Machine Learning, 45, 219–228, 2001

Reinterpreting the Category Utility Function

BORIS MIRKIN mirkin@dcs.bbk.ac.uk
School of Computer Science and Information Systems, Birkbeck College, Malet Street, London, WC1E 7HX, UK; Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), Rutgers University, Piscataway, NJ 08854-8018, USA

Editor: Douglas Fisher

Abstract. The category utility function is a partition quality scoring function applied in some clustering programs of machine learning. We reinterpret this function in terms of the data variance explained by a clustering, or, equivalently, in terms of the square-error classical clustering criterion that administers the K-Means and Ward methods. This analysis suggests extensions of the scoring function to situations with differently standardized and mixed scale data.

Keywords: clustering, data standardization, contingency coefficient, correlation ratio, weighting features, mixed-scale data

1. Introduction

The category utility function introduced by Gluck and Corter (1985) has been applied in Cobweb (Fisher, 1987), a tool for incremental clustering with categorical features, and related systems. The original framework has been expanded to both nonincremental clustering and mixed scale data (e.g., Gennari, 1990; Reich & Fenves, 1991; Devaney & Ram, 1997).

The category utility function is defined in terms of the bivariate distributions of a clustering and each of the features, which is outwardly different from more traditional clustering criteria adhering to similarities and dissimilarities between instances. This paper shows that, however, the category utility function is equivalent, up to a denominator, to the square-error criterion in traditional clustering, when a standard encoding of categories is applied. More generally, the paper illustrates that a well-known, so-called "conceptual clustering" approach is intrinsically related to the "classical clustering" paradigm. As an intermediate step, we show in the next section that the category utility function is firmly related to some conventional statistical association measures for cross-classifications.

The framework is useful in deriving a theory-driven extension of the category utility function to the case of mixed scale data.

2. The category utility function

Consider a partition, $C = \{C_k\}$ $(k = 1, \ldots, n)$, found by a clustering algorithm based on given attributes A_i $(i = 1, \ldots, m)$. The attributes are assumed nominal so that each A_i has a set of attribute values or categories, $\{V_{ij}\}$. The category utility function scores partition C

against the set of variables according to formula:

$$CU(C) = \frac{1}{n}\sum_{k=1}^{n} P(C_k)\left[\sum_i \sum_j P(A_i = V_{ij} \mid C_k)^2 - \sum_i \sum_j P(A_i = V_{ij})^2\right] \quad (1)$$

The term in square brackets is the increase in the expected number of attribute values that can be predicted given a class, C_k, over the expected number of attribute values that could be predicted without using the class; the prediction strategy assumed follows a probability-matching approach. The term $P(C_k)$ weights the classes according to their sizes, and the division by n takes into account the difference in partition sizes.

The category utility function is closely related to a statistical contingency measure of decrease in the proportion of incorrect predictions introduced by Goodman and Kruskal (1954), following a suggestion by Wallis:

$$\Delta(C, A_i) = \sum_{k=1}^{n} P(C_k) \sum_j P(A_i = V_{ij} \mid C_k)^2 - \sum_j P(A_i = V_{ij})^2 \quad (2)$$

The category utility function is obviously the averaged sum of the delta coefficients:

$$CU(C) = \frac{1}{n}\sum_i \Delta(C, A_i). \quad (3)$$

Gennari (1990) and Fisher et al. (1993) have considered exactly the same part of the total category utility score as $\Delta(C, A_i)$, to measure the relative "salience" of feature A_i.

In the case when A_i forms a partition on the set of instances, the Goodman-Kruskal-Wallis index can also be expressed as a "goodness-of-fit" of the bivariate distribution to the statistical independence (Goodman & Kruskal, 1954):

$$\Delta(C, A_i) = \sum_{k=1}^{n} \sum_j \frac{[P(A_i = V_{ij} \,\&\, C_k) - P(A_i = V_{ij})P(C_k)]^2}{P(C_k)} \quad (4)$$

