Entropy based principle and generalized contingency tables

Vincent Vigneron

ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

ENTROPY-BASED PRINCIPLE AND GENERALIZED CONTINGENCY TABLES

Vincent Vigneron

MATISSE-SAMOS CNRS UMR 8595,
90 rue de Tolbiac,
75634 Paris cedex 13, France.
Email: vigneron@univ-paris1.fr

Abstract. It is well known that the entropy-based concept of mutual information provides a measure of dependence between two discrete random variables. There are several ways to normalize this measure in order to obtain a coefficient similar *e.g.* to Pearson's coefficient of contingency. This paper presents a measure of independence between categorical variables and is applied for clustering of multidimensional contingency tables. We propose and study a class of measures of directed discrepancy. Two factors make our divergence function attractive: first, the coefficient we obtain a framework in which a *Bregman divergence* can be used for the objective function ; second, we allow speciafication of a larger class of constraints that preserves varous statistics.

1 Formulation and analysis

Clustering is the problem of partitoning a finite set of points in a multidimensional space into classes (called *clusters*) so that points belonging to the same class are *similar*. An important step in designing a clustering technique is defining a way to measure the quality of partitioning in terms of the above objective. Given such a measure, an appropriate partition can be computed by optimizing some quantity (*e.g.* the sum of the distances of the points to the cluster centroids). However, if the data vectors contain *categorical variables*, geometric approaches are inappropriate and other strategies have to be found [1]. This is often the case in applications where the data are described by binary attributes.
Many algorithms have been designed for clustering analysis of categorical data [2, 3, 4, 5]. For instance, entropy-type metrics for *similarity* among objects have been developped from early on. In this paper, we address the following 2 questions: (i) what class of discrepancy function admit efficient clustering algorithms ? (ii) how to vizualize the classes and the explanatory variables ?
In this paper, we show that an entropy-based clustering criterion can be formally derived from the heterogeneity of clusters and interpreted as a Bregman measure. Bregman information principle generalizes the maximum entropy principle.

Definition 1 (Bregman divergence) *Let ϕ be a real valued strictly convex function defined on the convex set $S \subseteq \mathbb{R}$, the domain of ϕ such that ϕ is differentiable on int(S) the interior of S. The Bregman divergence $\mathcal{B}_\phi : S \times int(S) \mapsto \mathbb{R}^+$ is defined as $\mathcal{B}_\phi(z_1, z_2) = \phi(z_1) - \phi(z_2) - (z_1 - z_2, \nabla\phi(x_2))$, where $\nabla\phi$ is the gradient of ϕ.*

For instance, let $\phi(z) = a \log z$. For $z_1, z_2 \in \mathbb{R}^+, \mathcal{B}_\phi(z_1, z_2) = z_1 \log(z_1/z_2) - (z_1 - z_2)$. Based on Bregman divergences, it is possible to define as useful concept of *Bregman information* which captures the information in a random variable and, hence, formulate the clustering problem in these terms. This paper is not (directly) concerned in mumerical estimates of multidimensional entropy such as sample-spacings, kernel density plug-in estimates, splitting data estimates, etc.
The rest of the paper is organized as follows: section 2 set down notations and shows the equivalence between the entropy-based criterion and Bregman divergence, section 3 formulates the problem of categorical clustering of variables, section 4 establishes presents our experimental results.

