PROCEEDINGS OF THE ROYAL SOCIETY A

CrossMark click for updates

Proc. R. Soc. A (2012) **468**, 2630–2651
doi:10.1098/rspa.2011.0704
Published online 18 April 2012

Statistical approach to normalization of feature vectors and clustering of mixed datasets

By Maria M. Suarez-Alvarez[1,*], Duc-Truong Pham[2,†], Mikhail Y. Prostov[3] and Yuriy I. Prostov[4]

[1]*School of Engineering, Cardiff University, Cardiff CF24 0AA, UK*
[2]*School of Mechanical Engineering, University of Birmingham, Birmingham B15 2TT, UK*
[3]*Faculty of Mechanics and Mathematics, Moscow State University, Moscow 119991, Russia*
[4]*Department of Higher Mathematics, Moscow Institute of Radio Engineering, Electronics and Automation, Technical University, 78 Vernadskogo pr., Moscow 117454, Russia*

Normalization of feature vectors of datasets is widely used in a number of fields of data mining, in particular in cluster analysis, where it is used to prevent features with large numerical values from dominating in distance-based objective functions. In this study, a unified statistical approach to normalization of all attributes of mixed databases, when different metrics are used for numerical and categorical data, is proposed. After the proposed normalization, the contributions of both numerical and categorical attributes to a specified objective function are statistically the same. Formulae for the statistically normalized Minkowski mixed p-metrics are given in an explicit way. It is shown that the classic z-score standardization and the min–max normalization are particular cases of the statistical normalization, when the objective function is, respectively, based on the Euclidean or the Tchebycheff (Chebyshev) metrics. Finally, clustering of several benchmark datasets is performed with non-normalized and introduced normalized mixed metrics using either the k-prototypes (for $p=2$) or another algorithm (for $p\neq 2$).

Keywords: clustering; normalization; standardization; Minkowski metrics; statistics

1. Introduction

Data mining (DM) is a process of extracting relations and patterns from data; therefore, DM transforms raw collections of data into information. It is well known (Jain & Dubes 1988; Cios *et al.* 2007) that data can have diverse formats and can be stored through a variety of different storage models. In a number of fields of machine intelligence, e.g. in texture clustering, image retrieval, speech recognition and clustering of databases, an object is often represented by a vector variable,

*Author for correspondence (suarez-alvarezm@cardiff.ac.uk).

†Also a visiting professor at King Saud University (KSU) in Riyadh, Saudi Arabia.

Received 1 December 2011
Accepted 20 March 2012

namely the feature vector **A**. The collection of objects described by the same features is called a dataset. The goal of clustering is to discover similarities and patterns within a large dataset by splitting data into clusters (groups). Because it is assumed that the data are unlabelled, clustering is often considered as the most important unsupervised learning problem (Cios *et al.* 2007).

Depending on the clustering goals, different formalization and different methods for clustering are involved (Mirkin 2005). Clustering techniques can be divided into three main categories: hierarchical clustering (HC); model-based clustering; and objective function-based clustering (Cios *et al.* 2007; Gan *et al.* 2007). HC is based on creating a hierarchical (multi-level) decomposition of the set of data points using some criteria or models. The clustering results are represented in a form of a graph (dendrogram). The standard HC methods can handle data with numeric and categorical values. However, it is generally accepted (Jain & Dubes 1988; Cios *et al.* 2007) that the quadratic computational cost makes HC unacceptable for clustering large datasets.