Both formulations, (2) and (4), can be reformulated in terms of contingency tables between C and each of the attributes. The contingency table, P^i, for C and A_i has rows corresponding to concepts (classes) of partition C and columns to values (categories) of attribute A_i; its entries are $p^i_{kj} = P(A_i = V_{ij} \,\&\, C_k)$ while $P(C_k)$ and $P(A_i = V_{ij})$ are, respectively, sums of its rows and columns denoted as $p_{k+} = \sum_j p^i_{kj}$ and $p^i_{+j} = \sum_k p^i_{kj}$ (under the assumption of mutual exclusivity of all concepts and all categories). Using these notations, (2) and (4) are:

$$\Delta(C, A_i) = \sum_{k=1}^{n} \sum_j \frac{\left(p^i_{kj}\right)^2}{p_{k+}} - \sum_j \left(p^i_{+j}\right)^2 = \sum_{k=1}^{n} \sum_j \frac{\left(p^i_{kj} - p_{k+}p^i_{+j}\right)^2}{p_{k+}} \quad (5)$$

Table 1. Segmented numerals presented with seven binary variables corresponding to presence/absence of the corresponding segment in figure 1.

Digit	e1	e2	e3	e4	e5	e6	e7
1	0	0	1	0	0	1	0
2	1	0	1	1	1	0	1
3	1	0	1	1	0	1	1
4	0	1	1	1	0	1	0
5	1	1	0	1	0	1	1
6	1	1	0	1	1	1	1
7	1	0	1	0	0	1	0
8	1	1	1	1	1	1	1
9	1	1	1	1	0	1	1
0	1	1	1	0	1	1	1

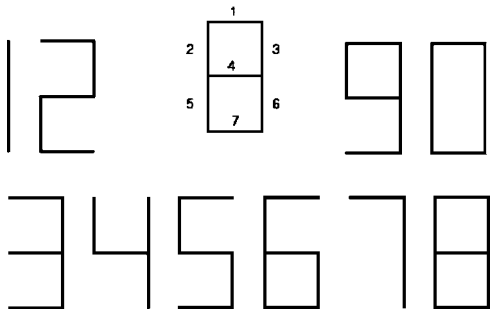

Figure 1. Integer digits presented by segments of the rectangle.

Example. Let us consider data in Table 1 referring to ten numeral digits according to Figure 1, so that attributes show presence/absence of the seven segments in the drawn digits.

Let the partition $C = \{C_1, C_2\}$ be given by attribute $e7$ with $C_1 =$ "$e7$ is present" covering 2, 3, 5, 6, 8, 9, and 0, and $C_2 =$ "$e7$ is absent" covering 1, 4, and 7.

The cross-classification of C and $e2$ in Table 2 yields $\Delta(C, e2) = 0.06$. Similar calculations for the other six attributes, e1, e3, ..., e7, lead to the total $\sum_{i=1}^{7} \Delta(C, ei) = 0.963$ and, thus, to $CU(C) = 0.481$ according to (3) with $n = 2$.

Table 2. Cross-tabulation of C against e1.

C	e1 = 1	e1 = 0	Total
C_1	5	2	7
C_2	1	2	3
Total	6	4	10

3. The square-error clustering criterion

Let us denote the set of instances by H and the set of features by L. In classical clustering, the data are presented as an instance-to-feature matrix $X = (x_{hl})$, $h \in H$, $l \in L$, where rows $x_h = (x_{hl})$ correspond to entities (instances) $h \in H$ and their components x_{hl} are corresponding values of features $l \in L$.

Such an instance-to-feature matrix is traditionally preprocessed into $Y = (y_{hl})$ where

$$y_{hl} = \frac{x_{hl} - a_l}{b_l}, \quad h \in H, \quad l \in L. \tag{6}$$

Here, a_l is the mean of column l; the scaling parameter b_l can be either the standard deviation or other user-specified value (see discussion in Section 5).

Assume that the clustering structure for a preprocessed data set $Y = (y_{hl})$, $h \in H$, $l \in L$, is a partition of H into n nonoverlapping clusters, C_k, as above. In classical clustering, each cluster C_k is supplied with its intensional description, the "standard" point $c_k = (c_{kl})$, which is a vector of the feature means within cluster C_k: $c_{kl} = \sum_{h \in C_k} y_{hl}/|C_k|$ $(k = 1, \ldots, n)$.

