A Clustering Method using Entropy for Grouping Students

Byoung Wook KIM[a*], Jon MASON[b] & Jin Gon SHON[c*]
[a]*Dept. of Computer Science Education, Korea University, Republic of Korea*
[b]*International Graduate Centre of Education, Charles Darwin University, Australia*
[c]*Dept. of Computer Science, Korea National Open University, Republic of Korea*
*jgshon@knou.ac.kr

Abstract: This study suggests a novel clustering method using entropy in information theory for setting cut-scores. Based on item response vectors from the examinees, we construct the Ordered Item Booklets (OIBs) based on the Rasch model which is a kind of Item Response Theory (IRT). The approach of the proposed method is to partition the scores into n-clusters and to construct probability distribution tables separately for each cluster from the item response vector. Using these probability distribution tables, mutual information and relative entropy (Kullback-leibler divergence) were computed between each of the clusters and then cut-scores were determined by the cluster's partition to minimize mutual information values. Experimental results show that the approach of this proposed entropy method has a realistic possibility of application as a clustering evaluation method.

Keywords: entropy, clustering, item response data, test score

1. Introduction

This short paper reports on a method involving a specific set of mathematical operations on performance data created by examinees. Specifically, item test scores are analysed and ordered in terms of difficulty (as perceived by examinees) to determine abilities of the examinees. It also makes use of entropy, or disorder, within the data collected. Within computer science and using standardized evaluation criteria educational evaluation can be broadly classified as either a criterion-referenced evaluation or a norm-referenced evaluation. A norm-referenced evaluation (or test) is a type of evaluation method that produces an estimate of a ranked position of a tested individual within a defined population, with respect to a particular trait being measured. This estimate is derived from by combining an analysis of test scores with other relevant data from a sample drawn from the population. That is, this type of test identifies whether the test taker has performed better or worse than other test takers, but not whether the test taker actually knows more or less material than is necessary for the given purpose.

A criterion-referenced test is one that provides a translation of test scores into a statement about the behavior to be expected of a person with that score or their relationship to a specified subject matter. A criterion-referenced assessment can be contrasted with norm-referenced assessment and both have advantages and disadvantages. Typically, it is the context and purpose of evaluation that determines the choice of which method is used.

Recently, it has become more important to classify students according to their test scores in school for instruction with a variable curriculum, usually by achievement level enabling selection of excellent students. A great deal of research across numerous courses is currently being carried out to teach students appropriately for their ability and level. The main drawback, however, in classifying students according to their test scores is a lack of reasonable criteria that reflects ability.

This study presents a clustering method in order to split the class into groups according to similarity of individual student scores. An instance of a clustering problem consists of a set of objects and a set of properties for each object into groups using only the characteristic vectors.

2. Analysis of Item Response Data

Research relevant to the proposed method uses analysis of item response data. Kim et al. (2005) proposed information-based pruning for identifying interesting association rules in item response data through data mining techniques. Park et al. (2002) applied goodness to classify scores in item response using relaxation error which is one type of clustering method. But, their proposed algorithm is a type of greedy algorithm, thus creating a local minimum problem. Richard (1979) applied goodness to classify scores in item response also using relaxation error as a clustering method. In general, only certain aspects of the characteristic vectors will be relevant, and extracting these relevant features is one field where clustering plays a major role (Tishby, 1999). Clustering is a fundamental tool in unsupervised learning that is used to group together similar objects (Jain and Dube, 1988), and has practical importance in a wide variety of applications such as text, web-log and market-basket data analysis. The crucial point of all clustering algorithms is the choice of a proximity measure.

3. Clustering Evaluation Methods

After a set of cluster is found, we need to assess the goodness of the clusters. Unlike classification, where it is easy to measure accuracy using labeled test data, for clustering nobody knows what the correct clusters are given.

3.1 Sum of Squared Error

SSE is the sum of the squared differences between each observation and its group's mean. It can be used as a measure of variation within a cluster.

$$SSE = \sum_{j=1}^{k} \sum_{x \in C_j} disk(x, m_j)^2 \qquad (1)$$

k is the number of required clusters, is the *j*th cluster, is the centroid of cluster (the mean vector of all the data points in), and is the distance between data point x and centroid.

In Euclidean space, the mean of a cluster is computed with:

$$m_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \qquad (2)$$

where $|C_j|$ is the number of data points in cluster C_j. The distance from a data point to a cluster mean is computed with Eq.(3).

$$dist(x_i, m_j) = \| x_i, m_j \| = \sqrt{(x_{i1} - m_{j1})^2 + \cdots + (x_{ir} - m_{jr})^2} \qquad (3)$$

3.2 Relaxation Error

Relaxation Error is a method that measure ‘goodness’ of conceptual clustering.

$$RE(C) = \sum_{i=1}^{n} \sum_{j=1}^{n} P(x_i) P(x_j) | x_i - x_j | \qquad (4)$$

where x_i, x_j are property values and $P(x_i)$, $P(x_j)$ are probability of in cluster *C*.

RE(*C*) is error value calculated between each cluster, this error is high means a goodness of clustering is low. If Partitions of C is $P = \{C_1, C_2, \cdots, C_n\}$, *RE*(*P*) is calculated by Eq.(5).