2 Maximum entropy and minimum Chi-square

Consider the general class of measures of directed divergence $D(p\|q) = \sum_{i=1}^{n} f(p_i, q_i)$ where $p = \{p_i\}, q = \{q_i\}$ are probabilities sets. Important class of such measures is given by $D(p\|q) = \sum_{i=1}^{n} q_i f(\frac{p_i}{q_i}), q_i > 0$ where f is twice differentiable and a strictly convex function. When $f(x) = -x \ln x, f'(x) = 1 + \ln x, f''(x) = \frac{1}{x} > 0$ if $x > 0$. Accordingly, $D(p\|q) = \sum_{i=1}^{n} q_i \frac{p_i}{q_i} \ln(\frac{p_i}{q_i}) = \sum_i p_i \ln(\frac{p_i}{q_i})$. This is the *Kullback-Leibler measure of directed divergence*. This measure is non-negative and vanishes iff $q_i = p_i, \forall i$[1].
Let J and I two finite sets indexing two categorical variables and let M be a $I \times J$ table of frequencies (Tab. 2). Let f_{ij} be the frequency in the cell in the ith row and jth column of an $m \times n$ contingency table and let $f_J = (f_{\cdot j})_{j \in J}$ and $f_I = (f_{i\cdot})_{i \in I}$ be the sums of elements in the ith row and jth column respectively, *i.e.* $p_{ij} = \frac{f_{ij}}{f_0}, p_{i\cdot} = \frac{f_{i\cdot}}{f_0}, p_{\cdot j} = \frac{f_{\cdot j}}{f_0}$, where $f_0 = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{m} f_{i\cdot} = \sum_{j=1}^{n} f_{\cdot j}$.

f_{11}	f_{12}	$\cdots$	f_{1n}	$f_{1\cdot}$
f_{21}	f_{22}	$\cdots$	f_{2n}	$f_{2\cdot}$
$\vdots$		$\ddots$		
f_{m1}	f_{m2}	$\cdots$	f_{mn}	$f_{m\cdot}$
$f_{\cdot 1}$	$f_{\cdot 2}$	$\cdots$	$f_{\cdot n}$	f_0

$\rightarrow$

p_{11}	p_{12}	$\cdots$	p_{1n}	$p_{1\cdot}$
p_{21}	p_{22}	$\cdots$	p_{2n}	$p_{2\cdot}$
$\vdots$		$\ddots$		
p_{m1}	p_{m2}	$\cdots$	p_{mn}	$p_{m\cdot}$
$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot n}$	1

Table 1: $m \times n$ contingency tables.

From elementary courses in statistics, we know that for any contingency table with given row and column sums, the maximum entropy value of $S_{12} = -\sum_i^m \sum_j^n \frac{f_{ij}}{f_0} \ln(\frac{f_{ij}}{f_0})$ $= \frac{1}{f_0}(f_0 \ln f_0 - \sum_i^m \sum_j^n f_{ij} \ln f_{ij})$ is obtained when $f_{ij} = \frac{f_{i\cdot} f_{\cdot j}}{f_0}$ or $p_{ij} = p_i p_j$, so that $\max S_{12} = -\sum_{i=1}^{m} \sum_{j=1}^{n} p_{i\cdot} p_{\cdot j} \ln p_{i\cdot} p_{\cdot j} = S_1 + S_2$. This shows that $S_{12} \leq S_1 + S_2$. The non-negative quantity $S_{12} - S_1 - S_2$ can therefore be regarded as a measure of the dependence of the 2 attributes. Now,

$$S_{12} - S_1 - S_2 = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \ln \frac{p_{ij}}{p_{i\cdot} p_{\cdot j}} \tag{1}$$

can also be interpreted in terms of Kullback-Leibler's measure of directed divergence. Let us find its value for a small departure from independence e_{ij}. Let $p_{ij} = p_{i\cdot} p_{\cdot j} + e_{ij}$, then from (1),

$$S_1 + S_2 - S_{12} = \sum_{j=1}^{p} \sum_{i=1}^{n} f_{i\cdot} f_{\cdot j} \ln \left(1 + \frac{e_{ij}}{f_{i\cdot} f_{\cdot j}}\right) + \sum_{j=1}^{p} \sum_{i=1}^{n} e_{ij} \ln \left(1 + \frac{e_{ij}}{f_{i\cdot} f_{\cdot j}}\right) \tag{2}$$

Using Taylor's development of $\ln(1 + x)$, we have:

$$S_1 + S_2 - S_{12} = \sum_{j,i} \left[e_{ij} - \frac{e_{ij}^2}{2 f_{i\cdot} f_{\cdot j}} + \frac{e_{ij}^3}{3(f_{i\cdot} f_{\cdot j})^2}\right] + \sum_{j,i} \left[\frac{e_{ij}^2}{f_{i\cdot} f_{\cdot j}} - \frac{e_{ij}^3}{2(f_{i\cdot} f_{\cdot j})^2}\right] + \dots \tag{3}$$

