

# On Clustering Binary Data

Tao Li\*

Shenghuo Zhu<sup>†</sup>

## Abstract

Clustering is the problem of identifying the distribution of patterns and intrinsic correlations in large data sets by partitioning the data points into similarity classes. This paper studies the problem of clustering binary data. This is the case for market basket datasets where the transactions contain items and for document datasets where the documents contain “bag of words”. The contribution of the paper is two-fold. First a new clustering model is presented. The model treats the data and features equally, based on their symmetric association relations, and explicitly describes the data assignments as well as feature assignments. An iterative alternating least-squares procedure is used for optimization. Second, a unified view of binary data clustering is presented by examining the connections among various clustering criteria.

## 1 Introduction

The problem of clustering data arises in many disciplines and has a wide range of applications. Intuitively, clustering is the problem of partitioning a finite set of points in a multi-dimensional space into classes (called clusters) so that (i) the points belonging to the same class are *similar* and (ii) the points belonging to different classes are *dissimilar*.

In this paper, we focus our attention on binary datasets. Binary data have been occupying a special place in the domain of data analysis. Typical applications for binary data clustering include market basket data clustering and document clustering. For market basket data, each data transaction can be represented as a binary vector where each element indicates whether or not any of the corresponding item/product was purchased. For document clustering, each document can be represented as a binary vector where each element indicates whether a given word/term was present or not.

The first contribution of the paper is the introduction of a new clustering model along with a clustering algorithm. A distinctive characteristic of the binary data is that the features (attributes) they include have the same nature as the data they intend to account for: both are binary. This characteristic implies the symmetric association relations between data and features: if the set of data points is associated to

the set of features, then the set of attributes is associated to the set of data points and vice versa. The association relation suggests a new clustering model where the data and features are treated equally. Our new clustering model, *BMD* (Binary Matrix Decomposition), explicitly describes the data assignments (assigning data points into clusters) as well as feature assignments (assigning features into clusters). The clustering problem is then presented as binary matrix decomposition, which is solved via an iterative alternating least-squares optimization procedure. The procedure simultaneously performs two tasks: data reduction (assigning data points into clusters) and feature identification (identifying features associated with each cluster). By explicitly feature assignments, *BMD* produces interpretable descriptions of the resulting clusters. In addition, by iterative feature identification, *BMD* performs an implicit adaptive feature selection at each iteration and flexibly measures the distances between data points. Therefore it works well for high-dimensional data.

The second contribution of this paper is the presentation of a unified view for binary data clustering by examining the connections among various clustering criteria. In particular, we show the equivalence among the matrix decomposition, dissimilarity coefficients, minimum description length and entropy-based approach.

## 2 BMD Clustering

In this section, we describe the new clustering algorithm. Section 2.1 introduces the cluster model. Section 2.2 and Section 2.3 present the optimization procedure and the refining methods, respectively. Section 2.4 gives an example to illustrate the algorithm.

**2.1 The Clustering Model** Suppose the dataset  $X$  has  $n$  instances, having  $r$  features each. Then  $X$  can be viewed as a subset of  $R^r$  as well as a member of  $R^{n \times r}$ . The cluster model is determined by two matrices: the data matrix  $D_{n \times K} = (d_{ik})$  and the feature matrix  $F_{r \times K} = (f_{jk})$ , where  $K$  is the number of clusters.

$$d_{ik} = \begin{cases} 1 & \text{Data point } i \text{ belongs to cluster } k \\ 0 & \text{Otherwise} \end{cases}$$
$$f_{jk} = \begin{cases} 1 & \text{Attribute } j \text{ belongs to cluster } k \\ 0 & \text{Otherwise} \end{cases}$$

\*School of Computer Science, Florida International University, taoli@cs.fiu.edu.

<sup>†</sup>NEC Labs America, Inc., zsh@sv.nec-labs.com. Major work was completed when the author was in University of Rochester.

The data (respectively, feature) matrix specifies the cluster memberships for the corresponding data (respectively, features).

