Publication 8

Importance of Individual Variables in the k−Means Algorithm

Juha Vesanto

In *Proceedings of the Pacific−Asia Conference Advances in Knowledge Discovery and Data Mining (PAKDD2001)*,
Springer−Verlag, pp. 513−518, 2001.

Importance of Individual Variables in the k-Means Algorithm

Juha Vesanto

Neural Networks Research Centre,
Helsinki University of Technology,
P.O.Box 5400, 02015 HUT, Finland
Juha.Vesanto@hut.fi

Abstract. In this paper, quantization errors of individual variables in k-means quantization algorithm are investigated with respect to scaling factors, variable dependency, and distribution characteristics. It is observed that Z-norm standardation limits average quantization errors per variable to unit range. Two measures, quantization quality and effective number of quantization points are proposed for evaluating the goodness of quantization of individual variables. Both measures are invariant with respect to scaling/variances of variables. By comparing these measures between variables, a sense of the relative importance of variables is gained.

Keywords: k-means, quantization, scaling, normalization, standardation

1 Introduction

Unsupervised clustering algorithms are important tools in exploratory data analysis. Because clustering criteria is usually based on some distance measure between individual data vectors, they are highly sensitive to the scale, or dispersion, of the variables. It is easy to come up with examples where the clustering result can be considerably changed by a simple linear rescaling of the variables (see e.g. [5] p.5). Therefore, apart from the case when the original values of the variables are somehow meaningful with respect to each other, some kind of rescaling or standardation procedures are normally recommended prior to the clustering [8,5,1].

The most common standardation procedure is to treat all variables independently and transform each to so-called Z-scores by substracting the mean and dividing by the standard deviation of each variable. Another widely used method is to normalize the range of the variable to unit interval. Also other nonlinear, multidimensional, and even local standardation operations are possible [6,5,8,4]. However, what these more complex may gain in flexibility, they loose in interpretative power.

In this paper, vector quantization of numerical data sets is investigated with respect to the quality of quantization of individual components. The aim is to derive easily understandable measures of quantization quality to be used as part of a data understanding framework.

D. Cheung, G.J. Williams, and Q. Li (Eds.): PAKDD 2001, LNAI 2035, pp. 513–518, 2001.

2 Definitions

Let D be a $n \times d$ sized numerical matrix such that each row of the matrix corresponds to one data sample $\mathbf{x}_i$, and each column to one variable $\mathbf{v}_j$. Without loss of generality, let each $\mathbf{v}_j$ be Z-normed to zero mean and unit variance. Scaling D with a set of scaling factors $\{f_j\}$ is a simple multiplication operation: $\hat{\mathbf{v}}_j = f_j \mathbf{v}_j$, which results in a new data set $\hat{D}$. The variances of the new variables are equal to squared scaling factors $\hat{\sigma}_j^2 = f_j^2 \sigma_j^2 = f_j^2$.

The scaled data set $\hat{D}$ is quantized (or clustered). In this paper, the batch k-means algorithm is used for quantization [3,7]. The algorithm finds a set of k prototype vectors $\hat{\mathbf{m}}_l$ which minimize the representation error:

$$E = \frac{1}{n}\sum_{i=1}^{n} ||\hat{\mathbf{x}}_i - \hat{\mathbf{m}}_{b_i}||^2 = \sum_{j=1}^{d} \frac{1}{n}\sum_{i=1}^{n} |\hat{x}_{ij} - \hat{m}_{b_i j}|^2 = \sum_{j=1}^{d} E_{\hat{\mathbf{v}}_j} = \sum_{j=1}^{d} \hat{E}_j, \qquad (1)$$

where $b_i = \arg_l \min(||\hat{\mathbf{x}}_i - \hat{\mathbf{m}}_l||)$, $||\cdot||$ is euclidean distance metric, and $\hat{E}_j$ is the average quantization error of scaled variable $\hat{\mathbf{v}}_j$.