[1]When $0 < \alpha < 1, \sum_{i=1}^{n} q_i^{1-\alpha} p_i^{\alpha}$ is a concave function and so its logarithm is also a concave function. We cal use $\frac{1}{1-\alpha} \sum_{i=1}^{n} q_i^{1-\alpha} p_i^{\alpha}, 0 < \alpha < 1$ as a measure of discrepancy. This measure was suggested by Renyi in 1961.

where we have omitted $\sum_{j,i} \frac{e_{ij}^4}{(f_{i\cdot}f_{\cdot j})^3}, \sum_{j,i} \frac{e_{ij}^5}{(f_{i\cdot}f_{\cdot j})^4}, \ldots$ Now, $\sum_{j,i} e_{ij} = \sum_{j,i}(f_{ij} - f_{i\cdot}f_{\cdot j}) = 0$, so that up to this order of approximation:

$$S_1 + S_2 - S_{12} \approx \sum_{j,i} \left[\frac{e_{ij}^2}{2f_{i\cdot}f_{\cdot j}} - \frac{e_{ij}^3}{6(f_{i\cdot}f_{\cdot j})^2} \right] = \sum_{j,i} \left[\frac{(f_{ij} - f_{i\cdot}f_{\cdot j})^2}{2f_{i\cdot}f_{\cdot j}} - \frac{(f_{ij} - f_{i\cdot}f_{\cdot j})^3}{6(f_{i\cdot}f_{\cdot j})^2} \right] \quad (4)$$

In (4), as such upto a first approximation, $S_1 + S_2 - S_{12} = \sum_{j,i} \frac{(f_{ij} - f_{i\cdot}f_{\cdot j})^2}{2f_{i\cdot}f_{\cdot j}} = \frac{1}{2}\chi^2$. The above proof gives an interesting interpretation for the Chi-square which is now seen to represent twice the difference between the observed and the maximum entropy. This shows that Chi-square is intimately connected with entropy maximization despite many lamentations of statisticians that Chi-square does not represent anything meaningful. Good [6] gave a comprehensive discussion of the use of maximum entropy principle in the case of multidimensional contingency tables. Tribus [7] brought out the relationship between Chi-square test and maximization of entropy in contingency tables.

A measure of divergence (or deviation to independence) can be derived from (3) if we observe that $\Delta S = S_1 + S_2 - S_{12} = \sum_{i,j} \sum_k^\infty \frac{(-1)^k}{k(k-1)} p_{i\cdot}p_{\cdot j} \left(\frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{i\cdot}p_{\cdot j}} \right)^k$. Now, $d_{IJ} = \sum_{i,j} \sum_{k=1}^\infty p_{i\cdot}p_{\cdot j} \frac{(-1)^k x^k}{k(k-1)}$, where $\sum_k^\infty \frac{(-1)^k x^k}{k(k-1)}$ is the infinite series of the second derivative of the function $\phi(x) = \frac{1}{1+x}$. A primitive of ϕ is $\psi(x) = (x+1)\ln(x+1) - x$. Hence, the discrimination measure

$$\sum_{i,j} p_{ij} \ln \left(\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}} \right) - p_{ij} + p_{i\cdot}p_{\cdot j} \quad (5)$$

which is a Bregman divergence in the sense of definition 1. Fig. 1 depicts the deviation between functions $f_1(x) = \frac{x^2}{2}$ and $f_2(x) = (x+1)\ln(1+x) - x$ around zero.

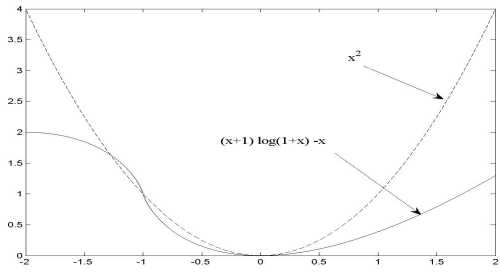

Fig. 1: Deviation between Chi-square metric and our Bregman-like divergence.

