Finding Dependent Test Items: An Information Theory Based Approach

Xiaoxun Sun
Australian Council for Educational Research
xiaoxun.sun@gmail.com

ABSTRACT

In this paper, we propose a new approach to find the most dependent test items in students' response data by adopting the concept of entropy from information theory. We define a distance metric to measures the amount of mutual independency between two items, and it is used to quantify how independent two items are in a test. Based on the proposed measurement, we present a simple yet efficient algorithm to find the best dependency tree from the students' response data, which shows the hierarchical relationship between test items. The extensive experimental study has been performed on synthetic datasets, and results show that the proposed algorithm for finding the best dependency tree is fast and scalable, and the comparison with item correlations has been made to confirm the effectiveness of the approach. Finally, we discuss the possible extension of the method to find dependent item sets and to determine dimensions and sub-dimensions from the data.

1. INTRODUCTION

Data mining is the analysis step of the knowledge discovery in databases process, and it is the process of discovering novel and potentially useful information and patterns from large data sets. There are different data mining technologies lying at the intersection of artificial intelligence, machine learning, statistics and database systems. The goal of data mining is to extract useful and previously unknown information out of large complex data collections. Data mining techniques have been applied to many other fields. In the context of educational research, educational data mining refers to developing methods for exploring the unique types of data that come from educational settings, and using existing data mining or developing new methods to better diagnose students' performance and design tests that better suits students.

Students' response data contain the responses of students to a set of test questions. It can be used to determine the knowledge of a student has learned, and it can also be used to discover the relationship between the test items latent or underlying attributes. Such relationship may take the form of attempting to find out which variables are most strongly associated with a single variable of particular interest, or may take the form of attempting to discover which relationships between any two variables are strongest. Students' response data are beneficial to both test developers and course instructors. Students' response data contains valuable information that can be used to improve the effectiveness of test items, and for course instructors, students' performance on the test is importance to instructors for the guidance and improvement of teaching.

2. INFORMATION THEORY BASED METHOD

In the information theory, the main concept is entropy. It is defined to measure the expected uncertainty or the amount of information provided by a certain event. We feel more surprised when an unlikely event happens than a likely one occurs. One useful measure of the extent of surprise of an event is to use the logistic function. Suppose the probability of an event happening is p, then the extent of surprise of such event can be defined as $-\log_k p$, in which k refers to the base of the logistic function. From this definition, it can be seen that the less the probability is, the higher the amount of information the event would provide. Given the example of students' response data, items that have been answered correctly by a small portion of students contains much more useful information for course instructors than the items that have been answered fully correct.

We adopt the conditional entropy to measure the mutual information, which is a distance metric.

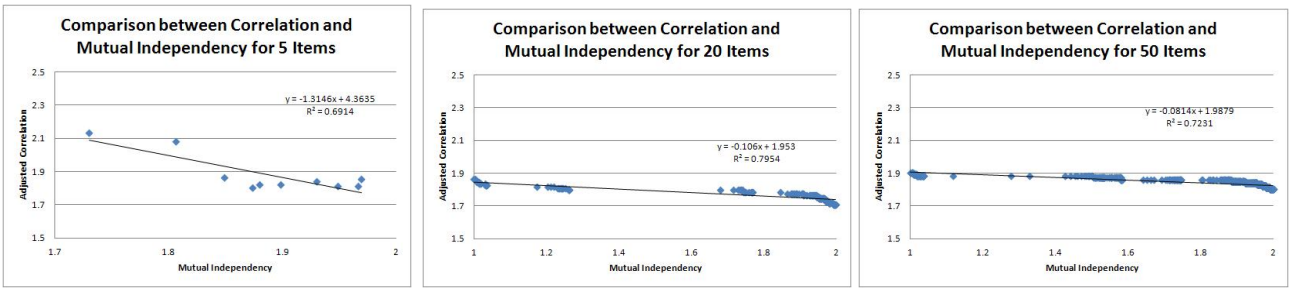
Definition 1 (Mutual Information Measure). *The mutual information measure with regard to two random variables A and B is defined as:*

$$MI(A, B) = H(A|B) + H(B|A) \quad (1)$$

Mutual information measure is a measure of how independent are the two random variables when the value of each random variable is known. Two events A and B are independent if and only if their mutual information measure achieves the maximum $H(A) + H(B)$. Therefore, the less the value of the mutual information measure is, the more dependent the two random variables are. According to this measure, A is said to be more dependent on B than C, if $MI(A, B) \leq MI(A, C)$.

