Machine Learning, 62, 199–215, 2006
2006 Springer Science + Business Media, Inc. Manufactured in The Netherlands.
DOI: 10.1007/s10994-005-5316-9

A Unified View on Clustering Binary Data

TAO LI taoli@cs.fiu.edu
School of Computer Science, Florida International University, 11200 SW 8th Street, Miami, FL, 33199

Editor: Jennifer Dy

Published online: 29 January 2006

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. Binary data have been occupying a special place in the domain of data analysis. A unified view of binary data clustering is presented by examining the connections among various clustering criteria. Experimental studies are conducted to empirically verify the relationships.

Keywords: clustering, binary data, unified view

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* (Hartigan, 1975; Kaufman & Rousseeuw, 1990). 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 (Agrawal & Srikant, 1994). 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 (Li et al., 2004a; Li, 2005).

Generally clustering problems are determined by four basic components: a) the (physical) representation of the given data set; b) the distance/dissimilarity measures between data points; c) the criterion/objective function which the clustering solutions should aim to optimize; and, d) the optimization procedure. For a given data clustering problem, the four components are tightly coupled. Various methods/criteria have been proposed over the years from various perspectives and with various focuses (Barbara et al., 2002; Gibson et al., 1998; Huang, 1998; Ganti et al., 1999; Guha et al., 2000; Gyllenberg et al., 1997; Li et al., 2004b). However, few attempts have been made to establish the connections between them while highlighting their differences. In this paper, we aim to provide a unified view of binary data clustering by examining the connections among various clustering criteria. In particular, we show the relationships among the entropy criterion, dissimilarity coefficients, mixture models, matrix decomposition, and minimum description length.

Table 1. Notation.

n, number of data points	p, number of features
K, number of clusters	$C = (C_1, C_2, \ldots, C_K)$, Clustering
$X = (x_{ij})_{n\times p}$, the dataset	n_k, the cardinality of the class C_k
$N = n \times p$	$N_k = n_k \times p$
$N_k^{j1} = \sum_{i\in C_k} x_{ij}$	$N_k^{j0} = n_k - N_k^{j1}$
$N^{j1} = \sum_{i=1}^{n} x_{ij}$	$N^{j0} = n - N^{j1}$
x_t, a point variable	y_j, a feature variable
$\hat{H}$, Estimated Entropy	

The rest of the paper is organized as follows: Section 2 sets down some notations used throughout the paper; Section 3 presents the unified view on binary data clustering by examining the connections among various clustering criteria. In particular, Section 3.1 introduces the traditional entropy-based clustering criterion; Section 3.2 establishes the relations between entropy-based criterion with dissimilarity coefficients; Section 3.3 shows the equivalence between the entropy-based criterion with the classification likelihood; Section 3.4 illustrates the connections between a matrix perspective and dissimilarity coefficients; Section 3.5 describes minimum description length approach and its relation with the matrix perspective. Section 4 presents experimental studies to empirically verify the relationships and finally Section 5 concludes.

2. Notations

Given a binary dataset $X = (x_{ij})_{n\times p}$ where $x_{ij} = 1$ if the jth feature is present in the ith instance and $x_{ij} = 0$ otherwise, we want to find a partition of X into classes $C = (C_1, C_2, \ldots, C_K)$ such that the points within each class are *similar* to each other. Let n_k be the cardinality of the class C_k, $1 \le k \le K$. We will use N for np, N_k for $n_k p$, N_k^{j1} for $\sum_{i\in C_k} x_{ij}$, N_k^{j0} for $n_k - N_k^{j1}$, N^{j1} for $\sum_{i=1}^{n} x_{ij}$, N^{j0} for $n - N^{j1}$, and $\hat{H}$ for the estimated entropy, where x_t is a point variable and y_j is a feature variable.

Table 1 summarizes the notation that will be used throughout the paper.

Consider a discrete random vector $Y = (y_1, y_2, \ldots, y_p)$ with p independent components y_i where y_i take its value from a finite set V_i. The entropy of Y is defined as

$$\begin{aligned} H(Y) &= -\sum p(Y)\log p(Y) = \sum_{i=1}^{p} H(y_i) \\ &= -\sum_{i=1}^{p}\sum_{t\in V_i} p(y_i = t)\log p(y_i = t) \end{aligned}$$

3. A unified view

In summary, the connections between various methods/criteria for binary clustering are presented in Figure 1. In the rest of the section, we will further illustrate the relationships in detail.

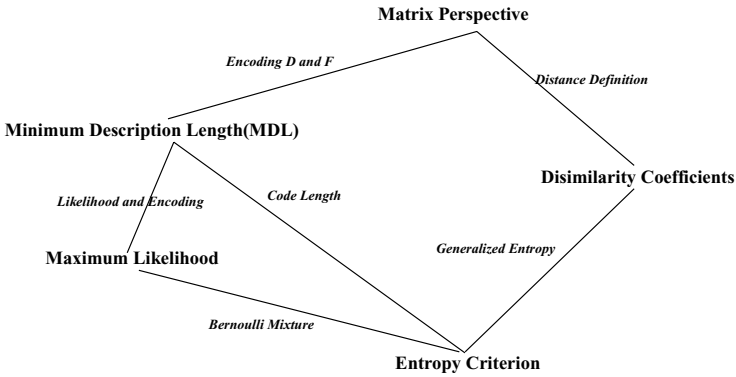

Figure 1. Summary of relations for various clustering criteria. The words beside the arrows describe connections between the criteria.

