CHAPTER 8

Association analysis for a web-based educational system

B. Minaei-Bidgoli¹, P. Tan², G. Kortemeyer³ & W.F. Punch²

¹Computer Engineering Department, Iran University of Science and Technology, Iran.
²Computer Science & Engineering Department, Michigan State University, USA.
³Lyman Briggs School of Science, Michigan State University, USA.

Abstract

An important goal of data mining is to discover the unobvious relationships among the objects in a data set. 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 research 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. Our research provides a new algorithm for mining contrasting rules that can improve web-based educational systems for both teachers and students – allowing for greater learner improvement and more effective evaluation of the learning process. We apply this technique to an online educational system developed at Michigan State University called LON-CAPA. A larger advantage of developing this approach is its wide application in any other data mining application.
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 study 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 research 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 $2 \times 2$ contingency table as shown in Table 1.

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 shown in Fig. 1 (where $s$ and $c$ are the support and confidence

<table>
<thead>
<tr>
<th></th>
<th>Passed</th>
<th>Failed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review materials</td>
<td>95</td>
<td>14</td>
<td>109</td>
</tr>
<tr>
<td>Do not review</td>
<td>41</td>
<td>50</td>
<td>91</td>
</tr>
<tr>
<td>Total</td>
<td>136</td>
<td>64</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 1: A contingency table of student success vs. study habits for an online course.
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 Figs 2 and 3.

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

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?

We present a general formulation of contrast rules and propose a new algorithm for mining interesting contrast rules. The rest of this study 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.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review materials ⇒ Passed</td>
<td>s = 47.5%, c = 87.2%</td>
<td></td>
</tr>
<tr>
<td>Review materials ⇒ Failed</td>
<td>s = 7.0%, c = 12.8%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: A contrast rule extracted from Table 1.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passed ⇒ Review materials</td>
<td>s = 47.5%, c = 69.9%</td>
<td></td>
</tr>
<tr>
<td>Failed ⇒ Review materials</td>
<td>s = 7.0%, c = 15.4%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: A contrast rule extracted from Table 1.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passed ⇒ Review materials</td>
<td>s = 47.5%, c = 69.9%</td>
<td></td>
</tr>
<tr>
<td>Passed ⇒ Do not review</td>
<td>s = 20.5%, c = 30.1%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: A contrast rule extracted from Table 1.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not review ⇒ Passed</td>
<td>s = 20.5%, c = 45.1%</td>
<td></td>
</tr>
<tr>
<td>Do not review ⇒ Failed</td>
<td>s = 25.0%, c = 54.9%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: A contrast rule extracted from Table 1.
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 \( X \subseteq I, Y \subseteq I \), and \( X \) and \( Y \) are disjoint, that is, \( X \cap Y = \emptyset \). \( X \) 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 contains \( X \) also contains \( Y \). Formally, support is defined as shown in eqn (1),

\[
s(X \Rightarrow Y) = \frac{s(X \cup Y)}{N},
\]

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

\[
c(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}.
\]

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

\[
\text{RuleCoverage} = \frac{s(X)}{N}.
\]

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 \([5, 6]\). Silberschatz and Tuzhilin suggest that interestingness measures can be categorized into two classes: objective and subjective measures \([7]\).

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 \([8]\). 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 [7]. Different subjective measures have been addressed to discover the interestingness of a rule [7]. For example, a rule template [9] is a subjective technique that separates only those rules that match a given template. Another example is neighborhood-based interestingness [10], 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]. Zaïane [11] suggested the use of web mining techniques to build an agent that recommends online learning activities in a web-based course. Ma et al. [12] 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 [13] 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. [14] 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 eqn (5), where $1 \leq i \leq K$ and $1 \leq j \leq L$, and the degree of freedom is $(K-1)(L-1)$.
Table 2: A contingency table proportional to Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Passed</th>
<th>Failed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>40</td>
<td>49</td>
<td>89</td>
</tr>
<tr>
<td>Female</td>
<td>60</td>
<td>51</td>
<td>111</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>(b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>400</td>
<td>490</td>
<td>890</td>
</tr>
<tr>
<td>Female</td>
<td>600</td>
<td>510</td>
<td>1110</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
<td>1000</td>
<td>2000</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{4}
\]

However, a drawback of this test is that the \( \chi^2 \) value is not invariant under the row-column scaling property \[8\]. For example, consider the contingency table shown in Table 2(a). If \( \chi^2 \) 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 \( \chi^2 \) 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 \( \chi^2 \) is proportional to \( N \).

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 \( \times \) 2 contingency table as shown in Table 3.

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

We assume that \( B \) is a target variable and \( A \) is a conjunction of explanatory attributes. Let \( \mu \) 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 \[8\]. In Fig. 6, we provide a formal definition of 'contrast rule'.
Table 3: A contingency table for the binary case.