2.1 Finding the Best Dependency Tree

Dependency tree was introduced by Chow and Liu [1] and it has been used in finding dependency structure in the features which improve the classification accuracy of the Bayes network classifiers [3]. [2] used the dependency tree to represent a set of frequent patterns, which can be used to summarize patterns into few profiles. [4] presented a large node



(a) (b) (c)

Figure 1: Comparisons between mutual independency measure and the correlations

dependency tree, in which the nodes are subsets of variables of dataset. The large node dependency tree is applied to density estimation and classification.

2.2 Extensions

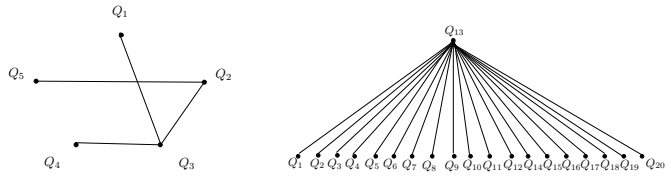
The method described in the paper can be easily extended to be capable of handling two item sets, each of which consists of different items. This extension is useful in the sense that it could provide the dependency relationship in a higher level and the extended method is able to generate the dependency between different sub-strands, which makes entropy-based dependency method superior to the traditional correlation method. The generated best dependency tree could be used to determine dimensions and sub-dimensions of the data. This can be done by summarizing patterns from the best dependency tree.

3. SIMULATION RESULTS

In this study, we compare the strength of the correlation against the mutual information measure between two variables. In general, the more independent two variables are, the more related two variables should be. In this sets of experiments, we graph the relationship between mutual independency measure and the correlation of two test items.

Figure 1 shows the relationship between the mutual independency measure and the correlations. Figure 1(a) plots the relationship among items in a short test, Figure 1(b) plots the relationship among items in a medium test, and Figure 1(c) plots the relationship among items in a long test. From Figure 1(a), since 5 items are included in the simulated data, the number of 2-item combinations are $C_5^2 = \frac{5x4}{2} = 10$, for Figure 1(b), 20 items will produce $C_{20}^2 = \frac{20x19}{2} = 190$ different combinations of 2-item sets, and for Figure 1(c), there are $C_{50}^2 = \frac{50x49}{2} = 1225$ combinations for 50 items. From all figures, it can be seen that the slopes of the regression line are negative, which confirms the fact that the more the mutual independency between two variables, the less correlated they are. The R^2 of the regression line is the indication of how strong the linear relationship is. In all cases, the values of R^2 are greater than 0.65, and in Figure 1(a), the R^2 indicates a strong linear relationship, while that relationship is stronger in Figure 1(b) and Figure 1(c).

Figure 2 displays the best dependency tree structure calculated from the simulated data1 and data2. There are two patterns {$Q1$, $Q3$, $Q4$}, {$Q2$, $Q5$}) observed from Figure 2(a) and in Figure 2(b), all the questions are highly dependent on $Q13$.



(a).Best dependency tree constructed from data1

(b).Best dependency tree constructed from data2

Figure 2: The dependency tree structures of two tests

4. CONCLUSIONS AND FUTURE WORK

In this paper, we apply the concept of entropy to propose a distance metric to evaluate the amount of mutual information among records in students' response data, and propose a method of constructing dependency tree from the data. The experimental results confirm the effectiveness and efficiency of the proposed method.

There are some potential work on the research agenda. First, the information theory based method presented in this paper finds the dependent item pairs, and it can be extended to calculate the dependency between item sets. Second, the simulation results conducted in this paper are on synthetic data, and applying to real students' response data is necessary.

5. REFERENCES

[1] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[2] B. Cui, Y. Li, and Z. Zhang. Summarizing frequent patterns using profiles. *In Database Systems for Advanced Applications, 11th International Conference, DASFAA*, 2006.

[3] N. Friedman, D. Geiger, and M. Goldszmid. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[4] K. Huang, I. King, and M. Lyu. Constructing a large node chow-liu tree based on frequent itemsets. *In Proceedings of the International Conference on Neural Information Processing*, 2002.