3 Generalized contingency table

3.1 Notations

We consider the situation in which N individuals answer to Q questions (variables). Each question has m_q possible answers (or modalities). The individuals answer each question q $(1 \leq q \leq Q)$ by choosing only one modality among the m_q modalities. If we assume that $Q = 3$ and $m_1 = 3$, $m_2 = 2$ and $m_3 = 3$, then an answer of an individual could be $(0, 1, 0|0, 1|1, 0, 0)$, where 1 corresponds to the chosen modality for each question. Let us denote by M the total number of all the modalities: $M = \sum_{q=1}^Q m_q$. To simplify, we can enumerate all the modalities from 1 to M and denote by

$Z_i,\ (1 \leq i \leq M)$ the column vector constructed by the N answers to the i-th modality. The k-th element of the vector Z_i is 1 or 0, according to the choice of the individual k. Let $K_{(N\times M)} = \{k_{ij}\}$ the complete disjonctive table where $k_{ij} = 1$ if the individual i chooses the modality j and 0 otherwise (see Tab.2). The marginals of the rows of K are constant and equal to the number Q of questions, *i.e.* $k_{i\cdot} = \sum_{j=1}^{M} k_{ij} = Q$. K is essential if we want to remember who answered what, but if we only have to study the *relations between the Q variables* (or questions), we can sum up the data in a crosstabulations table, called *Burt matrix*, defined by $B = K^T K$, where K^T is the transposed matrix of K (see Tab.2).

m_1			m_2		m_3			
0	1	0	0	1	0	0	0	1
0	1	0	1	0	0	0	1	0
0	0	1	1	0	0	1	0	0
1	0	0	0	1	0	0	0	1
1	0	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1	0
0	0	1	1	0	1	0	0	0
1	0	0	1	0	1	0	0	0
0	1	0	1	0	0	1	0	0
0	1	0	0	1	0	0	1	0
0	0	1	0	1	0	1	0	0
1	0	0	1	0	0	0	0	1

$\rightarrow B_{(9\times 9)} =$

4	0	1	2	2	1	0	1	2
0	5	0	2	3	0	1	3	1
0	0	3	2	1	1	2	0	0
2	2	2	6	0	1	2	1	1
2	3	1	0	6	0	1	3	2
1	0	1	2	0	2	0	1	0
0	1	2	2	1	0	3	0	0
1	3	0	1	3	0	0	4	0
2	1	0	1	2	0	0	0	3

Table 2: Left: disjunctive table $K_{(12\times 3)}$. Right: Burt table $B_{(9\times 9)}$ from $K_{(12\times 3)}$.

B is a $(M \times M)$ symmetrical matrix, composed of $Q \times Q$ blocks, such that the $(q \times r)$ block B_{qr} $(1 \leq q, r \leq Q)$ contains the N answers to the question r. The block B_{qq} is a diagonal matrix, whose diagonal entries are the numbers of individuals who have respectively chosen the modalities $1, \ldots, m_q$ for the question q. The Burt table $B_{(M\times M)}$ has to be seen as a *generalized contingency table*, when more than 2 kinds of variables are to be studied simultaneously (see [8]). In this case, we loose a part of the information about the individuals answers, but we keep the information regarding the relations between the modalities of the qualitative variables. Each row of the matrix B characterizes *a modality of a question* (or variable). Let us denote by f_{ij} the entries of the matrix B, then the total sum of all the entries of B is $b = \sum_{i,j} b_{ij} = Q^2 N$. One defines successively (i) F the table of the relative frequencies, with entry $p_{ij} = \frac{b_{ij}}{b}$ with margins $p_{i\cdot} = \sum_j p_{ij}$ and $p_{\cdot j} = \sum_i p_{ij}$, (ii) R the table of the profiles which sum to 1, with entry $R_{ij} = \frac{p_{ij}}{p_{i\cdot}}$.