The variable-wise quantization errors $\hat{E}_j$ can be represented as functions of the number of effective number of quantization points k_j, ie. the number of quantization points which are needed to get the same error when only $\mathbf{v}_j$ is quantized. The k-means algorithm finds a local minima of E in the space defined by k_j. The derivative of E gives insight to what happens when the importance of a variable increases:

$$\frac{\delta E}{\delta k_{j'}} = \sum_{j=1}^{d} \hat{\sigma}_j^2 \frac{\delta E_j}{\delta k_{j'}} = \sum_{j=1}^{d} \hat{\sigma}_j^2 \frac{\delta E_j}{\delta k_j}\frac{\delta k_j}{\delta k_{j'}}. \qquad (2)$$

This shows that the allocation of quantization points is dependent on three factors: the variance of each variable $\hat{\sigma}_j^2$, distribution characteristics of the variable $\frac{\delta E_j}{\delta k_j}$ and dependencies between variables: $\frac{\delta k_j}{\delta k_{j'}}$. In general, since the total supply of quantization points k is limited, the partial derivatives $\frac{\delta k_j}{\delta k_{j'}}$, $j \neq j'$ are negative. On the other hand, for those variables which are (highly) dependent on variable j', $\frac{\delta k_j}{\delta k_{j'}}$ is positive, and thus increased $k_{j'}$ actually benefits them.

The function $\hat{E}_j(k)$ can be estimated directly from data by making a number of quantizations of each variable with varying values of k. This is relatively light operation compared to quantization of the whole data space. For uniform distribution, the function is also easy to derive analytically: $\hat{E}_j(k) = \hat{\sigma}_j^2 k^{-2}$, in which case $k_j = (\hat{\sigma}_j^2/\hat{E}_j)^{0.5}$. A similar formula can also be used for other continuous distributions, for example for gaussian distribution $\hat{E}_j(k) \approx \hat{\sigma}_j^2 k^{-1.7}$.

For each variable-wise quantization error $\hat{E}_j$, the minima is reached when $k_j = k$, and maxima when $k_j = 1$. In the latter case the data is quantized using just its mean, in which case $\hat{E}_j = \hat{\sigma}_j^2$. Thus, the quantization errors $\hat{E}_j$ are limited by:

$$\hat{E}_j(k) \leq \hat{E}_j \leq \hat{\sigma}_j^2. \qquad (3)$$

Since $\hat{E}_j = f_j^2 E_j$ and $f_j = \hat{\sigma}_j$ the limits can also be defined with respect to the original unscaled variables:

$$E_j(k) \le E_j \le 1. \tag{4}$$

The quantization quality of a variable can be estimated as how close E_j is to the minimum possible error $E_j(k)$:

$$q_j = \frac{E_j - E_j(k)}{1 - E_j(k)} = \frac{\hat{E}_j - \hat{E}_j(k)}{\hat{\sigma}_j^2 - \hat{E}_j(k)} \tag{5}$$

With sufficiently large k, $\hat{E}_j(k) = 0$ and thus $q_j \approx \hat{E}_j / \hat{\sigma}_j^2$. Since quantization quality is intricately linked with variable importance — the more important variable, the better it is quantized — q_j acts as a measure of variable importance.

3 Experiments

To get further insight to how the proposed methods work in practice, some tests were made using artificial data sets. The number of data points was 1000, and they were quantized using 100 quantization points. To ensure good quantization, ten k-means runs with 50 epochs were made and the best one was utilized. The data points were from four different kinds of distributions: gaussian, uniform, exponential and "2-spikes", which was formed as a mixture of two supergaussian distributions with equal prior probabilities.

Figure 1 studies the effect of scaling factors in a 10-dimensional case. Instead of a steady increase in quantization error $\hat{E}_j$ of the scaled variable, which might be expected, there is a transfer area for scaling factors in range $[1, 10]$, where the error of the scaled variable is about equal to the quantization error of all the other variables. The behaviour is due to the limits imposed on $\hat{E}_j$ by the possible values of k_j. With small scaling factors, $k_1 \approx 1$, while for large scaling factors $k_1 \approx 100$.