It is well known (e.g., Jain & Dubes, 1988, p. 95; Mirkin, 1996, p. 301) that the following equation holds for any clustering $(C, c) = \{C_k, c_k \mid k = 1, \ldots, n\}$:

$$\sum_{h \in H} \sum_{l \in L} y_{hl}^2 = \sum_{k=1}^{n} \sum_{l \in L} c_{kl}^2 |C_k| + \sum_{k=1}^{n} \sum_{h \in C_k} \sum_{l \in L} (y_{hl} - c_{kl})^2 \tag{7}$$

This equation decomposes the data scatter on the left, which is constant, into the explained and unexplained parts. This applies also to the total data variance which is just the lefthand term in (7) divided by $|H|$. The unexplained (or within-group) variance on the right in (7) is the well-known square-error classical clustering criterion (Jain & Dubes, 1988) to be minimized with respect to the clustering, (C, c), that is sought. The square-error criterion is nothing but the sum of the Euclidean distances squared between row-vectors and corresponding standard points.

We are interested, primarily, in the explained part of the scatter (variance),

$$F(C, c) = \sum_{k=1}^{n} \sum_{l \in L} c_{kl}^2 |C_k| \tag{8}$$

that is to be maximized over all clusterings (C, c) to minimize the square-error criterion.

4. The square-error criterion adjusted to categorical data

To adjust the case of qualitative attributes A_i to the classical clustering approach, each of the categories V_{ij} is formatted as a binary feature represented by a column l. Identifying a column l with corresponding V_{ij}, the column's elements are routinely coded as $x_{h,ij} = 1$, when h falls under V_{ij}, or $x_{h,ij} = 0$, otherwise, for all instances $h \in H$.

The mean of the binary feature corresponding to category V_{ij} (i.e.,to column $l = V_{ij}$) is equal to p^i_{+j} in the denotations of Section 2. Assuming the scaling parameter $b_{ij} = 1$ leads to the preprocessing formula (6) specified as

$$y_{h,ij} = x_{h,ij} - p^i_{+j}, \quad h \in H, \quad V_{ij} \in L. \tag{9}$$

It is not difficult to prove now that, for any preprocessed column $y_{ij} = (y_{h,ij})$ and for any class C_k in the clustering (C, c), its within-class mean (i.e., the standard point c_k's ij-th component) is equal to

$$c_{k,ij} = \frac{p^i_{kj}}{p_{k+}} - p^i_{+j} \tag{10}$$

By substituting expressions (10) for c_{kl} in (8) (with $l = V_{ij}$), the explained part of the variance becomes:

$$F(C, c) = |H| \sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j} \left(\frac{p^i_{kj}}{p_{k+}} - p^i_{+j} \right)^2 p_{k+} = |H| \sum_{i=1}^{m} \Delta(C, A_i) \tag{11}$$

because of (5). Taking into account (3), this proves the following.

Statement 1. *Under the data standardization specified above, the explained variance* $F(C, c)$ *in* (8) *is proportional to the sum of Goodman-Kruskal-Wallis coefficients between clustering* C *and attributes* A_i, *that is,* $nCU(C)$.

Formula (11) provides one more meaningful reformulation of the category utility function in terms of frequencies.

Example. The data matrix from Table 1 after preprocessing its columns is in Table 3. However, it is not exactly the matrix Y above because both Y and X must have 14 columns corresponding to each of the 14 categories reflected in Table 1. Columns corresponding to the category "ei is absent" in all features $i = 1, 2, \ldots, 7$ are not included in Table 3, because they provide no additional information.

The data scatter of this matrix is the summary column variance times $|H| = 10$, which is 13.1. However, to get the data scatter in the lefthand side of (7), this must be doubled to 26.2 to reflect the "missing half" of the data matrix Y.

The part of the data scatter taken into account by the partition C is the total of $\Delta(C, ei)$ over $i = 1, \ldots, 7$ times $|H| = 10$, according to (11), that is, 9.63 or 36.7% of 26.2.

The statement means that maximizing $F(C, c)/n$ and maximizing $CU(C)$ are equivalent.