3.3 Entropy

Entropy measures the amount of impurity or disorder in the data. For each cluster, we can measure its entropy as follows:

$$entropy(D_i) = -\sum_{j=1}^{k} P_i(c_j)\log_2 P_i(c_j) \qquad (6)$$

where $P_i(c_j)$ is the probability of class c_j in data set D_i. The total entropy of the whole clustering (which considers all clusters) is:

$$entropy_{total}(D) = -\sum_{i=1}^{k} \frac{|D_i|}{|D|} \times entropy(D_i) \qquad (7)$$

4. Experimental Classification Results and Analysis

In general, item response data such as that represented in Table 1 can be obtained as an outcome through one test. In the item response data, each transaction (i.e., student) has its own score and is considered differently, with '1' indicating a correct answer and the '0' indicating a wrong one.

Table 1: Item response data

Student	I1	I2	I3	…	In	In	score
S1	0	1	0	…	1	1	TS1
S2	1	1	0	…	1	1	TS2
S3	1	0	1	…	0	0	TS3
…	…	…	…	…	…	…	…
S_{m-1}	1	0	0	…	1	1	TS_{m-1}
S_m	0	1	1	…	0	0	TS_m

For our experiments, we used a sample dataset for 13611 students which is an item response data of ICT literacy Test carried out in 2007. Figure 1 shows score frequent analysis in sample data. As shown the Figure 1, test score is a continuous numeric attribute.

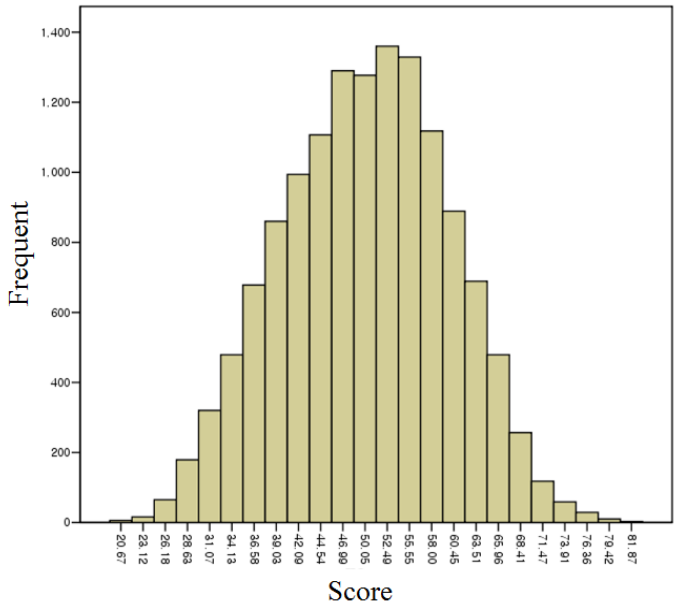

Figure 1. Score frequent in sample data

SSE and Relaxation Error uses one variable is score ($TS_{1,\dots,}TS_m$) in item response data.

Table 2: Confusion matrix with each method with instance No.153

Cluster	SSE	Relaxation error
1	68114	210.92
2	92569	
3	68944	
total	229627	210.92

But Entropy required two variables for calculating.

Table 3: Confusion matrix with each method on Item No.1 with instance No.153

Cluster	Correct	Wrong	entropy
1	483	3114	0.150
2	1354	5009	0.349
3	1309	2342	0.253
total	3146	10465	0.752

Table 4 shows a result of clustering evaluation entropy, SSE and RE.

Table 4: Top 20 Rank Results of Clustering Evaluation (# is a instance No.)

#	entropy	#	SSE	#	RE
153	18.19	152	215015	140	187.52
152	18.20	138	220811	154	187.88
138	18.20	153	229627	141	194.44
166	18.20	166	244336	125	198.09
167	18.21	165	246347	126	199.69
154	18.23	139	256420	155	199.79
139	18.23	167	257092	167	200.41
137	18.24	154	258582	153	210.92
165	18.24	123	259733	139	214.22
151	18.26	137	265565	168	215.82
123	18.27	151	273607	166	216.33
179	18.27	168	277523	142	217.93
122	18.27	179	278272	127	219.62
168	18.27	178	281492	110	220.55
178	18.29	180	283124	109	222.10
180	18.29	122	287683	156	225.92
140	18.31	155	293991	124	226.39
155	18.32	177	297593	111	236.27
124	18.33	107	304079	169	243.22
106	18.34	140	304329	179	245.66

As shown in the Table 4, the entropy method is more similar to the SSE than the RE method. This result means that entropy is suitable for clustering continuative numeric attribute than RE.

5. Conclusion

We have developed a method for clustering item response data using entropy within a continuative numeric attribute as test score. In order to prove the effectiveness of this method, we have compared Sum of Squared Error (SSE) and Relaxation Error (RE). Experimental results, as shown, highlight more similarity with SSE rather than RE. Building on this approach we can verify a result's similarity of entropy and SSE statistically in future work.

Acknowledgements

We would like to thank all the people who provided comments on successive versions of this document.

References

Jain, A. K. & Dubes, R. C. (1988). Algorithms for Clustering Data. Englewood Cliffs, NJ, 1988: Prentice-Hall.

Kim, Hyeoncheol, Kwak, Eun-Young. (2005). Information-Based Pruning for Interesting Association Rule Mining in the Item Response Dataset, *LNCS 3681, Knowledge-Based Intelligent Information and Engineering Systems*. Berlin / Heidelberg: Springer pp. 372-378.

Park, EunJin, Chung Hong, Jang DukSung. (2002). A Grading Method for Student's Achievements Based on the Clustering Technique. *Korean Institute of Intelligent Systems, 12*(2), 151-156.

Tishby, N., Pereira, F., and Bialek, W. (1999). The information bottleneck method, In 37th *Annual Allerton Conference on Communication, Control and Computing*, pp. 368.

Vawter, R. (1979). Entropy state of a multiple choice examination and the evaluation of understanding, *American Journal of Physics, 47*(4), 320-324.