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>(\overline{B})</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(f_{11})</td>
<td>(f_{12})</td>
<td>(f_{1+})</td>
</tr>
<tr>
<td>(\overline{A})</td>
<td>(f_{21})</td>
<td>(f_{22})</td>
<td>(f_{2+})</td>
</tr>
<tr>
<td>Total</td>
<td>(f_{+1})</td>
<td>(f_{+2})</td>
<td>(N)</td>
</tr>
</tbody>
</table>

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

Definition (General Formulation of Contrast Rules):

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

1. \(br \subseteq \Omega\) is the base rule,
2. \(\nu(br) \subseteq \Omega\) is a neighborhood to which the base rule \(br\) is compared,
3. \(M = <m_{base}, m_{neighbor}>\) is an ordered pair of measures where \(m_{base}\) and \(m_{neighbor}\) measure the rules in \(br\) and \(\nu(br)\), respectively,
4. \(\Delta(m_{base}(br), m_{neighbor}(\nu(br)))\) is a comparison function between \(m_{base}(r)\) and \(m_{neighbor}(\nu(br))\).

A contrast rule, \(cr\), is interesting if and only if \(\Delta(m_{base}(br), m_{neighbor}(\nu(br))) \geq \sigma\), where \(\sigma\) 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 Fig. 6, the contrast rule definition is based on a paired set of rules, base rule \(br\) and its neighborhood \(\nu(br)\). 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.

3.1 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 Fig. 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:

- \(br\): \([A \Rightarrow B]\)
- \(\nu(r)\): \([A \Rightarrow \overline{B}]\)
- \(M\): \((\text{confidence}, \text{confidence})\)
- \(\Delta\): absolute difference.
The evaluation criterion for this example is shown in eqn (5). This criterion can be used for ranking different pairs of contrast rules.

\[
\Delta = |c(r) - c(\nu(r))| \\
= |c(A \Rightarrow B) - c(A \Rightarrow \overline{B})| \\
= \left| \frac{f_{11}}{f_{1+}} - \frac{f_{12}}{f_{1+}} \right| = \left| \frac{f_{11} - f_{12}}{f_{1+}} \right|, \tag{5}
\]

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,

\[
\Delta = |c(A \Rightarrow B) - c(A \Rightarrow \overline{B})| \\
= |2c(A \Rightarrow B) - 1| \\
\propto c(A \Rightarrow B).
\]

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

3.2 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 Fig. 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:

- \( br: \{B \Rightarrow A\} \)
- \( \nu(br): \{\overline{B} \Rightarrow A\} \)
- \( M: \langle \text{confidence, confidence} \rangle \)
- \( \Delta: \text{absolute difference}. \)

The evaluation criterion for this example is shown in eqn (6), where \( \Delta \) is defined as follows:

\[
\Delta = |c(r) - c(\nu(r))| \\
= |c(B \Rightarrow A) - c(\overline{B} \Rightarrow A)| \\
= \left| \frac{f_{11}}{f_{1+}} - \frac{f_{12}}{f_{1+}} \right| = \left| \rho(A \Rightarrow B) - \rho(A \Rightarrow \overline{B}) \right|, \tag{6}
\]

where \( \rho \) is the rule proportion \([15]\) and is defined in eqn (7)

\[
\rho(A \Rightarrow B) = \frac{P(AB)}{P(B)} = c(B \Rightarrow A). \tag{7}
\]
3.3 Example 3: cr₃ (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:

\[
corr = \frac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{f_{1+}f_{2+}f_{+1}f_{+2}}}
\]  

(8)

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

- \( br \) is \( \{ A \Rightarrow B, B \Rightarrow A, A \Rightarrow \bar{B}, B \Rightarrow \bar{A} \} \)
- \( v(br) \) is \( \{ A \Rightarrow \bar{B}, B \Rightarrow A, A \Rightarrow B, B \Rightarrow \bar{A} \} \)
- \( M \) : \( \langle \text{confidence, confidence} \rangle \)
- \( \Delta \): the difference in the square root of confidence products [see eqn (9)]

\[
\Delta = \sqrt{c_1c_2c_3c_4} - \sqrt{c_5c_6c_7c_8},
\]

(9)

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

\[
\Delta = \sqrt{P(AB)P(\bar{A}\bar{B})P(\bar{A}\bar{B})} - \sqrt{P(\bar{A}\bar{B})P(\bar{A}\bar{B})P(\bar{A}\bar{B})} - \sqrt{P(A\bar{B})P(A\bar{B})P(A\bar{B})} + \sqrt{P(A\bar{B})P(A\bar{B})P(A\bar{B})}.
\]

(10)

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

\[
corr = \sqrt{\frac{\chi^2}{N}}.
\]

(12)

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

3.4 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 shown in Fig. 7.