In model-based clustering methods, it is assumed that the data are generated by a mixture of underlying probability distributions in which each component represents a different cluster (Fraley & Raftery 2002; Cios *et al.* 2007). The expectation–maximization (EM) algorithm is a popular approach in model-based clustering when the observations can be viewed as incomplete data (Dempster *et al.* 1977). It admits both categorical and continuous attributes. The distance-based techniques, including k-prototypes; and model-based techniques, including the EM algorithm, have their advantages and drawbacks. For example, the k-prototypes algorithm is computationally easy, fast and memory-efficient; however, its procedure does not guarantee the convergence to a global extremum (Mirkin 2005) and it works well mainly when resulting clusters are close to hyperspherical in shape. The model-based clustering is more flexible and EM algorithms do not require the specification of distance measures; however, there are various reasons that may make the EM algorithm not practical for models with very large numbers of components and its rate of convergence can be very slow (Fraley & Raftery 2002). Because the distance-based techniques, including k-prototypes, are currently the major clustering techniques, this study will deal only with distance-based objective function clustering.

Normalization (standardization) of feature vectors is a long-standing problem of DM. Normalization of attributes has been discussed by many authors (Kaufman & Rousseeuw 2005; Larose 2005; Gan *et al.* 2007; Kamath 2009). A direct application of geometric measures (distances) to attributes with large ranges will implicitly assign bigger contributions to the metrics than the application to attributes with small ranges. In addition, the attributes should be dimensionless because the numerical values of the ranges of dimensional attributes depend on the units of measurements and, therefore, the choice of the units of measurements may greatly affect the results of clustering. Hence, one should not use distance measures such as the Euclidean distance without normalization of data (Aksoy & Haralick 2001; Larose 2005). In fact, two kinds of normalizations have been discussed in the overwhelming majority of papers, namely the min–max normalization and z-score standardization. Aksoy & Haralick (2001) gave a review of normalization techniques that have to be applied to numerical data before conducting a cluster analysis. In essence, the goal of all normalization procedures reviewed was the normalization of each feature component to the

[0, 1] range, i.e. the min–max normalization. Furthermore, one needs to apply normalization not only to numerical attributes but also to categorical attributes. As noted by Kamath (2009), normalization is not a trivial task, nevertheless it has to be discussed because 'in the real world, dirty data sets need cleaning; raw data need to be normalized; outliers need to be checked' (Larose 2005, p. xiii). Normalization relies on the use of various mathematical concepts, and, hence, it is important to develop appropriate mathematical tools for this procedure.

In general, unsupervised DM with no *a priori* information about the importance of attributes should follow the 'principle of equal importance of features' (Mirkin 2005, p. 218). Therefore, normalization should give all attributes equal influence on characterizing overall dissimilarity between pairs of objects (Hastie *et al.* 2001; Pham *et al.* 2006). However, Hastie *et al.* (2001), after introducing a correct interpretation of the normalization procedure, gave an example where standardization obscured the two well-separated groups. They argued that giving all attributes equal influence in this case would 'tend to obscure the groups to the point where a clustering algorithm cannot uncover them' (Hastie *et al.* 2001, p. 458). In certain cases, clustering without normalization has given good results (Jain & Dubes 1988; Hastie *et al.* 2001). However, the fact that the attributes had proper weights was purely by chance.

A statistical approach for normalization of mixed metrics will be presented in this study. Although statistical approaches to DM are very popular and it is well known that cluster analysis can be based on probability and statistical models (Bock 1996; Mirkin 1996; Fraley & Raftery 2002), these statistical approaches were not targeted to normalization of metrics. Almost no papers were devoted to discussion of normalization of categorical and mixed datasets. For example, the k-prototypes algorithm was applied to a non-normalized metric by Huang (1998). Although Larose (2005) underlined that when measuring distance the numerical attribute values should be normalized, he suggested to use for categorical attributes the matching dissimilarity measure without normalization or as he mentioned, 'perhaps, when mixing categorical and continuous variables, the min–max normalization may be preferred' (Larose 2005, p. 101). It seems to the authors that only Mirkin (1996, 1997, 1998, 2005) discussed standardization of mixed features based on their contributions to the quadratic data scatter. He said that 'with feature rescaling, feature scales become balanced according to the principle of equal importance of each feature brought into the data table' (Mirkin 2005, p. 64). In spite of Mirkin's work, normalization was neither mentioned in application to mixed data (Huang 1998; Bachner 2000) nor used in heuristic mixed metrics (Gibert & Cortés 1997). In 1998, Mirkin stated that methods for analysis of data in mixed feature space are still an issue.

