Assessment of Entropy Based Method of Variable Choice in Cluster Analysis

Jerzy Korzeniewski

Abstract— **In the paper, we investigate the efficiency of an algorithm for the choice of variables in cluster analysis built on the entropy approach (*Dash & Liu,* 2000). The assessment of this algorithm is carried out on synthetic data sets in the form of the mixtures of normal distributions. It turns out that the method is not working so well as the Authors of the entropy based approach suggested. The method fails in presence of correlation between masking variables.**

***Index Terms*— cluster analysis, entropy, variable choice.**

I. Introduction

It is widely acknowledged that not all variables characterising data set observations contribute the same weight to the data set cluster structure. Some are more important than other (true variables), some are less important and some may be an obstacle (masking variables) in detecting the data set cluster structure. In recent years there has been an offspring of methods designed to choose the best subset of variables describing the data set cluster structure. There are about a dozen different approaches to the task. Steinley and Brusco (2008) examined eight methods in a broad empirical experiment. The conlcusions which follow are rather negative to nearly all model based methods as the best methods turned out to be non-model approaches i.e. VS-KM method by Brusco and Cradit (2001), relative clusterability weighting with VAF selection by Steinley and Brusco (2007) and HINoV by Carmone et al. (1999). Of the three model based methods only the feature saliency method by Law et al. (2004) did relatively well. There are other methods which one could apply to the same task and which were not considered in this experiment e.g. the entropy based method by Dash and Liu (2000). This approach can be used to construct a number of algorithms to choose variables. The Authors suggest two algorithms. The first one is to calculate the entropy of all sets consisting of all variables excluding one. The variables representing these sets which have higher entropy are more likely to be true variables. The second algorithm is a wrapper approach and consists in running a *k*-means algorithm to group the data for all possible subsets of entropy-ranked variables and to assess the grouping by means of a criterion. The subject of this article is to investigate the efficiency of both algorithms.

II. ENTROPY BASED METHODS

The entropy of the set of observations $x_1,...,x_n$ is defined as (see *Dash & Liu* , 2000)

$$E(x_1,...,x_n) = \sum_{x_1 x_2 ... x_n} p(x_1,...,x_n)\log(p(x_1,...,x_n)) \qquad (1)$$

The higher the entropy the more uniform the distribution of variables, the more distinct data set cluster structure the lower the entropy. If there is a distinct cluster structure the distances between two points are either big or small – the smaller the number of medium size distances, although it is dependent on the very structure. Let us assume that the two point distances have been normalized separately on each variable by means of dividing the distance by the maximum distance for a given variable. The eentropy of two observations being at distance d from each other can be approximated in the following way

$$E(x_1, x_2) = -d\log d - (1-d)\log(1-d) \qquad (2)$$

so that the maximum value of 1 the entropy would assume for mean distance i.e. d=0.5 , while the minimum value of 0 the entropy would assume for the smallest possible distance i.e. d=0 and for the biggest possible distance i.e. d=1. Thus, the entropy of the whole data set is equal to

$$E = -\sum_{x_1, x_2}\left[d(x_1,x_2)\log d(x_1,x_2) + (1-d(x_1,x_2))\log(1-d(x_1,x_2))\right] \qquad (3)$$

where the summation is over all pairs of data set observations. Subsequantly, we switch from distances $d(x_1, x_2)$ between observations to similarities between observations $S(x_1, x_2)$ assuming values from interval $[0,1]$ by means of formula

$$S(x_1, x_2) = \exp(-\alpha \cdot d(x_1, x_2)), \qquad (4)$$

where α is such that the arithmetic mean of all pairwise distances would correspond to similarity 0.5 i.e. $0.5 = \exp(-\alpha \cdot \bar{d})$. Switching from distances to similarities results in entropy being low if similarity is either high or low i.e. close to 0 or 1. If similarity is of medium value i.e. in the neighbourhood of 0.5 the entropy is high. Thus, the entropy of the whole data set will be given by the formula

$$E = -\sum_{x_1, x_2}\left[S(x_1,x_2)\log S(x_1,x_2) + (1-S(x_1,x_2))\log(1-S(x_1,x_2))\right] \qquad (5)$$

The first algorithm we want to investigate consists in comparing all entropies corresponding to all variables apart from one i.e.

$$E(v_i) = E(1,\ldots,v_{i-1},v_{i+1},\cdots,V) \qquad (6)$$

For example, if $E(v_2) > E(v_1)$ it suggests that variable v_2 is more important to data set cluster structure than variable v_1. Calculating all V entropies according to formula 6 we can arrange their sequence in nondecreasing order. The only thing that remains to be settled is to decide where to divide

Manuscript received March 29, 2010.

J. Korzeniewski is with the University of Lodz, Poland (fax: +4842-6355032; e-mail: jurkor@ wp.pl).

this sequence into two groups representing true and masking variables. Instead of the greatest jump criterion (used by some researchers, e.g *Steinley & Brusco,* 2008) we applied the *k*-means grouping of variables (for *k*=2) with starting points being two extreme entropies. The class of variables assigned to the lowest entropy will be discarded as masking variables, while the class corresponding to the highest entropy will represent true variables.