The 2-spikes distribution (Figure 1 top right corner) has a sudden decrease in quantization error for $1 < f_1 < 5$. This decrease shows the effect of increasing k_1 over the threshold of 2: at this point, the quantization points are divided to two groups, one for either spike of the variable.

Both importance measures q_j and k_j work very well in all cases showing how the importance of the first variable increases with increasing scaling factor in the significant range of scaling factors, and levels off when the scaling factor does not really matter.

Figure 2 studies the effect of variable dependencies. In the test, 10-dimensional data sets with 1 to 9 identical variables were quantized. The quantization errors behave exactly as if there were actually 10 to 2 variables, respectively: the sum of errors of the dependent variables is equal to the errors of each of the independent variables.

As an example of the usage of the proposed indicators, the IRIS data set [2] was quantized using 15 prototypes, see Table 1. Both measures q_j and k_j indicate

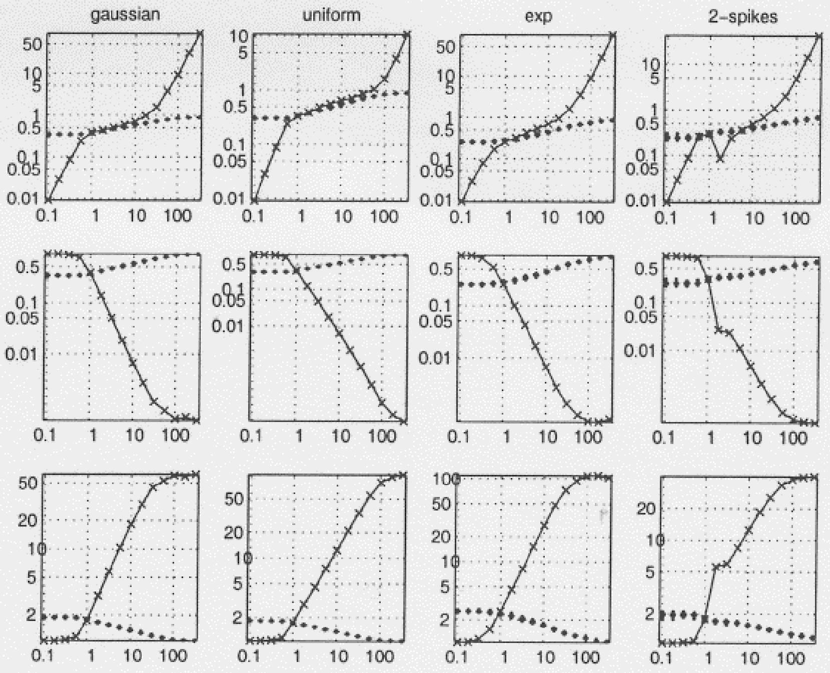

Fig. 1. Quality of quantization. $\hat{E}_j$ on top row, q_j in the middle, and k_j on the bottom. The data consists of 10 similar variables (gaussians on the left, uniform and exponential on the middle, and 2-spikes distributions on the right) the first of which is scaled by a factor of $f_1 \in [0.1, 316]$. The solid line corresponds to the first variable, and the separate markers (actually: boxplots) to the other 9 variables. Note that all axes are logarithmic.

that petal length variable was the most important variable in this quantization, although its descaled quantization error (Δ column) is the biggest of the four variables.

Table 2 shows results for a modified version of IRIS data set. Four new variables have been added: one discreet variable which indicates the class information of the sample (values 1, 2 and 3 for the three different Iris subspecies) and three random uniformly distributed variables. The random variables are clearly the least important.

Table 1. Quantization of the IRIS data set with 15 prototypes. Both min, max and Δ values are in original value range in order to faciliate interpretation by domain experts.