It is natural to follow the principle of equal importance of features and assume that the average contribution of the jth feature component to the total similarity measure is equal to its mean, and, therefore, the goal of a normalization procedure is the equalization of the attribute contributions to the measure. If it is known *a priori* that some attributes are irrelevant to the problem under consideration, then they can be removed from the feature vector. A unified statistical treatment to both numerical and categorical features of mixed datasets is used in this study. New normalized metrics are introduced such that the average contributions of all attributes to the measures are equal to each other

from a statistical point of view. Although this idea has been recently discussed in the literature (Hastie *et al.* 2001; Pham *et al.* 2006), nothing was said about the statistical consistency of the proposed estimators. In addition, Hastie *et al.* (2001) used biased estimators. Finally, the previously proposed approaches were not applicable to some metrics. Here, mathematically rigorous treatment of the normalization procedure is presented and examples of normalized metrics (e.g. the Minkowski and Tchebycheff metrics) are given in an explicit way. Besides the proposed approach is extended to the case of mixed metrics, i.e. when different metrics are used for numerical and categorical data, respectively. Advantages of the introduced normalized metrics are demonstrated on examples when the k-prototypes algorithm is applied to both non-normalized and properly normalized mixed metrics. It will be shown that the accuracy usually increases when clustering is performed using normalized metrics. The questions concerning the mathematical proof of the consistency and unbiasedness of the proposed estimators are discussed elsewhere (Suarez-Alvarez 2010).

2. Preliminaries

(*a*) *Mathematical representation of datasets*

Datasets may be stored as flat files. Flat (rectangular) files are the most common way to store the datasets and further only flat files are considered. The rows represent objects (also known as records, individuals, patterns, data points) and the columns represent features. In application to databases, the features are called attributes, and hence $\mathbf{A}$ can be defined as a set of attributes, $\mathbf{A} = (A_1, A_2, \ldots A_{m+l})$. In many traditional applications, it is assumed usually that all the features are the same type. However, real-life datasets are often mixed; they consist of both numerical and categorical types, i.e. each attribute can take on a finite (categorical attribute) or infinite (continuous) number of possible values (numerical attribute). Only these two data types of attributes are considered here because other types of attributes can be mapped to these two types. In this study, it is accepted the term *categorical* as a general term that can be split into nominal and ordinal. If categorical variables have ordered scales they are called ordinal variables, whereas the variables having no ordered scales are called nominal variables. Binary variables are considered here as categorical.

From the mathematical point of view, the vector of features $\mathbf{A}$ for mixed data can be split into $\mathbf{A} = (\mathbf{A^n}, \mathbf{A^c})$, namely the vector of numerical features $\mathbf{A^n} = (A_1^n, \ldots, A_m^n)$ and the vector of categorical features $\mathbf{A^c} = (A_1^c, \ldots, A_l^c)$. For numerical or quantitative features, the feature domain $\mathrm{Dom}(A_j)$ can be represented on the real line, whereas for categorical features (sometimes these features are also called nominal or qualitative), the domain is a finite set of different states. Evidently, categorical features may be represented by numerical codes of possible different states of the feature.

A database can be represented as a matrix of size $N \times (m + l)$ where N is the number of records, and $(m + l)$ is the total number of attributes, i.e. the ith row of the matrix represents the ith record of the database $(1 \leq i \leq N)$. This row is a vector $(x_{i1}, \ldots, x_{im}, y_{i1}, \ldots, y_{il})$, whose values $x_{i1}, \ldots, x_{im}$ are numerical, whereas the values $y_{i1}, \ldots, y_{il}$ are categorical.