The sequence of ranked variables determines to some extent the performance of the second algorithm proposed by the Authors which consists in running a grouping method based only on m variables counting from the beginning of the sequence. For each grouping we compute the value of a criterion based on $tr\left(W^{-1}B\right)$, where $B=\sum_{k=1}^{K}\left(\bar{x}_k-\bar{x}\right)\left(\bar{x}_k-\bar{x}\right)^T$ is the between cluster variance matrix and $W=\sum_{k=1}^{K}\sum_{x_i\in C_k}\left(x_i-\bar{x}_k\right)\left(x_i-\bar{x}_k\right)^T$ is the within cluster variance matrix. The trace has the following interpretation: the higher its value the more distinct are the data set clusters. The number of variables with the highest value of the criterion is considered as the set of true variables. The grouping method originally used was the *k*-means grouping.

III. SIMULATION EXPERIMENT

In order to assess comparatively the entropy based algorithms with other existing methods we applied similar experiment pattern to the one used by Steinley and Brusco (2008) with respect to the number of variables (true and masking), overlap size, type of distributions, number of clusters etc.. The pattern was even broader with respect to the number of clusters considered as we included sets with 3 clusters.

We generated several thousands data sets, each consisting of 200 data items which constituted a couple of clusters (each cluster generated from a normal distribution) differing with respect to the following factors.

The first factor, the number of clusters in the data set was examined at four levels – 3, 4, 6 and 8 clusters.

The second factor, number of items in clusters was examined at three levels: (a) an equal number of objects in each cluster; (b) 10% of objects and (c) 60% of objects in one cluster and the remaining objects equally divided across the remaining clusters.

The third factor, the number of true variables was tested at three levels – 2, 4 and 6.

The fourth factor, the probability of overlap between clusters on each true variable was tested at five levels – 0, 0.1, 0.2, 0.3, 0.4. The overlap was of the “chain” type (see Steinley and Henson, 2005) and so, on each dimension, there were *k*-1 pairs of overlapping clusters (*k* – number of clusters).

The fifth factor, the degree of within-cluster correlation had two variants: (a) the covariance matrix for each cluster was the identity matrix ; (b) each cluster had the same covariance matrix with ones on the diagonal and the off-diagonal elements drawn from a continuous distribution on the interval [0.3; 0.8].

The sixth factor, the number of masking variables, was tested at three levels – 2, 4 and 6.

The seventh factor, the distribution of the masking variables was tested at five levels: (a) all masking variables were independently generated from a skewed distribution (the gamma with one degree for the numerator and denominator); (b) all masking variables were independently generated from the normal distribution with zero mean and identity covariance matrix; (c) all masking variables were independently generated from the normal distribution with zero mean and covariance matrix with ones on the diagonal and 0.25 off the diagonal; (d) all masking variables were independently generated from the normal distribution with zero mean and covariance matrix with ones on the diagonal and 0.5 off the diagonal; (e) all masking variables were independently generated from the normal distribution with zero mean and covariance matrix with ones on the diagonal and 0.75 off the diagonal. In addition, every pattern was repeated 2 times which gave 10800 data sets.

To assess the method we used two criteria (see: *Steinley and Brusco* 2008, p. 135):

Recall: The number of relevant variables in the chosen subset of variables divided by the total number of relevant variables.

Precision: The number of relevant variables in the chosen subset of variables divided by the total number of variables selected.

Recall and precision were computed for every data set and, subsequently, the arithmetic mean of the two measures was computed from all data sets.

IV. RESULTS AND CONCLUSIONS

The first algorithm performed quite well in all cases apart from cases c), d), and e) of the masking variables distribution (see Table 1). In cases a) and b) the precision and recall ranged from 0.75 to 0.92 which is a good performance. However, in all cases in which masking variables were correlated the method performed very badly. In case c) with very small amount of correlation i.e. 0.25, the precision and recall were around 0.65, in cases d) around 0.54 an in case e) hardly 0.45. It is also very important to point out that in these 3 cases the quality of ranking of the entropies was in total mass. If by the proper succession of ranking we denote every possible ranking in which all true variables come first (in arbitrary succession) and all masking variables second (in arbitrary succession), it turns out that the proper ranking came up about 5% of all rankings. Therefore, the limited version of a wrapper approach based on the ranked list of variables i.e. the second algorithm would have to peform incorrectly.

These results contradict the Dash and Liu statement that this method seems to be doing well in presence of correlation between masking variables.

Table 1 Precision and racall for different types of the masking distribution.

Type of masking distribution	a	b	c	d	e
Precision	0.92	0.83	0.63	0.52	0.45
Recall	0.86	0.75	0.66	0.55	0.44

Source: own calculations.

REFERENCES

[1] M. Brusco, J. D. Cradit, A variable-selection heuristics for K-means clustering, *Psychometrika* 66, 2001.

[2] F. J. Carmone Jr., Kara Ali , S. Maxwell, HINoV: A New Model to Improve Market Segment Definition by Identifying Noisy Variables , *Journal of Marketing Research,* Vol. 36, No. 4, 1999.

[3] M. Dash, H. Liu, Feature selection for clustering , *Proceedings of Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining,* (PAKDD), 2000.

[4] M. Law, A. Jain , M. Figueiredo, Simultaneous feature selection and clustering using mixture models , *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 2004.

[5] D. Steinley, M. Brusco, Selection of Variables in Cluster Analysis: An Empirical Comparison of Eight Procedures , *Psychometrika* 73 No. 1, 2008.

[6] D. Steinley, R. Henson, OCLUS: An analytic method forgenerating clusters with known overlap. *Journal of Classification*, 22, 2005.

[7] D. Steinley, M. Brusco, A new variable weighting and selection procedure for K-means cluster analysis, *Multivariate Behavioural Research* 43, Taylor and Francis, 2008.