3.2 Clustering row profiles

The classical multiple correspondence analysis (MCA) ([9]) is a *weighted* principal component analysis (PCA) performed on the row profiles or column-profiles of the matrix R, each row being weighted by $p_{i\cdot}$. MCA would provide a simultaneous representation of the M vectors on a low dimensional space which gives some information about the relations between the Q variables and minimize χ^2. In [5], Cottrell *et al.* consider the Euclidean distance between rows, each being weighten by $p_{i\cdot}$, to analyse multidimensional data, involving qualitative variables and feed a Kohonen map with these row vectors. We can do better: from (5), is comes that the distance between two rows $r(i)$ and $r(i')$ of the table R is "exactly" given by $d\{r(i), r(i')\} = \sum_k^{\infty} \sum_{j=1}^{M} \frac{(-1)^k}{(p_{\cdot j})^{k-1} k(k-1)} \left(\frac{p_{ij}}{p_{i\cdot}} - \frac{p_{i'j}}{p_{i'\cdot}}\right)^k$. Let $x = \left(\frac{p_{ij}}{p_{i\cdot} p_{\cdot j}} - \frac{p_{i'j}}{p_{i'\cdot} p_{\cdot j}}\right)$.

Now, $d\{r(i), r(i')\} = \sum_j p_{\cdot j} \sum_{k=1}^{\infty} \frac{(-1)^k x^k}{k(k-1)}$, which is the infinite series of the second derivative of the function $\phi(x) = \frac{1}{1+x}$. A primitive of ϕ is $\psi(x) = (x+1)\ln(x+1) - x$. Hence, the *total deviation rate* to independence of Q categorical variables comes as above from Pearson's approximation of independence:

$$d_Q = \sum_{i,i'} \sum_j p_{\cdot j} \{(\alpha_{ij} + 1)\ln(\alpha_{ij} + 1) - \alpha_{ij}\}, \tag{6}$$

with $\alpha_{ij} = \frac{p_{ij}}{p_{i\cdot} p_{\cdot j}} - \frac{p_{i'j}}{p_{i'\cdot} p_{\cdot j}}$. So it is equivalent to compute a profile matrix C whose entry is $c_{ij} = \frac{p_{ij}}{p_{\cdot j} p_{i\cdot}}$ and to consider the "distance" $d\{r(i), r(i')\}$ between its rows.
A remark has to be made at this stage: two modalities or more will be close if there is a large proportion of individuals that choose them simultaneously. We would like to get these individuals grouped in the same region.

4 Experiments

It is possible at this stage to use a Kohonen algorithm to get such a representation (for which there is no more constraint of linearity of the projection), as it has been already proposed by [10]. we propose to train a Kohonen network with these *row-profiles* as inputs and to study the resulting map to extract the relevant information about the relations between the Q. See [11] for further details on the Kohonen algorithm. The difference with the usual Kohonen algorithm sets in the search of the winner unit $\omega_0 = \arg\min_u \psi(\omega(u), c_i)$, where each unit u is represented in the R^M space by its *weight-vector* $\omega(u)$ and $c_i = (\frac{p_{1j}}{p_{\cdot j} p_{1\cdot}}, \ldots, \frac{p_{Mj}}{p_{\cdot j} p_{M\cdot}})$, among all the units of the lattice using the fonction ψ which rules now the metric space. ψ is now the Bregman measure to take advantage of the convexity of the criterion.
Using a black and white image of rice grains, one can illustrates a process on binary variables. The image I in Fig. 2 is a $(100 \times 256)-$matrix containing only $0/1$ ((pixels).

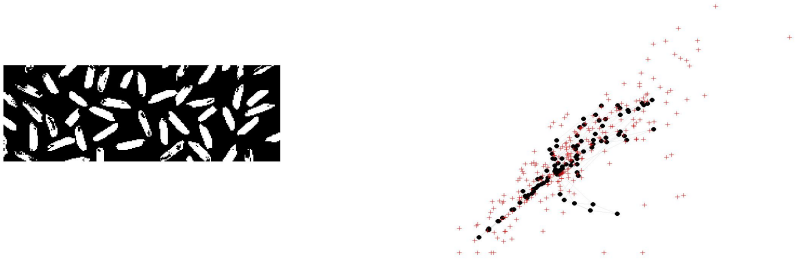

Fig. 2: Left: image of rice grains. Right : Kohonen map of columns of pixel.