This implies that when the number of clusters, n, is prespecified, the square-error clustering criterion is equivalent to the category utility criterion. However, in conceptual clustering the number of clusters is not prespecified, which shows the difference between these two criteria. Even with the square-error clustering criterion modified by relating to n, the criteria

Table 3. Data in Table 1 1/0 coded with the follow-up centering of the columns.

e1	e2	e3	e4	e5	e6	e7
−.8	−.6	.2	−.7	−.4	.1	−.7
.2	−.6	.2	.3	.6	−.9	.3
.2	−.6	.2	.3	−.4	.1	.3
−.8	.4	.2	.3	−.4	.1	−.7
.2	.4	−.8	.3	−.4	.1	.3
.2	.4	−.8	.3	.6	.1	.3
.2	−.6	.2	−.7	−.4	.1	−.7
.2	.4	.2	.3	.6	.1	.3
.2	.4	.2	.3	−.4	.1	.3
.2	.4	.2	−.7	.6	.1	.3

still can be incompatible, because the modified classical clustering criterion refers to the data scatter in Eq. (7) also divided by n and thus not constant.

5. Extension to the mixed-scale data case

Contributions of quantitative features to the explained part of the data scatter in (7) also measure association.

When a quantitative feature A_l in a data matrix X is preprocessed according to formula (7) with a scaling factor b_l, its part in the criterion $F(C, c)$ is equal to

$$\sum_{k=1}^{n} c_{kl}^2 |C_k| = \sum_{h\in H} y_{hl}^2 - \sum_{k=1}^{n} \sum_{h\in C_k} (y_{hl} - c_{kl})^2 = |H| \left(\sigma_l^2 - \sum_{k=1}^{n} p_{k+} \sigma_{kl}^2 \right) \Big/ b_l^2 \qquad (12)$$

Here $\sigma_l^2 = \sum_{h\in H} (x_{hl} - a_l)^2 / |H|$ and $\sigma_{lk}^2 = \sum_{h\in C_k} (x_{hl} - a_{lk})^2 / |C_k|$ are the variance and within-class variance of the original variable A_l, respectively; a_l and a_{lk} denote its respective grand mean and within-class mean, and $p_{k+} = |C_k|/|H|$ is the proportion of instances in C_k as above.

This is closely associated with a well known measure of statistical association, the so-called correlation ratio (squared) between C and quantitative A_l defined by:

$$\eta^2(C, A_l) = \frac{\sigma_l^2 - \sum_{k=1}^{n} p_{k+} \sigma_{kl}^2}{\sigma_l^2} \qquad (13)$$

The correlation ratio is between 0 and 1, and the coefficient is equal to 1 only when all the within-class variances are zero (the case of "complete" association between C and A). The greater the within-category variances, the smaller the correlation ratio.

The contribution of clustering C into the scatter of a quantitative feature A_l (12) is obviously $|H|\eta^2(C, A_l)\sigma_l^2/b_l^2$. In particular, this becomes just $|H|\eta^2(C, A_l)$ when the scaling factor b_l in (9) is the standard deviation σ_l.

These observations lead to the following.

Statement 2. *The square-error partitioning criterion with data containing both categorical* (*unordered*) *attrbutes and quantitative features*, *preprocessed as defined above*, *is equivalent to the criterion of maximizing the sum of pairwise correlation coefficients that are equal to* $\Delta(C, A_i)$ *for categorical attributes* A_i *or* $\eta^2(C, A_l)\sigma_l^2/b_l^2$, *for quantitative features* A_l.

The statement suggests the following extension of the category utility function to the case when features can be both quantitative and categorical:

$$CUM(C) = \frac{\sum_i \phi(C, A_i)}{n} \tag{14}$$

where n is the number of classes (concepts) in C and $\phi(C, A_i)$ is either $\Delta(C, A_i)$ or $\eta^2(C, A_i)\sigma_i^2/b_i^2$.

This modification of the category utility function differs from those suggested earlier by Gennari (1990) and Reich and Fenves (1991). No underlying probabilistic distribution of quantitative variables needs to be assumed here.

Specifying the scaling factor b_l is, in fact, equivalent to weighting the feature A_l with respect to the other features in formula (14). Decomposition (7) may give guidance in this.