3.1. Classical entropy criterion

As measures for uncertainty presented in random variables, entropy-type criteria for the heterogeneity of object clusters have been used since the early times of cluster analysis (Block, 1989). In this section, we first study the entropy-based criteria in categorical clustering. In particular, we will show that the entropy-based clustering criteria can be formally derived in the formal framework of probabilistic clustering models.

Entropy criterion The classical clustering criterion (Bock, 1989; Celeux & Govaert, 1991) is to find the partition C such that

$$\begin{aligned} O(C) &= \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{t=0}^{1}\frac{N_k^{jt}}{N}\log\frac{NN_k^{jt}}{N_kN^{jt}} \\ &= \sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{t=0}^{1}\frac{N_k^{jt}}{N}\left(\log\frac{N_k^{jt}}{n_k}-\log\frac{N^{jt}}{n}\right) \\ &= \frac{1}{N}\sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{t=0}^{1}n_k\frac{N_k^{jt}}{n_k}\log\frac{N_k^{jt}}{n_k}-\sum_{j=1}^{p}\sum_{t=0}^{1}\frac{N^{jt}}{N}\log\frac{N^{jt}}{n} \\ &= \frac{1}{p}\left(\hat{H}(X)-\frac{1}{n}\sum_{k=1}^{K}n_k\hat{H}(C_k)\right) \end{aligned} \tag{1}$$

is maximized.[1] 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. Given a dataset, $\hat{H}(X)$ is then fixed, to maximize $O(C)$ is to minimize the expected entropy of the partition

$$\frac{1}{n}\sum_{k=1}^{K}n_k\hat{H}(C_k). \tag{2}$$

Intuitively, lower expected entropy means less uncertainty, and hence lead to better clustering.

Kullback-Leibler measure The entropy criterion above can also be thought of as a Kullback-Leibler (K-L) measure. The idea is as follows: suppose the observed dataset is generated by a number of classes. We first model the unconditional probability density function and then seek a number of partitions whose combination yields the density function (Roberts et al., 1999; Roberts 2000). The K-L measure then tries to measure the difference between the unconditional density and the partitional density.

Given two distributions $p(y)$ and $q(y)$,

$$KL(p(y)||q(y)) = \int p(y) \log\left(\frac{p(y)}{q(y)}\right) dy.$$

For each $t \in \{0, 1\}$, let $p(y_j = t) = \frac{N^{jt}}{n}$ and $q(y_j = t) = p_k(y_j = t) = \frac{N_k^{jt}}{n_k}$, $t = \{0, 1\}$. Note that here $p(y)$ represents the unconditional density while $q(y)$ (or $p_k(y)$) is the partitional density. Then $KL(p(y)||q(y)) \approx \sum p(y)\log(p(y)) - \sum p(y)\log(q(y))$.

Proposition 1. *The entropy criterion $O(C)$ given in Equation (1) can be approximated by*

$$\frac{1}{p}[(1-K)\hat{H}(X) - \sum_{k=1}^{K} KL(p(y)||p_k(y))]$$

where $p(y_j = t) = \frac{N^{jt}}{n}$ and $q(y_j = t) = p_k(y_j = t) = \frac{N_k^{jt}}{n_k}$, $t = \{0, 1\}$.

Proof: Observe that

$$\begin{aligned}
KL(p(y) \parallel q(y)) &\approx \sum p(y)\log(p(y)) - \sum p(y)\log(q(y)) \\
&= \sum_{j=1}^{p} \sum_{t\in\{0,1\}} p(y_j = t)\log(p(y_j = t)) \\
&\quad - \sum_{j=1}^{p} \sum_{t\in\{0,1\}} p(y_j = t)\log(q(y_j = t)) \\
&= -\hat{H}(X) - \sum_{j=1}^{p} \sum_{t\in\{0,1\}} \frac{N^{jt}}{n} \log\left(\frac{N_k^{jt}}{n_k}\right) \\
&= -\hat{H}(X) + \frac{n_k}{n}\hat{H}(C_k)
\end{aligned}$$

Thus, using Equation (1), $O(C)$ is equal to

$$\frac{1}{p}\left[(1-K)\hat{H}(X) - \sum_{k=1}^{K} KL(p(y) \parallel p_k(y))\right]. \tag{3}$$

□

Hence, minimizing the K-L measure is equivalent to minimizing the expected entropy of partition over the observed data.

3.2. Entropy and dissimilarity coefficients

In this section, we show the relationship between the entropy criterion and the dissimilarity coefficients. A popular partition-based criterion (within-cluster) for clustering is to minimize the summation of dissimilarities inside the cluster. Let $C = (C_1, \ldots, C_K)$ be the partition, then the within-cluster criterion can be described as minimizing

$$D(C) = \sum_{k=1}^{K} \frac{1}{n_k} \sum_{x_i, x_j \in C_k} d(x_i, x_j), \tag{4}$$

where $d(x_i, x_j)$ is the distance measure between x_i and x_j and n_k is the size of cluster k. In general, the distance function can be defined using L_p norm. For binary clustering, however, the dissimilarity coefficients are popular measures of the distances.