(*b*) *Clustering, objective functions and similarity measures*

A very general category of clustering is concerned with building partitions (clusters) of datasets on the basis of some performance index known also as an objective or cost function. This is a function associated with an optimization problem where the best element from some set of available alternatives is chosen to minimize or maximize the function (Zhigljavsky & Žilinskas 2008). The value of this function determines how good the chosen solution is. There are various clustering algorithms for objective function-based clustering (MacQueen 1967; Huang 1997; Pham *et al.* 2011) because it is practically unfeasible to find a global optimum for the objective function by considering all possible combinations of elements (exhaustive search).

Here, one needs to distinguish hard and fuzzy cluster methods. The clustering is hard if the obtained groups satisfy the following requirements of a partition (i) each group must contain at least one object and (ii) each object must belong to exactly one group (Kaufman & Rousseeuw 2005). In this case, data are divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels (Bezdek 1980).

Any measure of the degree of closeness is called similarity measure. In clustering analysis of numerical datasets, it is very common to calculate the similarity or dissimilarity between two feature vectors $\mathbf{x}_1 = (x_{11}, \ldots, x_{1m})$ and $\mathbf{x}_2 = (x_{21}, \ldots, x_{2m})$ using a square distance measure. Indeed, it is very natural to use the Euclidean metric (distance) ρ_{E} (or L_2 metric)

$$\rho_{\mathrm{E}}(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_2 = \left(\sum_{j=1}^{m} (x_{1j} - x_{2j})^2 \right)^{1/2} \tag{2.1}$$

as a measure for continuous numerical features because this metric is in everyday use. Often, other similarity measures are also used. For example, the city block distance (or L_1 metric)

$$\rho_{\mathrm{B}}(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{x}_1 - \mathbf{x}_2||_1 = \sum_{j=1}^{m} |x_{1j} - x_{2j}|, \tag{2.2}$$

the Minkowski distance ρ_p (or L_p metric)

$$\rho_p(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{x}_1 - \mathbf{x}_2||_p = \left(\sum_{j=1}^{m} |x_{1j} - x_{2j}|^p \right)^{1/p}, \tag{2.3}$$

where p is a positive number, $1 \leq p < +\infty$; and the Tchebycheff (Chebyshev) or maximum norm metric

$$\rho_{\max}(\mathbf{x}_1, \mathbf{x}_2) = \|\mathbf{x}_1 - \mathbf{x}_2\|_{\max} = \max_j |x_{1j} - x_{2j}|. \tag{2.4}$$

The Euclidean metric and city block distance are particular cases of the Minkowski metric for $p = 2$ and $p = 1$, respectively. The Tchebycheff metric can be obtained from the Minkowski metric as the following limit $p \to \infty$. Other

metrics are also applied to numerical datasets. However, there are no such natural similarity measures for categorical data and for mixed (numeric and categorical) data. Therefore, two different similarity measures are often combined for clustering of mixed data (Gibert & Cortés 1997).

(i) *Objective functions*

If the objective (cost) function J is based on the Euclidean distance, then the k-means algorithm (MacQueen 1967) can be used in application to numerical data, while an extension of this algorithm, namely the k-prototypes algorithm (Huang 1997) can be used in application to mixed data. These clustering algorithms minimize the cost function (Bezdek 1980; Huang 1997) for 'hard' (non-fuzzy) k-partition of a dataset into k clusters

$$\left.\begin{aligned} & J=\sum_{j=1}^{k}\sum_{i=1}^{N} u_{ij}\rho_{\mathrm{E}}^{2}(A_i,Q_j), \quad u_{ij}\in\{0,1\}, \quad 1\le i\le N, \quad 1\le j\le k \\ \text{and} \quad & \sum_{j=1}^{k} u_{ij}=1, \quad \sum_{i=1}^{N} u_{ij}>0. \end{aligned}\right\} \tag{2.5}$$

Here, ρ_{E} is a similarity metric and u_{ij} is an element of the partition matrix. The condition $u_{ij}=1$ means that the record A_i is assigned to cluster j with prototype Q_j.