For clustering, it is customary to assume that each data point is assigned to one and only one cluster, i.e.,  $\sum_{k=1}^K d_{ik} = 1$  holds for  $j = 1, \dots, n$ . Given representation  $(D, F)$ , basically,  $D$  denotes the cluster assignments of data points and  $F$  indicates the feature representations of clusters. The  $ij$ -th entry of  $DF^T$  is the dot product of the  $i$ -th row of  $D$  and the  $j$ -th row of  $F$ , and indicates whether the  $j$ -th feature will be present in the  $i$ -instance. Hence,  $DF^T$  can be interpreted as the approximation of the original data  $X$ . Our goal is then to find a  $(D, F)$  that minimizes the squared error between  $X$  and its approximation  $DF^T$ .

$$(2.1) \quad \operatorname{argmin}_{D, F} O = \frac{1}{2} \|X - DF^T\|_F^2,$$

where  $\|X\|_F$  is the Frobenius norm of the matrix  $X$ , i.e.,  $\sqrt{\sum_{i,j} x_{ij}^2}$ . With the formulation, we transform the data clustering problem into the computation of  $D$  and  $F$  that minimizes the criterion  $O$ .

**2.2 Optimization Procedure** The objective criterion can be expressed as

$$(2.2) \quad \begin{aligned} O_{D, F} &= \frac{1}{2} \|X - DF^T\|_F^2 \\ &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m \left( x_{ij} - \sum_{k=1}^K d_{ik} f_{kj} \right)^2 \\ &= \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K d_{ik} \sum_{j=1}^m (x_{ij} - f_{kj})^2 \\ &= \frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K d_{ik} \sum_{j=1}^m (x_{ij} - y_{kj})^2 \\ &\quad + \frac{1}{2} \sum_{k=1}^K n_k \sum_{j=1}^m (y_{kj} - f_{kj})^2, \end{aligned}$$

where  $y_{kj} = \frac{1}{n_k} \sum_{i=1}^n d_{ik} x_{ij}$  and  $n_k = \sum_{i=1}^n d_{ik}$  (note that we use  $f_{kj}$  to denote the entry of  $F^T$ ). The objective function can be minimized via an alternating least-squares procedure by alternatively optimizing one of  $D$  or  $F$  while fixing the other.

Given an estimate of  $F$ , new least-squares estimates of the entries of  $D$  can be determined by assigning each data point to the closest cluster as follows:

$$(2.3) \quad \hat{d}_{ik} = \begin{cases} 1 & \text{if } \sum_{j=1}^m (x_{ij} - f_{kj})^2 < \sum_{j=1}^m (x_{ij} - f_{lj})^2 \\ & \text{for } l = 1, \dots, K, l \neq k \\ 0 & \text{Otherwise} \end{cases}$$

When  $D$  is fixed,  $O_{D, F}$  can be minimized with respect to  $F$  by minimizing the second part of Equation (2.2):

$$O'(F) = \frac{1}{2} \sum_{k=1}^K n_k \sum_{j=1}^m (y_{kj} - f_{kj})^2.$$

Note that  $y_{kj}$  can be thought of as the probability that the  $j$ -th feature is present in the  $k$ -th cluster. Since each  $f_{kj}$  is binary<sup>1</sup>, i.e., either 0 or 1,  $O'(F)$  is minimized by:

$$(2.4) \quad \hat{f}_{kj} = \begin{cases} 1 & \text{if } y_{kj} > 1/2 \\ 0 & \text{Otherwise} \end{cases}$$

In practice, if a feature has similar association to all clusters, then it is viewed as an outlier at the current stage.

The optimization procedure for minimizing Equation (2.2) alternates between updating  $D$  based on Equation (2.3) and assigning features using Equation (2.4). After each iteration, we compute the value of the objective criterion  $O(D, F)$ . If the value is decreased, we then repeat the process; otherwise, the process has arrived at a local minimum. Since the *BMD* procedure monotonically decreases the objective criterion, it converges to a local optimum. The clustering procedure is shown in Algorithm 1.