Dissimilarity coefficients Let a set X of n data points and a set A of p binary attributes be given. Given two data points x_1 and x_2, there are four fundamental quantities that can be used to define the similarity between the two (Baulieu, 1997):

- $a = card(x_{1j} = x_{2j} = 1)$,
- $b = card(x_{1j} = 1 \& x_{2j} = 0)$,
- $c = card(x_{1j} = 0 \& x_{2j} = 1)$,
- $d = card(x_{1j} = x_{2j} = 0)$,

where $j = 1, \ldots, p$ and *card* represents *cardinality*. The presence/absence based dissimilarity measure that satisfies a set of axioms (such as non-negative, range in [0, 1], rationality whose numerator and denominator are linear and symmetric) can be generally written as $D(a, b, c, d) = \frac{b+c}{\alpha a+b+c+\delta d}, \alpha > 0, \delta \geq 0$ (Baulieu, 1997). 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. (Jardine & Sibson, 1971). If the joint absence of the attribute is ignored[2], i.e., setting $\delta = 0$, then the binary dissimilarity measure can be generally written as $D(a, b, c, d) = \frac{b+c}{\alpha a+b+c}, \alpha > 0$. Table 2 listed several common dissimilarity coefficients and the corresponding similarity coefficients.

Global equivalence on coefficients In cluster applications, the ranking based on a dissimilarity coefficient is often of more interest than the actual value of the dissimilarity coefficient. The following propositions proved in Baulieu (1997) establish the equivalence results among dissimilarity coefficients.

Definition 1. *Two dissimilarity coefficients D and D' are said to be* **globally order equivalent** *provided $\forall (a_1, b_1, c_1, d_1), (a_2, b_2, c_2, d_2) \in (Z^+)^4$, we have $D(a_2, b_2, c_2, d_2) < D(a_1, b_1, c_1, d_1)$ if and only if $D'(a_2, b_2, c_2, d_2) < D'(a_1, b_1, c_1, d_1)$.*

Table 2. Binary dissimilarity and similarity coefficients. The "Metric" column indicates whether the given dissimilarity coefficient is a metric or not. A 'Y' stands for 'YES' while an 'N' stands for 'No'.

Name	Similarity	Dissimilarity	Metric
Simple matching Coeff.	$\frac{a+d}{a+b+c+d}$	$\frac{b+c}{a+b+c+d}$	Y
Jaccard's Coeff.	$\frac{a}{a+b+c}$	$\frac{b+c}{a+b+c}$	Y
Dice's Coeff.	$\frac{2a}{2a+b+c}$	$\frac{b+c}{2a+b+c}$	N
Russel & Rao's Coeff.	$\frac{a}{a+b+c+d}$	$\frac{b+c+d}{a+b+c+d}$	Y
Rogers & Tanimoto's Coeff.	$\frac{\frac{1}{2}(a+d)}{\frac{1}{2}(a+d)+b+c}$	$\frac{b+c}{\frac{1}{2}(a+d)+b+c}$	Y
Sokal & Sneath's Coeff. I	$\frac{\frac{1}{2}a}{\frac{1}{2}a+b+c}$	$\frac{b+c}{\frac{1}{2}a+b+c}$	Y
Sokal & Sneath's Coeff. II	$\frac{2(a+d)}{2(a+d)+b+c}$	$\frac{b+c}{2(a+d)+b+c}$	N

Proposition 2. *Given two dissimilarity coefficients* $D = \frac{b+c}{\alpha a+b+c+\delta d}$ *and* $D' = \frac{b+c}{\alpha' a+b+c+\delta' d}$. *If* $\alpha\delta' = \alpha'\delta$, *then* D *and* D' *are* **globally order equivalent**.

Corollary 1. If two dissimilarity coefficients can be expressed as $D = \frac{b+c}{\alpha a+b+c}$ and $D' = \frac{b+c}{\alpha' a+b+c}$, then D and D' are ***globally order equivalent***. □

In other words, if the paired absences are to be ignored in the calculation of dissimilarity values, then there is only one single dissimilarity coefficient up to the global order equivalence: $\frac{b+c}{a+b+c}$. With the equivalence results, our following discussion is then based on the single dissimilarity coefficient.

Entropy and dissimilarity coefficients Consider the coefficient $\frac{b+c}{a+b+c}$. Note that $b+c$ in the numerator is the number of mismatches between binary vectors.

Now let's take a closer look at the within-cluster criterion in Equation 4 defined by the dissimilarity coefficients.

$$\begin{aligned} D(C) &= \sum_{k=1}^{K} \frac{1}{n_k} \sum_{x_{t_1}, x_{t_2} \in C_k} d(x_{t_1}, x_{t_2}) \\ &= \sum_{k=1}^{K} \sum_{x_{t_1}, x_{t_2} \in C_k} \sum_{j=1}^{p} \frac{1}{n_k} d(x_{t_1 j}, x_{t_2 j}) \\ &= \frac{1}{p} \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{x_{t_1}, x_{t_2} \in C_k} \frac{1}{n_k} |x_{t_{1j}} - x_{t_{2j}}| \end{aligned}$$

$$= \frac{1}{p}\sum_{k=1}^{K}\sum_{j=1}^{p}\frac{1}{n_k}n_k p_k^j n_k(1-p_k^j)$$

$$= \frac{1}{p}\sum_{k=1}^{K}\sum_{j=1}^{p}n_k p_k^j(1-p_k^j) \tag{5}$$

where p_k^j is the probability that the jth attribute is 1 in cluster k.