Indeed, the additive contribution of the column l to the total data scatter in (7) is $w_l = \sum_{h\in H} y_{hl}^2$. For a quantitative A_l, $w_l = \sigma_l^2/b_l^2$ and, for a category V_{ij}, $w_{ij} = p^i_{+j}(1-p^i_{+j})$. The total contribution of a nominal attribute A_i, thus, is

$$W(A_i) = \sum_{V_{ij}\in A_i} w_{ij} = 1 - \sum_{V_{ij}\in A_i} \left(p^i_{+j}\right)^2. \tag{15}$$

Values w_l, w_{ij} and $W(A_i)$ reflect an a priori weighting system for the features and attributes in the matrix Y. Discussion of general principles for further data preprocessing to control these is beyond the scope of this note. One such a principle, to equalise the weights of all meaningful chunks of the data, promoted by the author in his monograph (Mirkin, 1996, p. 288), seems overly simplistic because it doesn't take into account the shapes of feature distributions and thus does not always work.

However, what can be discussed here is the relation between a quantitative feature A and a nominal variable A_t obtained by partitioning the range of A into t qualitative categories, with respect to a clustering (set of concepts) $C = \{C_k\}$. The relation can be captured by comparing the correlation ratio $\eta^2(C, A)$ with a corresponding adjusted contingency coefficient $\Delta(C, A_t)$. The adjustment should equalise the relative weights of the features A and A_t. Following from the discussion above, $\eta^2(C, A)$ is equal to the contribution of A to clustering C, related to the contribution of A to the data scatter. In the case of A_t, the analogous ratio

Table 4. Setting of the experiment.

Class	C_1	C_2	C_3	C_4
Number of observations	200	100	1000	1000
Variance	1.0	1.0	4.0	0.1
Initial mean	0.5	1.0	1.5	2.0
Final mean	10	20	30	40

is $\Delta(C, A_t)/W(A_t)$ which is the adjusted contingency coefficient. This adjusted coefficient is equal to the so-called Goodman-Kruskal's "Tau-b" introduced in Goodman and Kruskall (1954).

The relations between η^2, Δ and Δ/W depend on the bivariate distribution of A and C. However, when the distribution is organized in such a way that all the within-class variances of A are smaller than its overall variance, the pattern of association expressed in Δ and Δ/W generally follows that expressed in η^2.

Example. To illustrate this, an experiment has been set according to the data in Table 4: within each of the classes, $C_1, C_2, C_3,$ and C_4, a prespecified number of observations is randomly generated with the prespecified mean and variance. The totality of 2300 generated observations constitutes the quantitative feature A for which the correlation ratio $\eta^2(C, A)$ is calculated. Then, the range of A is divided in $t = 5$ of equally-spaced intervals (i.e., not necessarily intervals with an equal number of data) constituting categories of the corresponding attribute A_t, for which $\Delta(C, A_t)$ is calculated, as well as its adjusted version, $\Delta(C, A_t)/W(A_t)$.

The initial set of within-class means are not much different with respect to the corresponding variances. Multiplying each of the initial means by the same factor value, $f = 1, 2, \ldots, 20$, the means are step by step diverged in such a way that the within-class samples become more and more distinguishable from each other, thus increasing the association between C and A. The final means in Table 4 correspond to $f = 20$.

This is reflected in figure 2 where the horisontal axis corresponds to the divergence factor, f, and the vertical axis represents values of the three coefficients for the case when the within class random distribution of A is uniform. We can see that the patterns of delta and adjusted delta follow rather closely the pattern of the correlation ratio; the product-moment correlation between η^2 and Δ, in our experiments, is of the order of 0.9. The difference in values of Δ and η^2 is caused by two factors: first, by the coarse qualitative nature of A_t versus to the fine-grained quantitative character of A, and, second, by the difference in their contributions, $W(A_t) < 1$ by A_t and 1 by A, to the data scatter. The second factor is taken into account in the adjusted delta (dashed) line; still there is a difference between the dashed and solid lines because of the first factor.

Similar results are observed for the normal distribution and, to a lesser extent, for the exponential distribution. In the exponential distribution (with density $\alpha \exp^{-\alpha x}$), the variance (α^{-2}) must follow the mean (α^{-1}), so that the within-class variances, in this case, do not fit into the pattern of Table 4, which explains the observation.