In order to employ the MCR algorithm, several steps must be taken. During the pre-processing 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

---

**Mining Contrast Rules (MCR) Algorithm:**

| Input: $D$ – Input set of $N$ transactions |
| $B$ – Target variable, the basis of interesting contrasts |
| $\sigma$ – 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 $\psi(br)$ pair compute difference in measures, $\Delta$

Sort the rules with respect to $\Delta$

Select top $k$ rules

return $R$

---

Figure 7: MCR algorithm for discovering interesting candidate rules.
values, we divide the data set, $D$, into $C$ disjoint subsets based on the elements of the target variable, as shown in Fig. 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 (the code for this program is available at http://fuzzy.cs.uni-magdeburg.de/~borgelt/apriori.html) 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, $\sigma$; the lower the $\sigma$, 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

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

1. **Student attributes**: These 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.
• \textit{LtGPA} [Level Transferred (i.e. High School) GPA]: measured the same as GPA.

2. \textit{Problem attributes}: These are fixed for any problem. Among several attributes for the problems we selected the four following attributes:

• \textit{DoDiff} (degree of difficulty): This is a useful factor for an instructor to determine whether a problem has an appropriate level of difficulty. \textit{DoDiff} is computed by the total number of students’ submissions divided by the number of students who solved the problem correctly. Thus, \textit{DoDiff} is a continuous variable in the interval \([0, 1]\) which is discretized into terciles of roughly equal frequency: easy, medium, and hard.

• \textit{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 \textit{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.
Association Analysis for a Web-Based Educational System

(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.

3. **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.

4. **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 study 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 750 MB, corresponding to more than 190k transactions of students attempting to solve homework problems.

For this research 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

<table>
<thead>
<tr>
<th>Data set</th>
<th>Course title</th>
<th>No. of students</th>
<th>No. of problems</th>
<th>Size of activity log (MB)</th>
<th>No. of Transactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBS 271</td>
<td>Physics_I</td>
<td>200</td>
<td>174</td>
<td>152.1</td>
<td>32,394</td>
</tr>
<tr>
<td>BS 111</td>
<td>Biological Science</td>
<td>382</td>
<td>235</td>
<td>239.4</td>
<td>71,675</td>
</tr>
<tr>
<td>CEM141</td>
<td>Chemistry_I</td>
<td>2048</td>
<td>114</td>
<td>754.8</td>
<td>190,859</td>
</tr>
</tbody>
</table>
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 study, 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 unexpect-
atedness [6, 17–19, 20].

One of the techniques for mining interesting association rules based on unex-
pectedness. 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 unanti-
cipated 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 instruc-
tors 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: \textit{Succ} stands for success per student per problem,
\textit{LtGPA} stands for transfer GPA, \textit{DoDiff} stands for degree of difficulty of a particular
problem, and \textit{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
\((A \Rightarrow B \text{ and } A \Rightarrow \overline{B})\). 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.

<table>
<thead>
<tr>
<th>Contrast rules</th>
<th>Support and confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Succ = YES) ⇒ Passed</td>
<td>(s = 86.1%, c = 92.7%)</td>
</tr>
<tr>
<td>(Succ = YES) ⇒ Failed</td>
<td>(s = 6.8%, c = 7.3%)</td>
</tr>
</tbody>
</table>

Table 6: CEM_141 data set, difference of confidences measure.

<table>
<thead>
<tr>
<th>Contrast rules</th>
<th>Support and confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Lt_GPA = [1.5,2)) ⇒ Passed</td>
<td>(s = 0.6%, c = 7.7%)</td>
</tr>
<tr>
<td>(Lt_GPA = [1.5,2)) ⇒ Failed</td>
<td>(s = 7.1%, c = 92.3%)</td>
</tr>
</tbody>
</table>

Table 7: BS_111 data set, difference of proportion measure.

<table>
<thead>
<tr>
<th>Contrast rules</th>
<th>Support and confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male ⇒ (Lt_GPA = [3.5,4] &amp; Time &gt; 20_hours)</td>
<td>(s = 0.1%, c = 26.3%)</td>
</tr>
<tr>
<td>Female ⇒ (Lt_GPA = [3.5,4] &amp; Time &gt; 20_hours)</td>
<td>(s = 0.6%, c = 89.7%)</td>
</tr>
</tbody>
</table>

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.

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 \(\left(\frac{c_2}{c_1} = 12\right)\). 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 \(B ⇒ A\) and \(B ⇒ A\). Contrast rules in Table 7 suggest that historically strong students who 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 rule 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 for an interesting rule with low support.
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.

Contrast rules in Table 8 suggest that students with transfer GPAs in the range 3.0–3.5 who were male and answered homework problems on the first try were more likely to pass the class than to fail ($c_1/c_2 = 4.8$). This rule could belong to the second category. We found this rule using the chi-square measure for CEM 141.

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.

Acknowledgments

We thank the National Science Foundation for grants supporting this work through the Information Technology Research (ITR 0085921) and the Assessment of Student Achievement (ASA 0243126) programs. During the earlier years of this project support was also received from the Alfred P Sloan and Andrew W Mellon Foundations. We are grateful to our own institution, Michigan State University and to its administrators for over a decade of encouragement and support.

References


156 Data Mining in E-Learning