Havrda and Charvat (1967) proposed a generalized entropy of degree s for a discrete probability distribution $P = (p_1, p_2, \ldots, p_n)$

$$H^s(P) = (2^{(1-s)} - 1)^{-1}\left(\sum_{i=1}^{n} p_i^s - 1\right), \quad s > 0, s > 1$$

and

$$\lim_{s\to 1} H^s(P) = -\sum_{i=1}^{n} p_i \log p_i.$$

Degree s, as a scalar parameter, appears on the exponent in the expression equation and controls the sensitivity of the uncertainty calculation.

In the case $s = 2$, then

$$H^2(P) = -2\left(\sum_{i=1}^{n} p_i^2 - 1\right). \tag{6}$$

Proposition 3. *With the generalized entropy defined in Equation (6),* $D(C) = \frac{1}{n}\sum_{k=1}^{K} n_k \hat{H}(C_k)$.

Proof: Note that, with the generalized entropy defined in Equation (6), we have

$$\frac{1}{n}\sum_{k=1}^{K} n_k \hat{H}(C_k) = -\frac{2}{n}\sum_{k=1}^{K}\sum_{j=1}^{p} n_k[(p_k^j)^2 + (1-p_k^j)^2 - 1]$$

$$= \frac{2}{n}\sum_{k=1}^{K}\sum_{j=1}^{p} n_k p_k^j(1-p_k^j)$$

$$= \frac{2p}{n}D(C)(\text{Based on Equation (5).})$$

□

Thus we have established the connections between the entropy-criterion and the dissimilarity coefficients.

3.3. Entropy and mixture models

In this section, we show that the entropy-based clustering criterion can be formally derived using the likelihood principle based on Bernoulli mixture models. The basic

idea of the mixture model is that the observed data are generated by several different latent classes (McLachlan & Peel, 2000). In our setting, the observed data, characterized by the $\{0, 1\}^p$ valued data vectors, can be viewed as a mixture of multivariate Bernoulli distributions. In general, there will be many data points: $\mathcal{X} = \{x_t\}_{t=1}^n$. Each x_t is a p-dimensional binary vector denoted as $x_t = (x_{t1}, x_{t2}, \ldots, x_{tp})$. Viewing $\{x_t\}_{t=1}^n$ as sample values of a random vector whose probability distribution function is:

$$\begin{aligned} p(x_t) &= \sum_i \pi(i) p(x_t|i) \\ &= \sum_i \pi(i) \prod_{j=1}^{p} [a_i^j]^{x_{tj}} [1 - a_i^j]^{(1-x_{tj})} \end{aligned}$$

where $\pi(i)$ denotes the probability of selecting the ith latent class and $\sum_i \pi(i) = 1$. a_i^j gives the probability that attribute j is exhibited in class i. Let $\mathbf{a}_i = (a_i^1, \ldots, a_i^p)$, $i = 1, \ldots, K$ and use $\mathbf{a}$ to denote the parameters $\mathbf{a}_i$, $i = 1, \ldots, K$.

Maximum likelihood and classification likelihood Recall that for Maximum Likelihood Principle, the best model is the one that has the highest likelihood of generating the observed data. In the mixture approach, since the data points are independent and identically distributed, the maximum likelihood of getting the entire sample $\mathcal{X}$ can be expressed as:

$$\begin{aligned} L(\mathbf{a}) &= \log p(\mathcal{X}|\mathbf{a}) = \log \prod_{t=1}^{n} \pi p(x_t|\mathbf{a}) \\ &= \sum_{t=1}^{n} \log \left(\sum_{i=1}^{k} \pi_i p(x_t|\mathbf{a_i}) \right) \\ &= \sum_{t=1}^{n} \log \left(\sum_{i=1}^{K} \pi_i \prod_{j=1}^{p} [a_i^j]^{x_{tj}} [1 - a_i^j]^{(1-x_{tj})} \right) \end{aligned}$$

We introduce auxiliary vectors $u_t = (u_t^i, i = 1, \ldots, K)$ which indicate the origin/generation of the points: u_t^i is equal to 1 or 0 accordingly as x_t comes from the cluster C_i or not. These vectors are the missing variables. The classification likelihood (Symons, 1981) is then:

$$\begin{aligned} CL(\mathbf{a}, \mathbf{u}) &= \sum_{t=1}^{n} \sum_{i=1}^{K} u_t^i \log p(x_t|\mathbf{a_i}) \\ &= \sum_{t=1}^{n} \sum_{i=1}^{K} u_t^i \log \prod_{j=1}^{p} [a_i^j]^{x_{tj}} [1 - a_i^j]^{(1-x_{tj})} \end{aligned} \tag{7}$$