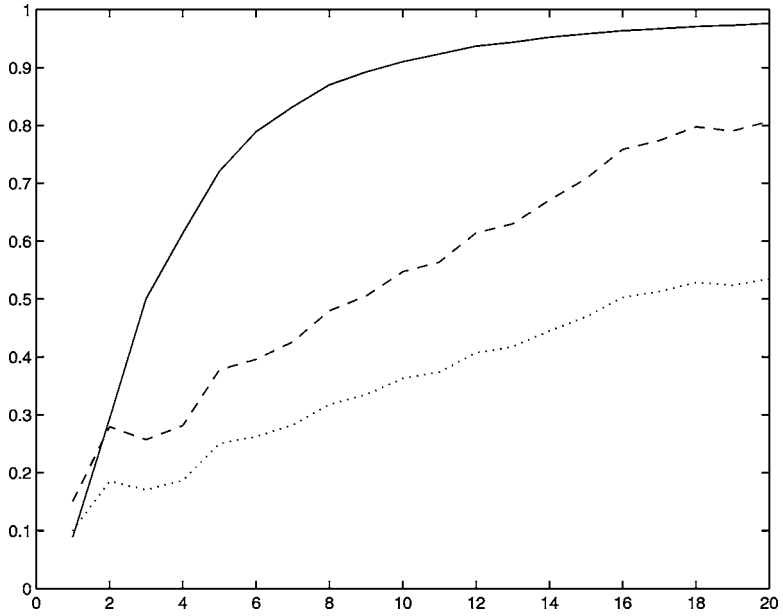

Figure 2. Growth of the correlation ratio (solid line), delta (dotted line) and adjusted delta (dashed line) with increase of the class divergence factor.

6. Conclusion

In this note, a firm relation between the category utility function and earlier, classical clustering criteria has been found. When the categories' binary features are scaled by range, $b = 1$ in (9), the square-error clustering corresponds to the category utility function, or the averaged sum of Wallis-Goodman-Kruskal contingency coefficients between the clustering sought and the attributes given.

The relation stated suggests a partition score function that could be used in the mixed data case. The quantitative features' contributions to the explained data scatter involve the correlation ratio, a well-known coefficient in statistics, that has not been explored in Cobweb associated clustering programs.

In the presented context, the problem of relative weighting of quantitative features and qualitative attributes is related to the standardization of corresponding columns in the data matrix. The Pythagorean decomposition of the data scatter in (7) may help in advancing into solution to this problem.

Acknowledgments

The author is grateful to Doug Fisher for thoroughly editing the paper and its revision, and to the referees for helpful comments. Doug Fisher also proposed to do the experiment described in the example of Section 5 (see Table 4 and figure 2). The author thanks the

Office of Naval Research for its support under grant number N00014-96-1-0208 to Rutgers University.

References

Devaney, M. & Ram, A. (1997). Efficient feature selection in conceptual clustering. In *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 92–97). Nashville, TN: Morgan Kaufmann.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning, 2*, 139–172.

Fisher, D., Xu, L., Carnes, J. R., Reich, Y., Fenves, S. J., Chen, J., Shiavi, R., Biswas, G., & Weinberg, J. (1993). Applying AI clustering to engineering tasks. In *IEEE Expert* (pp. 51–60). December.

Gennari, J. H. (1990). An experimental study of concept formation Technical Report 90-06, Department of Information and Computer Science, University of California, Irvine, CA.

Gluck, M. A. & Corter, J. E. (1985). Information, uncertainty, and the utility of categories. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 283–287). Irvine, CA: Lawrence Erlbaum Associates.

Goodman, L. A. & Kruskal, W. (1954). Measures of association for cross classifications. *Journal of American Statistical Association, 49*, 732–764.

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

Mirkin, B. (1996). *Mathematical Classification and Clustering*. Dordrecht: Kluwer Academic Publishers.

Reich, Y. & Fenves, S. J. (1991). The formation and use of abstract concepts in design. In D. H. Fisher and M. J. Pazzani (Eds.) *Concept formation: Knowledge and experience in unsupervised learning* (pp. 323–353). Los Altos, CA: Morgan Kaufmann.

Received November 1, 1999
Revised December 5, 2000
Accepted November 22, 2000
Final Manuscript November 22, 2000