---

**Algorithm 1 BMD: clustering procedure**

---

Input: (data points:  $X_{n \times r}$ , # of classes:  $K$ )

Output:  $D$ : cluster assignment;

$F$ : feature assignment;

**begin**

1. **Initialization:**

1.1 Initialize  $D$

1.2 Compute  $F$  based on Equation (2.4)

1.3 Compute  $O_0 = O(D, F)$

2. **Iteration:**

**begin**

2.1 Update  $D$  given  $F$  (via Equation (2.3))

2.2 Compute  $F$  given  $D$  (via Equation (2.4))

2.3 Compute the value of  $O_1 = O(D, F)$ ;

2.4 if  $O_1 < O_0$

2.4.1  $O_0 = O_1$

2.4.2 Repeat from 2.1

2.5 else

2.5.1 break; (Converges)

**end**

3. **Return**  $D, F$ ;

**end**

---

**2.3 Refining Methods** Clustering results are sensitive to initial seed points. The initialization step sets the initial values for  $D$  and  $F$ . Since  $D$  is a binary matrix and has at most one occurrence of 1 in each row, it is very sensitive to initial assignments. To overcome the sensitivity of initialization, we refine the procedure. Its idea is to use mutual information to measure the similarity between a pair

<sup>1</sup>If the entries of  $F$  are arbitrary, then the optimization here can be performed via singular value decomposition.

of clustering results. In addition, clustering a large data set may be time-consuming. To speed up the algorithm, a small set of data points, for example, 1% of the entire data set, may be selected as a *bootstrap* data set. The clustering algorithm is first executed on the bootstrap data set. Then, the algorithm is run on the entire data set using the data assignments obtained from the bootstrap data set (instead of using random seed points).

**2.4 An Example** To illustrate how *BMD* works, we show an artificial example, a dataset consisting of six sentences from two clusters: user interaction and distributed systems, as shown in Figure 1.

- 1(1) An system for user response
- 2(1) A survey of user interaction  
on computer response
- 3(1) Response for interaction
- 4(2) A multi-user distributed system
- 5(2) A survey of distributed computer system
- 6(2) distributed systems

Figure 1: The six example sentences. The numbers within the parentheses are the clusters: 1=user interaction, 2=distributed system.

After preprocessing, we get the dataset as in Table 1. In this example,  $D$  is a  $6 \times 2$  matrix and  $F$  is a  $7 \times 2$  matrix. Initially, the data points 2 and 5 are chosen as seed points, where the data point 2 is in class 1 and the data point 5 is in class 2. Initialization is then performed on the seed points to get the initial feature assignments. After *Step 1.2*, features  $a$ ,  $b$  and  $c$  are positive in class 1,  $e$  and  $f$  are positive in class 2, and  $d$  and  $g$  are outliers. In other words,  $F(a, 1) = F(b, 1) = F(c, 1) = 1$ ,  $F(e, 2) = F(f, 2) = 1$ , and all the other entries<sup>2</sup> of  $F$  are 0. Then *Step 2.1* assigns data points 1, 2 and 3 to class 1 and data points 4, 5 and 6 to class 2. then *Step 2.2* asserts  $a$ ,  $b$  and  $c$  are positive in class 1,  $d$ ,  $e$  and  $f$  are positive in class 2, and  $g$  is an outlier. In the next iteration, the objective criterion does not change. At this point the algorithm stops. The resulting clusters are: For data points, class 1 contains 1, 2, and 3 and class 2 contains 4, 5, and 6. For features,  $a$ ,  $b$  and  $c$  are positive in class 1,  $d$ ,  $e$  and  $f$  are positive in class 2 while  $g$  is an outlier.

We have conducted experiments on real datasets to evaluate the performance of our *BMD* algorithm and compare it with other standard clustering algorithms. Experimental results on suggest that *BMD* is a viable and competitive binary clustering algorithm. Due to space limit, we omitted the experiment details.