Note that

$$CL(\mathbf{a}, \mathbf{u}) = L(\mathbf{a}) - LP(\mathbf{a}, \mathbf{u})$$

where

$$LP(\mathbf{a}, \mathbf{u}) = -\sum_{t=1}^{n}\sum_{i=1}^{K} u_t^i \log\left(\frac{\pi_i\, p(x|\mathbf{a_i})}{\sum_{j=1}^{K}\pi_j\, p(x|\mathbf{a_j})}\right).$$

and $L(\mathbf{a})$ is given in Equation 7. Observe that $LP(\mathbf{a}, \mathbf{u}) \geq 0$ and it can be thought as corresponding to the logarithm of the probability of the partition induced by $\mathbf{u}$. Hence the classification likelihood is the standard maximum likelihood penalized by a term measuring the quality of the partition.

Maximizing the likelihood It holds that

$$\begin{aligned}
CL(\mathbf{a}, \mathbf{u}) &= \sum_{t=1}^{n}\sum_{i=1}^{K} u_t^i \log p(x_t|\mathbf{a_i}) \\
&= \sum_{t=1}^{n}\sum_{i=1}^{K} u_t^i \log \prod_{j=1}^{p}[a_i^j]^{x_{tj}}[1-a_i^j]^{(1-x_{tj})} \\
&= \sum_{i=1}^{K} \log \prod_{t\in C_i}\prod_{j=1}^{p}[a_i^j]^{x_{tj}}[1-a_i^j]^{(1-x_{tj})} \\
&= \sum_{i=1}^{K}\sum_{j=1}^{p}(N_k^{j1}\log a_i^j + N_k^{j0}\log[1-a_i^j]) \qquad (8)
\end{aligned}$$

Proposition 4. *Maximizing $CL(\mathbf{a}, \mathbf{u})$ in Equation 8 is equivalent to maximizing $O(C)$ in Equation 1.*

Proof: If $\mathbf{u}$ is fixed, maximizing $CL(\mathbf{a}, \mathbf{u})$ over $\mathbf{a}$ is then reduced to, for each $i = 1, \ldots, K$; $j = 1, \ldots, p$, choosing a_i^j to maximize $CL_{ij}(a_i^j) = N_k^{j1}\log a_i^j + N_k^{j0}\log[1-a_i^j]$.

Since $0 < a_i^j < 1$ and $N_k^{j0} + N_k^{j1} = n_k$, we have

$$\begin{aligned}
&\frac{\partial CL_{ij}}{\partial a_i^j} = 0 \\
\iff\ &\frac{N_k^{j1}}{a_i^j} - \frac{N_k^{j0}}{1-a_i^j} = 0 \\
\iff\ &\left(N_k^{j1} + N_k^{j0}\right)a_i^j = N_k^{j1} \\
\iff\ &a_i^j = \frac{N_k^{j1}}{n_k}.
\end{aligned}$$

Observe that $\frac{\partial^2(CL_{ij})}{\partial(a_i^j)^2} < 0$. By plugging $a_i^j = \frac{N_k^{j1}}{n_k}$, we have

$$CL(\mathbf{a}, \mathbf{u}) = -\sum_{i=1}^{K} n_i \hat{H}(C_i) \qquad (9)$$

Given a dataset, n, p, and $\hat{H}(X)$ are then fixed. Hence the criterion $CL(\mathbf{a}, \mathbf{u})$ is then equivalent to $O(C)$ in Equation 1 since both of them aim at minimizing the expected entropy over the partition. □

Note that a_i can be viewed as a "center" for the cluster C_i.

3.4. A matrix perspective

Recently, a number of authors (Anod & Lee, 2001; Dhillon & Modha, 2001; Li et al., 2004a; Soete & Douglas Carroll, 1994; Xu & Gong, 2004; Xu et al., 2003; Zha et al., 2001; Dhillon et al., 2003) have suggested clustering methods based on matrix computations and have demonstrated good performance on various datasets. These methods are attractive as they utilize many existing numerical algorithms in matrix computations. In our following discussions, we use the cluster model for binary data clustering based on a matrix perspective presented in (Li et al., 2004a; Li & Ma, 2004). In the cluster model, the problem of clustering is formulated as matrix approximations and the clustering objective is minimizing the approximation error between the original data matrix and the reconstructed matrix based on the cluster structures. In this section, we show the relations between the matrix perspective and other clustering approaches.

Introduction In Li et al. (2004a) and Li and Ma (2004), a new cluster model is introduced from a matrix perspective. Given the dataset $X = (x_{ij})_{n\times p}$, the cluster model is determined by two sets of coefficients: data coefficients $D = (d_{ij})$ and feature coefficients $F = (f_{ij})$. The data (respectively, feature) coefficients denote the degree to which the corresponding data (respectively, feature) is associated with the clusters. Note that X can be viewed as a subset of R^p as well as a member of $R^{n\times p}$. Suppose X has k clusters. Then the data (respectively, feature) coefficients can be represented as a matrix $D_{n\times k}$ (respectively $F_{p\times k}$) where d_{ij} (f_{ij}) indicates the degree to which data point i (respectively, feature i) is associated with cluster j.

