowledge discovery systems". *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970-974, December, 1996.

[10] Fu Y. and Han, J., "Meta-rule-guided mining of association rules in relational databases". *Proc. 1995 Int'l Workshop. on Knowledge Discovery and Deductive and Object-Oriented Databases (KDOOD'95)*, Dec. 1995, pp. 39-46.

[11] Dong, G., Li, J., "Interestingness of Discovered Association Rules in terms of Neighborhood-Based Unexpectedness", *Proceedings of Pacific Asia Conference on Knowledge Discovery in Databases (PAKDD)*, pp. 72-86. Melbourne, 1998.

[12] Osmar R. Zaïane, "Web Usage Mining for a Better Web-Based Learning Environment", in Proc. of *Conference on Advanced Technology for Education*, pp 60-64, Banff, Alberta, June 27-28, 2001.

[13] Ma, Y., Liu, B., Kian C., Wong, Yu, P.S., and Lee, S.M., "Targeting the Right Students Using Data Mining". *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2000, Industry Track)*, Aug, 2000, Boston, USA

[14] Bay, S. D. and Pazzani, M. J., "Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, 2001, Vol 5, No 3 213-246.

[15] Liu, B., Hsu, W. and Ma, Y, "Identifying Non-Actionable Association Rules", *Proc. Of seventh ACM SIGKDD International conference on Knowledge Discovery and Data Mining(KDD-2001)* San Francisco, USA.

[16] Borgelt, C., "Efficient Implementations of Apriori and Eclat**",** *Workshop of Frequent Item Set Mining Implementations* (FIMI) 2003.

[17] A.A. Freitas, "On rule interestingness measures.", Knowledge-Based Systems journal 12 (5-6), 309-315. Oct. 1999.

[18] R.Meo, "Replacing Support in Association Rule Mining", Rapporto Tecnico RT70-2003, Dipartimento di Informatica, Università di Torino, April, 2003

[19] Agresti, A. *Categorical data analysis*. 2nd edition, New York: Wiley, 2002.