### 3 Binary Data Clustering

In this section, a unified view on binary data clustering is presented by examining the relations among various binary

data point	feature	a	b	c	d	e	f	g
1		1	1	0	0	1	0	0
2		1	1	1	1	0	0	1
3		1	0	1	0	0	0	0
4		0	1	0	0	1	1	0
5		0	0	0	1	1	1	1
6		0	0	0	1	0	1	0

Table 1: A bag-of-word representation of the sentences.  $a$ ,  $b$ ,  $c$ ,  $d$ ,  $e$ ,  $f$ ,  $g$ , correspond to the presence of *response*, *user*, *interaction*, *computer*, *system*, *distributed* and *survey*, respectively.

clustering approaches. Section 3.1 sets down the notation, Section 3.2, Section 3.3 and Section 3.4 discuss the binary dissimilarity coefficients, minimum description length, and the entropy-based approach respectively. The unified view on binary clustering is summarized in Figure 2. Note that the relations of maximum likelihood principle with the entropy-based criterion and with minimum description length (MDL) are known in machine learning literature [8].

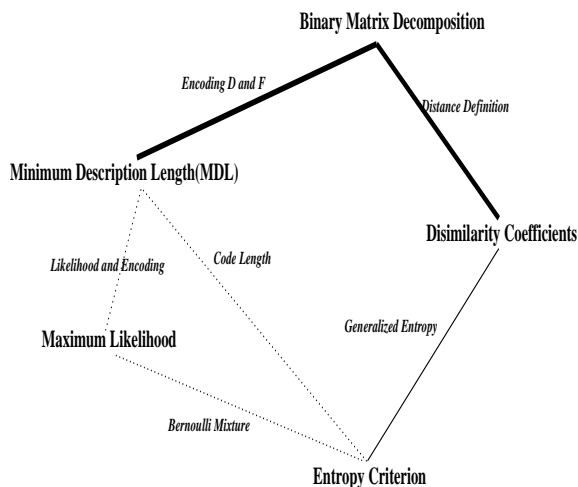


Figure 2: A Unified View on Binary Clustering. The thick lines are relations first shown in this paper, the dotted lines are well-known facts, and the thin line is first discussed in [7].

**3.1 Notation** We first set down some notation. Suppose that a set of  $n$   $r$ -dimensional binary data vectors,  $X$ , represented as an  $n \times r$  matrix,  $(x_{ij})$ , is partitioned into  $K$  classes  $C = (C_1, \dots, C_K)$  and we want the points within each class are *similar* to each other. We view  $C$  as a partition of the indices  $\{1, \dots, n\}$ . So, for all  $i$ ,  $1 \leq i \leq n$ , and  $k$ ,  $1 \leq k \leq K$ , we write  $i \in C_k$  to mean that the  $i$ -th vector belongs to the  $k$ -th class. Let  $N = nr$ . For each  $k$ ,  $1 \leq k \leq K$ , let  $n_k = \|C_k\|$ ,  $N_k = n_k r$ , and for each  $j$ ,  $1 \leq j \leq r$ , let  $N_{j,k,1} = \sum_{i \in C_k} x_{ij}$  and  $N_{j,k,0} = n_k - N_{j,k,1}$ . Also, for each  $j$ ,  $1 \leq j \leq r$ , let

<sup>2</sup>We use  $a, b, c, d, e, f, g$  to denote the rows of  $F$ .

$N_{j,1} = \sum_{i=1}^n x_{ij}$  and  $N_{j,0} = n - N_{j,1}$ . We use  $x_i$  as a point variable.

**3.2 Binary Dissimilarity Coefficients** A popular partition-based criterion (within-cluster) for clustering is to minimize the summation of distances/dissimilarities inside the cluster. The within-cluster criterion can be described as minimizing

$$(3.5) \quad S(C) = \sum_{k=1}^K \frac{1}{n_k} \sum_{i,i' \in C_k} \delta(x_i, x_{i'}),$$

or <sup>3</sup>

$$(3.6) \quad S(C) = \sum_{k=1}^K \sum_{i,i' \in C_k} \delta(x_i, x_{i'}),$$

where  $\delta(x_i, x_{i'})$  is the distance measure between  $x_i$  and  $x_{i'}$ . For binary clustering, the dissimilarity coefficients are popular measures of the distances.