Given representation (D, F), basically, D denotes the likelihood of data points associated with clusters and F indicates the feature representations of clusters. It should be noted that the number of clusters is determined by the number of columns of D (or F). The ijth entry of DF^T then indicates the possibility that the jth feature will be present in the ith instance, computed by the dot product of the ith row of D and the jth row of F. Hence after thresholding operations[3], DF^T Sour can be interpreted as the approximation of the original data X. The goal is then to find a D and F that minimizes the squared error between X and its approximation DF^T.

$$\arg\min_{D,F} O(D, F) = \frac{1}{2}||X - DF^T||_F^2, \tag{10}$$

where $\|X\|_F$ is the Frobenius norm of the matrix X, i.e., $\sqrt{\sum_{i,j} x_{ij}^2}$. With the formulation in Equation 10, we transform the data clustering problem into the computation of D and F that minimize the criterion O.

Matrix perspective and dissimilarity coefficients Given the representation (D, F), basically, D denotes the assignments of data points associated into clusters and F

indicates the feature representations of clusters. Observe that

$$\begin{aligned} O(D,F) &= \frac{1}{2}||X - DF^T||_F^2 \\ &= \frac{1}{2}\sqrt{\sum_{i,j}(x_{ij} - (DF^T)_{ij})^2} \\ &= \frac{1}{2}\sqrt{\sum_i \sum_j |x_{ij} - (DF^T)_{ij}|^2} \\ &= \frac{1}{2}\sqrt{\sum_k^K \sum_{i \in C_k} |x_{ij} - e_{kj}|^2} \\ &= \frac{1}{2}\sqrt{\sum_k^K \sum_{i \in C_k} d(x_i, e_k)}, \end{aligned} \tag{11}$$

where $e_k = (f_{k1}, \ldots, f_{kr}), i = 1, \ldots, K$ is the cluster "representative" of cluster C_i. Thus minimizing Equation 11 is the same as minimizing Equation 4 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 the number of their mismatches, which is the numerator of their dissimilarity coefficients.

3.5. Minimum description length

Minimum Description length (MDL) aims at searching for a model that provides the most compact encoding for data transmission (Rissanen, 1978) and is conceptually similar to minimum message length (MML) (Oliver & Baxter, 1994; Baxter & Oliver, 1994) and stochastic complexity minimization (Rissanen, 1989). The MDL approach can be viewed in the Bayesian perspective (Mumford, 1996; Mitchell, 1997): 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 3.4, in matrix perspective, the original matrix X can be approximated by the matrix product of DF^T. It should be noted that the number of clusters is determined by the number of columns of D (or F). 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 can be 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 X 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 $L(X|DF^T)$.

Proposition 5. *Minimizing $L(X|DF^T)$ is equivalent to minimizing $O(D, F) = \frac{1}{2}||X - DF^T||_F^2$.*

Proof: Use $\hat{X}$ to denote the data matrix generated 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 $X_{ij} = b$ conditioned upon the generated data $(\hat{x})_{ij}$, via DF^T, is c. Note that

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

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

$$H(X|\hat{X}(D, F)) = -\sum_{i,j \in \{0,1\}} p_{ij} \log p_{ij} \tag{12}$$

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 $O(D, F) = \frac{1}{2}||X - DF^T||_F^2$. □

4. Experiments

In this section, we present several experiments to show that the relationships described in the paper is also observed in practice.

4.1. Methods

The relationships among entropy, mixture models as well as minimum description length have been experimentally studied and evaluated in the machine learning literature (Mitchell, 1997; Celeux & Soromenho, 1996; Cover & Thomas, 1991). Here we conduct

Table 3 Document Data Set Descriptions.

Datasets	# documents	# classes
CSTR	476	4
WebKB4	4199	4
WebKB	8,280	7
Reuters	2,900	10

experiments to compare the following criteria: entropy, dissimilarity coefficient, and the matrix perspective. To minimize the entropy criterion defined in Equation 1, we use the optimization procedure introduced in (Li et al.,2004b). For minimizing the dissimilarity coefficient criterion defined in Equation 4, we use the popular K-means algorithm (Jain & Dubes, 1988). For the matrix perspective, we use the clustering method described in (Li & Ma, 2004).

4.2. Datasets

We perform our experiments on document datasets. In our experiments, documents are represented using binary vector-space model where each document is a binary vector in the term space and each element of the vector indicates the presence of the corresponding term.

We use a variety of datasets, most of which are frequently used in the data mining information retrieval research. The range of the number of classes is from 4 to 10 and the range of the number of documents is from 476 to 8280, which seem varied enough to obtain good insights on the comparison.

The descriptions of the datasets are listed as follows and their characteristics are summarized in Table 3.

- **CSTR:** This is the dataset of the abstracts of technical reports published in the Department of Computer Science at the University of Rochester between 1991 and 2002. The TRs are available at *http://www.cs.rochester.edu/trs*. It has been first used in (Li et al., 2003) for text categorization. The dataset contained 476 abstracts, which were divided into four research areas: Natural Language Processing (NLP), Robotics/Vision, Systems, and Theory.
- **WebKB:** The WebKB dataset contains webpages gathered from university computer science departments. There are about 8280 documents and they are divided into 7 categories: student, faculty, staff, course, project, department and other. The raw text is about 27MB. Among these 7 categories, student, faculty, course and project are the four most populous entity-representing categories. The associated subset is typically called **WebKB4**. In this paper, we perform experiments on both 7-category and 4-category datasets.
- **Reuters:** The Reuters-21578 Text Categorization collection contains documents collected from the Reuters newswire in 1987. It is a standard text categorization benchmark and contains 135 categories. In our experiments, we use a subset of the data collection which include the 10 most frequent categories.