Variable	[min,max]	Δ	q_j	k_j
sepal length	[4.3,7.9]	±0.2	0.063	5.07
sepal width	[2.0,4.4]	±0.1	0.096	4.29
petal length	[1.0,6.9]	±0.3	0.018	5.62
petal width	[0.1,2.5]	±0.1	0.033	3.87

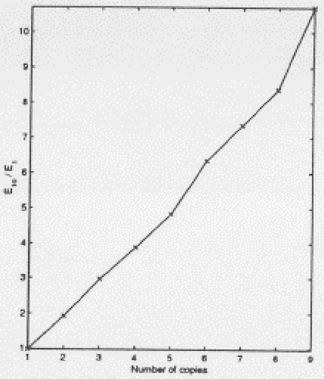

Fig. 2. Ratio between quantization errors $\hat{E}_j$ of the first and 10th variables (out of 10) when the first 1 to 9 variables were identical: $\hat{E}_{10} = c\hat{E}_1$ where c is the number of copies.

Table 2. Quantization of the augmented IRIS data set with 20 prototypes.

Variable	[min,max]	Δ	q_j	k_j
sepal length	[4.3,7.9]	±0.4	0.2	2.69
sepal width	[2.0,4.4]	±0.3	0.33	2.29
petal length	[1.0,6.9]	±0.3	0.036	3.60
petal width	[0.1,2.5]	±0.2	0.065	2.94
iris species	[1.0,3.0]	±0.2	0.062	2.75
random 1	[0.0,1.0]	±0.1	0.23	2.06
random 2	[0.0,1.0]	±0.2	0.33	1.86
random 3	[0.0,1.0]	±0.2	0.3	1.89

4 Discussion

Various studies have investigated and compared different kinds of standardation/scaling methods in clustering problems. For example in [6] several standardation procedures were compared to each other in an artificial clustering problem. Standardation based on range was often found to be superior to the Z-norm standardation. This is understandable since clustered distributions, for example dicreet variables, retain more of their variance than continuous variables in scaling by range. Thus they have, in the view of the results in this paper, bigger inherent scaling factors. However, Z-norm provides a more uniform starting point in quantization, since the maximum quantization errors are equal for all variables.

The importance of a variable can be viewed as the gain — decrease in the quantization error — the quantization algorithm achieves through increasing the effective number of quantization points k_j of some variables, and (therefore) decreasing k_j of the others. The allocation of k_j seems to depend primarily on three factors: scaling of the variables, their distribution characteristics, and their dependency on the other variables (see Eq. 2). Of these scaling has quite straightforward effect, and the effect of distribution characteristics can be assessed by calculating the 1-dimensional quantization errors to a range of k-values. The third factor is the most problematic, and also the most interesting, because it appears to allow a way to investigate variable dependencies through vector quantization.

Variable importance and quantization quality are important pieces of information when analysing and interpreting a quantization or a clustering result. The final quantization error of a variable — even when compared to errors of the other variables — does not by itself give very clear picture of the quantization quality of the variable. In this paper, two measures q_j and k_j have been proposed which are well suited for evaluating the quantization quality of single variables.

Acknowledgments. This work has been carried out in the "KÄD" project in Helsinki University of Technology. The financing of TEKES and Jaakko Pöyry Consulting is gratefully acknowledged.

References

1. Michael R. Anderberg. *Cluster Analysis For Applications.* Academic Press, 1973.
2. E. Anderson. The irises of the gaspe peninsula. *Bulletin of American Iris Society*, 1935.
3. Robert M. Gray. Vector quantization. *IEEE ASSP Magazine*, pages 4–29, April 1984.
4. Jari A. Kangas, Teuvo K. Kohonen, and Jorma T. Laaksonen. Variants of Self-Organizing Maps. *IEEE Transactions on Neural Networks*, 1(1):93–99, March 1990.
5. Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: and Introduction to Cluster Analysis.* John Wiley & Sons, Inc., 1990.
6. Glenn W. Milligan and Martha C. Cooper. A study of standardation of variables in cluster analysis. *Journal of Classification*, 5:181–204, 1988.
7. John Moody and Christian J. Darken. Fast Learning in Networks of Locally-Tuned Processing Units. *Neural Computation*, 1(2):281–294, 1989.
8. Dorian Pyle. *Data Preparation for Data Mining.* Morgan Kaufmann Publishers, 1999.