**3.2.1 Various Coefficients** Given two binary data points,  $w$  and  $w'$ , there are four fundamental quantities that can be used to define similarity between the two [1]:  $a = \|\{j \mid w_j = w'_j = 1\}\|$ ,  $b = \|\{j \mid w_j = 1 \wedge w'_j = 0\}\|$ ,  $c = \|\{j \mid w_j = 0 \wedge w'_j = 1\}\|$ , and  $d = \|\{j \mid w_j = w'_j = 0\}\|$ , where  $1 \leq j \leq r$ . It has been shown in [1] that the presence/absence based dissimilarity measure can be generally <sup>4</sup> written as  $D(a, b, c, d) = \frac{b+c}{\alpha a + b + c + \beta d}$ , where  $\alpha > 0$  and  $\beta \geq 0$ . Dissimilarity measures can be transformed into a similarity function by simple transformations such as adding 1 and inverting, dividing by 2 and subtracting from 1, etc. [6]. If the joint absence of the attribute is ignored, i.e.,  $\beta$  is set to 0, then the binary dissimilarity measure can be generally written as  $D(a, b, c, d) = \frac{b+c}{\alpha a + b + c}$ , where  $\alpha > 0$ .

In cluster applications, the rankings based on a dissimilarity coefficient is often of more interest than the actual value of the dissimilarity coefficient. It has been shown that [1], if the paired absences are ignored in the calculation of dissimilarity values, then there is only one single dissimilarity coefficient modulo the global order equivalence:  $\frac{b+c}{a+b+c}$ . Thus our following discussion is based on the single dissimilarity coefficient.

**3.2.2 BMD and Dissimilarity Coefficients** Given representation  $(D, F)$ , basically,  $D$  denotes the assignments of data points associated into clusters and  $F$  indicates the fea-

ture representations of clusters. Observe that

$$(3.7) \quad \begin{aligned} O(D, F) &= \frac{1}{2} \|X - DF^T\|_F^2 \\ &= \frac{1}{2} \sum_{i,j} (x_{ij} - (DF^T)_{ij})^2 \\ &= \frac{1}{2} \sum_{k=1}^K \sum_{i \in C_k} \sum_j |x_{ij} - e_{kj}|^2 \\ &= \frac{1}{2} \sum_{k=1}^K \sum_{i \in C_k} d(x_i, e_k), \end{aligned}$$

where  $e_k = (f_{k1}, \dots, f_{kr})$ ,  $i = 1, \dots, K$  is the cluster ‘‘representative’’ of cluster  $C_i$ . Thus minimizing Equation (3.7) is the same as minimizing Equation (3.6) where the distance is defined as  $d(x_i, e_k) = \sum_j |x_{ij} - (e_k)_{ij}|^2 = \sum_j |x_{ij} - (e_k)_{ij}|$  (the last equation holds since  $x_{ij}$  and  $(e_k)_{ij}$  are all binary). In fact, given two binary vectors  $X$  and  $Y$ ,  $\sum_i |X_i - Y_i|$  calculates their mismatches (the numerator of their dissimilarity coefficients).

**3.3 Minimum Description Length** Minimum Description length (MDL) aims at searching for a model that provides the most compact encoding for data transmission [10] and is conceptually similar to minimum message length (MML) [9, 2] and stochastic complexity minimization [11]. In fact, the MDL approach is a Bayesian method: the code lengths and the code structure in the coding model are equivalent to the negative log probabilities and probability structure assumptions in the Bayesian approach.