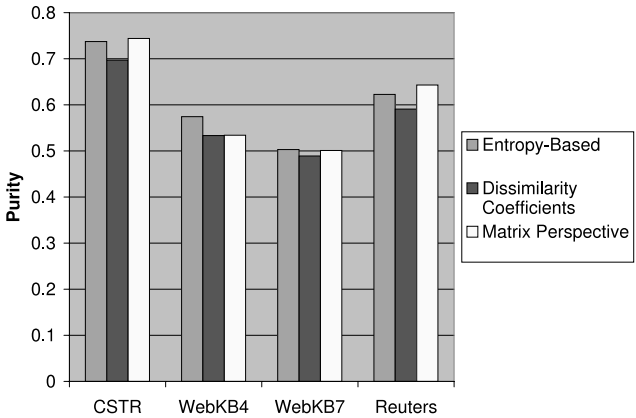

Figure 2. Purity comparison for various clustering criteria.

4.3. Result analysis

To pre-process the datasets, we remove the stop words using a standard stop list and perform stemming using a porter stemmer. All HTML tags are skipped and all header fields except subject and organization of the posted article are ignored. In all our experiments, we first select the top 200 words by mutual information with class labels. The feature selection is done with the rainbow package (McCallum, 1996).

All the datasets we use are standard labelled corpus and we can use the labels of the dataset as the objective knowledge to evaluate clustering. Since the goal of the experiments is to empirically reveal the relationships among the clustering criteria, hence we use the *purity* as the performance measure. Purity is an external subjective evaluation measure and it is intuitive to understand. The *purity* (Zhao & Karypis, 2001) aims at measuring the extent to which each cluster contained data points from primarily one class. The purity of a clustering solution is obtained as a weighted sum of individual cluster purities and is given by $Purity = \sum_{i=1}^{K} \frac{n_i}{n} P(S_i)$, $P(S_i) = \frac{1}{n_i} max_j(n_i^j)$ where S_i is a particular cluster of size n_i, n_i^j is the number of points of the ith input class that were assigned to the jth cluster, K is the number of clusters and n is the total number of points[4]. If two criteria are essentially optimizing equivalent functions up to some extent, then they should lead to similar clustering results and have similar purity values.

Figure 2 shows the purity comparisons of various criteria. The results are obtained by averaging 10 trials. We can observe that they lead to similar clustering results. For example, on *CSTR* dataset, the purity values for entropy-based criterion, dissimilarity coefficients and matrix perspective are 0.730, 0.697 and 0.740 respectively. On *WebKB4* dataset, the values are 0.574, 0.533 and 0.534 respectively. On *WebKB7* dataset, the values are 0.503, 0.489 and 0.501. On *Reuters* dataset, they are 0.622, 0.591 and 0.643. In summary, the purity results among the clustering criteria are close and the maximum difference is less than 4%. The differences are resulted from the subtle distinctions among various criteria as well as the inherent random nature of stochastic learning methods. The experimental study provides the empirical evidence on the relationships among various clustering criteria.

5. Conclusion

Binary data have been occupying a special place in the domain of data analysis. In this paper, we aim to provide a unified view of binary data clustering by examining the connections among various clustering criteria. In particular, we show the relationships among the entropy criterion, dissimilarity coefficients, mixture models, the matrix decomposition, and minimum description length.

The unified view provides an elegant base to compare and understand various clustering methods. In addition, the connections can provide many insights and motivations for designing binary clustering algorithms. For example, the equivalence between the information theoretical criterion and the maximum likelihood criterion suggests a way to assess the number of clusters when using the entropy criterion: to look at various techniques used in model-based approaches such as likelihood ratio tests, penalty methods, Bayesian methods, cross-validation (Biernacki & Govaert, 1997; Smyth, 1997). Moreover, the connections motivate us to explore the integration of various clustering methods.

Acknowledgment

The author is grateful to Dr. Shenghuo Zhu, Dr. Sheng Ma and Dr. Mitsunori Ogihara for their helpful discussions and suggestions. The author would also like to thank the anonymous reviewers for their invaluable comments. The work is partially supported by a 2005 IBM Faculty Award and a 2005 IBM Shared University Research(SUR) Award.

Notes

1. We adopt the convention that $0log0 = 0$ if necessary.
2. This can be found in many popular coefficients such as Jaccard's coefficients and Dice coefficients.
3. Thresholding operations make sure that each entry of DF^T is either 0 or 1. It can be performed as: if $DF^T_{ij} > \frac{1}{2}$, then set $DF^T_{ij} = 1$, otherwise $DF^T_{ij} = 0$, for i, j.
4. $P(S_i)$ is also called the individual cluster purity.