To represent the columns of I in $\mathbb{R}^{256}$, we train a Kohonen network with the rows of the Burt Matrix and using the Bregman divergence (see previous section). After training, each row profile can be represented by its corresponding winner unit : in Fig. 2, '$+$' represent the pixel columns, '$\bullet$' the units of the Kohonen grid. To evaluate the effect of the Bregman divergence in the representation space, we plot in Fig. 3 the kernel-density estimation of the distributions of the distances between row-profiles of B, *i.e.* `RowProfile(i,:)` and `RowProfile(j,:)` : `Euclidean` ('–'), `Citybloc` ('$\cdots$'), `Minkowski` with $p = 4$ ('$\cdot - \cdot -$') and our `Bregman` metric ('$- \cdot - \cdot$'). Clearly, the most favourable case is the Bregman because (i) the spread of the distribution is bigger, (ii) the distribution is centered.

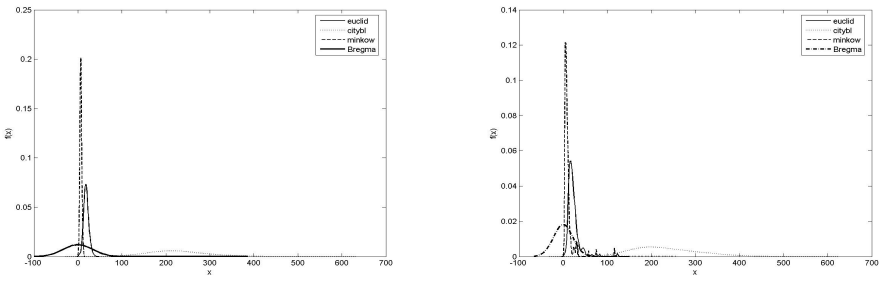

Fig. 3: kernel-density estimation of the distributions of the inter row-profiles distances. Left: row-profiles of B, left : row-profiles of R.

5 Conclusion

In this paper, we derive from the entropy-based criterion for categorical data clustering a Bregman divergence measure and illustrate its relation with other criteria. The Bregman measure is used as a metric in a Kohonen algorithm to take advantage of the convexity of the criterion. The experimental results indicates the effectiveness of the proposed method. The above formulation is applicable when the data matrix directly corresponds to an empirical joint distribution. However, there are important situation in which the data matrix is more general and may contain for instance, negative entries and a distorsion measure such as the Euclidean distance might be inappropriate.

References

[1] E.B. Andersen. *Introduction to the statistical analysis of categorical data*. Springer, 1989.

[2] Z. Huang. Extension to th k-means algorithm for clustering large data sets with categorical variables. *Data Mining and Knowledge Discovery*, 2:283–304, 1998.

[3] S. Guha, R. Rastogi, and K. Shim. ROCK : a robust clustering algorithm for categorical attributes. *Information Systems*, 23:345–366, 2000.

[4] V. Vigneron, H. Maaref, S. Lelandais, and A.P. Leitao. "Poor man" vote with m-ary non-parametric classifiers based on mutual information. application to iris recognition. In *4th AVBPA International Conference on Audio-Video Based Biometric Person Authentification*, London, june 2003.

[5] M. Cottrell, P. Letremy, and E. Roy. Analysing a contingency table with kohonen maps: a factorial correspondence analysis. In J. Cabestany, J. Mary, and A. Prieto, editors, *Proceedings of IWANN'93*, Lectures Notes in Computer Science, pages 305–311. Springer, 1993.

[6] I.J. Good. Maximum entropy for hypothesis formulation especially in multi-dimensional contingency tables. *Ann. Math. Stat.*, 34:911–934, 1965.

[7] M. Tribus. *Rational descriptions, decisions, and designs*. Pergamon Press, New York, 1979.

[8] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 1995.

[9] G. Saporta. *Probabilités, analyse de données et statistiques.* Technip, Paris, 1992.

[10] S. Ibbou and M. Cottrell. Multiple correspondence analysis of a crosstabulations matrix using the kohonen algorithm. In *Proceedings of ESANN'99*. Springer, 1999.

[11] T. Kohonen. *Self-organisation and Associative Memory*. Springer, 1989.