As described in Section 2, in BMD clustering, the original matrix  $X$  can be approximated by the matrix product of  $DF^T$ . Instead of encoding the elements of  $X$  alone, we then encode the model,  $D, F$ , and the data given the model,  $(X|DF^T)$ . The overall code length is thus expressed as

$$L(X, D, F) = L(D) + L(F) + L(X|DF^T).$$

In the Bayesian framework,  $L(D)$  and  $L(F)$  are negative log priors for  $D$  and  $F$  and  $L(X|DF^T)$  is a negative log likelihood of  $W$  given  $D$  and  $F$ . If we assume that the prior probabilities of all the elements of  $D$  and  $F$  are uniform (i.e.,  $\frac{1}{2}$ ), then  $L(D)$  and  $L(F)$  are fixed given the dataset  $X$ . In other words, we need to use one bit to represent each element of  $D$  and  $F$  irrespective of the number of 1’s and 0’s. Hence, minimizing  $L(X, D, F)$  reduces to minimizing  $L(X|DF^T)$ .

Use  $\hat{X}$  to denote the generated data matrix by  $D$  and  $F$ . For all  $i$ ,  $1 \leq i \leq n$ ,  $j$ ,  $1 \leq j \leq p$ ,  $b \in \{0, 1\}$ , and  $c \in \{0, 1\}$ , we consider  $p(x_{ij} = b \mid \hat{x}_{ij}(D, F) = c)$ , the probability of the original data  $W_{ij} = b$  conditioned upon the generated data  $(\hat{x})_{ij}$ , via  $DF^T$ , is  $c$ . Note that

$$p(x_{ij} = b \mid \hat{X}_{ij}(D, F) = c) = \frac{N_{bc}}{N_c}.$$

<sup>3</sup>Equation (3.5) computes the weighted sum using the cluster sizes.

<sup>4</sup>Basically, the presence/absence based dissimilarity measure satisfies a set of axioms such as non-negative, range in  $[0, 1]$ , rationality whose numerator and denominator are linear and symmetric, etc. [1].

Here  $N_{bc}$  is the number of elements of  $X$  which have value  $b$  where the corresponding value for  $\hat{X}$  is  $c$ , and  $N_{.c}$  is the number of elements of  $\hat{X}$  which have value  $c$ . Then the code length for  $L(X, D, F)$  is

$$\begin{aligned} L(X, D, F) &= - \sum_{b,c} N_{bc} \log P(x_{ij} = b \mid \hat{x}_{ij}(D, F) = c) \\ &= -np \sum_{b,c} \frac{N_{bc}}{np} \log \frac{N_{bc}}{N_{.c}} \\ &= npH(X|\hat{X}(D, F)) \end{aligned}$$

So minimizing the coding length is equivalent to minimizing the conditional entropy. Denote  $p_{bc} = p(x_{ij} = b \mid \hat{x}_{ij}(D, F) = c)$ . We wish to find the probability vectors  $\mathbf{p} = (p_{00}, p_{01}, p_{10}, p_{11})$  that minimize

$$(3.8) \quad H(X|\hat{X}(D, F)) = - \sum_{i,j \in \{0,1\}} p_{ij} \log p_{ij}$$

Since  $-p_{ij} \log p_{ij} \geq 0$ , with the equality holding at  $p_{ij} = 0$  or 1, the only possible probability vectors which minimize  $H(X|\hat{X}(D, F))$  are those with  $p_{ij} = 1$  for some  $i, j$  and  $p_{i_1 j_1} = 0, (i_1, j_1) \neq (i, j)$ . Since  $\hat{X}$  is an approximation of  $X$ , it is natural to require that  $p_{00}$  and  $p_{11}$  be close to 1 and  $p_{01}$  and  $p_{10}$  be close to 0. This is equivalent to minimizing the mismatches between  $X$  and  $\hat{X}$ , i.e., minimizing  $O(D, F) = \frac{1}{2} \|X - DF^T\|_F^2$ .