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conferenceon Very Large Data Bases (VLDB'94)* (pp. 487–499). Morgan Kaufmann Publishers.

Ando, R. K., & Lee, L. (2001). Iterative Residual Rescaling: An analysis and generalization of {LSI}. In *Proceedings of the 24th {SIGIR}* (pp. 154–162).

Barbara, D., Li, Y., & Couto, J. (2002). COOLCAT: An entropy-based algorithm for categorical clustering. *Proceedings of the eleventh international conference on Information and knowledge management (CIKM'02)* (pp. 582–589). ACM Press.

Baulieu, F.B. (1997). Two variant axiom systems for presence/absence based dissimilarity coefficients. *Journal of Classification*, *14*, 159–170.

Baxter, R.A., & Oliver, J.J. (1994). *MDL and MML: similarities and differences* (Technical Report 207). Monash University.

Biernacki, C., & Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics* (pp. 451–457).

Bock, H.-H. (1989). Probabilistic aspects in cluster analysis. In O. Opitz (Ed.), *Conceptual and numerical analysis of data*, (pp. 12–44). Berlin: Springer-verlag.

Celeux, G., & Govaert, G. (1991). Clustering criteria for discrete data and latent class models. *Journal of Classification*, *8*, 157–176.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*, 195–212.

Cover, T.M., & Thomas, J.A. (1991). *Elements of information theory*. John Wiley & Sons.

Dhillon, I.S., Mallela, S., & Modha, S.S. (2003). Information-theoretic co-clustering. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2003)* (pp. 89–98). ACM Press.

Dhillon, I.S., & Modha, D.S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, *42*, 143–175.

Ganti, V., Gehrke, J., & Ramakrishnan, R. (1999). CACTUS: Clustering categorical data using summaries. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'99)* (pp. 73–83). ACM Press.

Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Clustering categorical data: An approach based on dynamical systems. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB'98)* (pp. 311–322). Morgan Kaufmann Publishers.

Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, *25*, 345–366.

Gyllenberg, M., Koski, T., & Verlaan, M. (1997). Classification of binary vectors by stochastic complexity. *Journal of Multivariate Analysis*, *63*, 47–72.

Hartigan, J.A. (1975). *Clustering algorithms*. Wiley.

Havrda, J., & Charvat, F. (1967). Quantification method of classification processes: Concept of structural a-entropy. *Kybernetika*, *3*, 30–35.

Huang, Z. (1998). Extensions to the *k*-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*, 283–304.

Jain, A.K., & Dubes, R.C. (1988). *Algorithms for clustering data*. Prentice Hall.

Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy*. John Wiley & Sons.

Kaufman, L., & Rousseeuw, P.J. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley.

Li, T. (2005). A general model for clustering binary data. *Proceedings of Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2005)* (pp. 188–197).

Li, T., & Ma, S. (2004). IFD: iterative feature and data clustering. *Proceedings of the 2004 SIAM International conference on Data Mining (SDM 2004)* (pp. 472–476). SIAM.

Li, T., Ma, S., & Ogihara, M. (2004a). Document clustering via adaptive subspace iteration. *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)* (pp. 218–225).

Li, T., Ma, S., & Ogihara, M. (2004b). Entropy-based criterion in categorical clustering. *Proceedings of The 2004 IEEE International Conference on Machine Learning (ICML 2004)* 536–543.

Li, T., Zhu, S., & Ogihara, M. (2003). Efficient multi-way text categorization via generalized discriminant analysis. *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM 2003)* (pp. 317–324). ACM Press.

McCallum, A.K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. John Wiley.

Mitchell, T.M. (1997). *Machine learning*. The McGraw-Hill Companies, Inc.

Mumford, D. (1996). Pattern Theory: A Unifying Perspective. 25–62.

Oliver, J.J., & Baxter, R.A. (1994). *MML and Bayesianism: similarities and differences* (Technical Report 206). Monash University.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific Press.

Roberts, S., Everson, R., & Rezek, I. (1999). Minimum entropy data partitioning. *Proc. International Conference on Artificial Neural Networks* (pp. 844–849).

Roberts, S., Everson, R., & Rezek, I. (2000). Maximum certainty data partitioning. *Pattern Recognition*, *33*, 833–839.

Smyth, P. (1996). Clustering using monte carlo cross-validation. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (SIGKDD 1996)* (pp. 126–133).

Soete, G.D., & douglas Carroll, J. (1994). K-means clustering in a low-dimensional euclidean space. In *New approaches in classification and data analysis*, 212–219. Springer.

Symons, M.J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics*, *37*, 35–43.

Xu, W., & Gong, Y. (2004). Document clustering by concept factorization. *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval* (pp. 202–209). Sheffield, United Kingdom: ACM Press.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval((SIGIR'03))* (pp. 267–273). ACM Press.

Zha, H., He, X., Ding, C., & Simon, H. (2001). Spectral relaxation for k-means clustering. *Proceedings of Neural Information Processing Systems* (pp. 1057–1064).

Zhao, Y., & Karypis, G. (2001). *Criterion functions for document clustering: Experiments and analysis* (Technical Report). Department of Computer Science, University of Minnesota.

Received March 14, 2005
Final Revision September 30, 2005
Accepted October 3, 2005