### 3.4 Entropy-Based Approach

**3.4.1 Classical Entropy Criterion** The classical clustering criteria [3, 4] search for a partition  $C$  that maximizes the following quantity  $O(C)$ :

$$\begin{aligned} (3.9) \quad O(C) &= \sum_{k=1}^K \sum_{j=1}^r \sum_{t=0}^1 \frac{N_{j,k,t}}{N} \log \frac{NN_{j,k,t}}{N_k N_{j,t}} \\ &= \sum_{k=1}^K \sum_{j=1}^r \sum_{t=0}^1 \frac{N_{j,k,t}}{N} \left( \log \frac{N_{j,k,t}}{n_k} - \log \frac{N_{j,t}}{n} \right) \\ &= \frac{1}{r} \left( \hat{H}(X) - \frac{1}{n} \sum_{k=1}^K n_k \hat{H}(C_k) \right). \end{aligned}$$

Observe that  $\frac{1}{n} \sum_{k=1}^K n_k \hat{H}(C_k)$  is the entropy measure of the partition, i.e., the weighted sum of each cluster's entropy. This leads to the following criterion: Given a dataset, fix  $\hat{H}(X)$ , then maximizing  $O(C)$  is equivalent to minimizing the expected entropy of the partition:

$$(3.10) \quad \frac{1}{n} \sum_{k=1}^K n_k \hat{H}(C_k)$$

**3.4.2 Entropy and Dissimilarity Coefficients** Now examine the within-cluster criterion in Equation (3.5). We

have:

$$\begin{aligned} S(C) &= \sum_{k=1}^K \frac{1}{n_k} \sum_{i,i' \in C_k} \delta(x_i, x_{i'}) \\ &= \sum_{k=1}^K \frac{1}{n_k} \sum_{i,i' \in C_k} \frac{1}{r} \sum_{j=1}^r |x_{i,j} - x_{i',j}| \\ &= \frac{1}{r} \sum_{k=1}^K \sum_{j=1}^r n_k \rho_k^{(j)} (1 - \rho_k^{(j)}). \end{aligned}$$

Here for each  $k, 1 \leq k \leq K$ , and for each  $j, 1 \leq j \leq r$ ,  $\rho_k^{(j)}$  is the probability that the  $j$ -th attribute is 1 in  $C_k$ .

Using the generalized entropy<sup>5</sup> defined in [5],  $H^2(Q) = -2 \left( \sum_{i=1}^n q_i^2 - 1 \right)$ , we have

$$\begin{aligned} &\frac{1}{n} \sum_{k=1}^K n_k \hat{H}(C_k) \\ &= -\frac{1}{2n} \sum_{k=1}^K \sum_{j=1}^r n_k \left( (\rho_k^{(j)})^2 + (1 - \rho_k^{(j)})^2 - 1 \right) \\ &= \frac{1}{n} \sum_{k=1}^K \sum_{j=1}^r n_k \rho_k^{(j)} (1 - \rho_k^{(j)}) = \frac{r}{n} S(C). \end{aligned}$$

### References

- [1] F. B. Baulieu. Two variant axiom systems for presence/absence based dissimilarity coefficients. *Journal of Classification*, 14(1):159–170, 1997.
- [2] R. A. Baxter and J. J. Oliver. MDL and MML: similarities and differences. TR 207, Monash University, 1994.
- [3] H.-H. Bock. Probabilistic aspects in cluster analysis. In *Conceptual and Numerical Analysis of Data*, pages 12–44, 1989.
- [4] G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8(2):175–176, 1991.
- [5] J. Havrda and F. Charvat. Quantification method of classification processes: Concept of structural a-entropy. *Kybernetika*, 3:30–35, 1967.
- [6] N. Jardine and R. Sibson. *Mathematical Taxonomy*. John Wiley & Sons, 1971.
- [7] T. Li, S. Ma, and M. Ogihara. Entropy-based criterion in categorical clustering. In *ICML, 2004*. 536–543.
- [8] T. M. Mitchell. *Machine Learning*. The McGraw-Hill Companies, Inc., 1997.
- [9] J. J. Oliver and R. A. Baxter. MML and Bayesianism: similarities and differences. TR 206, Monash University, 1994.
- [10] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
- [11] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Press, Singapore, 1989.

<sup>5</sup>Note that  $H^s(Q) = (2^{(1-s)} - 1)^{-1} (\sum_{i=1}^n q_i^s - 1)$ .