If the cost function J is based on the general case of the Minkowski distance, then the k-means algorithm is not applicable and instead of the above standard cost function $\sum\rho_{\mathrm{E}}^{2}$, the cost function $\sum\rho_{p}^{p}$ will be used

$$\left.\begin{aligned} & J=\sum_{j=1}^{k}\sum_{i=1}^{N} u_{ij}\rho_{p}^{p}(A_i,Q_j), \quad u_{ij}\in\{0,1\}, \quad 1\le i\le N, \quad 1\le j\le k \\ \text{and} \quad & \sum_{j=1}^{k} u_{ij}=1, \quad \sum_{i=1}^{N} u_{ij}>0, \quad p\ge 1. \end{aligned}\right\} \tag{2.6}$$

Here, ρ_p is the Minkowski mixed p-metric.

(ii) *Statistical treatment of feature vectors*

With geometric similarity measures, usually no assumption is made about the probability distribution of the attributes and similarity (dissimilarity) is based on the distances between feature vectors in the feature space (Aksoy & Haralick 2001). However, each record (the row) of a database may be regarded as a random sample of a population under consideration (Mirkin 2005), i.e. one has a database of N observations (samples) and each sample (record) is a realization of possible

values of the feature vector **A**. Hence, it was suggested (Chen 1973) to use a weighted Euclidean distance that may take into account the dispersion of samples within a cluster.

For statistical treatment of feature vectors, one needs to know the probability distributions of their attributes. For a numerical attribute A_j^n, the probability distribution identifies the probability of the attribute value falling within a particular interval within the range of possible values. For a categorical attribute A_j^c, the probability distribution identifies the probability of certain states occurring.

It is assumed usually in the literature that each numerical feature has a normal (Gaussian) distribution with mean μ_j and variance σ_j. Their values may be estimated using standard statistical methods. However, in the general case, distribution functions are not known in advance and another function may be a better model for the attributes than the Gaussian distribution.

(*c*) *Classic normalization of numerical attributes*

The normalization procedure can be implemented in different ways. The most cited of these approaches normalizes the data by dividing the attribute value x_{ij} by its range using scaling with a shift (the min–max normalization)

$$x_{ij}^* = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}}. \tag{2.7}$$

Here, x_{ij}^* is the normalized attribute value in the database, and $x_{\max,j}$ and $x_{\min,j}$ are the maximum and the minimum values of attribute A_j, respectively. The results scaled by (2.7) do not depend on the original units of data measurements, and this linear scaling will transform the data to the range $[0, 1]$. However, in the general case, this normalization procedure does not achieve equalization of the attribute means. Hence, the application of the transformation (2.7) for normalization of real-world datasets and consequent clustering using the Minkowski norm (Doherty *et al.* 2004) does not give equal contributions of variables to the similarity measures because the means of the different normalized attributes are not necessarily equal to each other.

For numerical databases with attributes having normal (Gaussian) distributions, the most common normalization procedure is to transform the attribute A_j^n to a random variable with zero mean and unit variance (the z-score standardization) by

$$x_{ij}^* = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{2.8}$$

where μ_j and σ_j are the mean and standard deviation for values of the jth attribute A_j^n, respectively. As it will be shown, this scaling provides equal contributions of variables to the Euclidean similarity measure.

The standard preliminary transformation of the feature columns of arbitrary type of attributes a_{ij} can be written as

$$a_{ij}^* = \frac{a_{ij} - m_j}{b_j}, \tag{2.9}$$

where m_j is the grand mean, i.e. the average over the entire entity set, and b_j 'it can be either the standard deviation or range or other quantity reflecting the variables spread' (Mirkin 1997, p. 475, 2005, p. 65). Evidently, (2.8) is a particular case of (2.9). As Mirkin (2005) noted, the standardization (2.9) does not guarantee equal contributions of all features to the results.

In addition to the above methods, Milligan & Cooper (1988) (see also Gan *et al.* 2007) suggested several other ways for standardization of attributes, e.g.

$$x_{ij}^* = \frac{x_{ij}}{\sigma_j}, \tag{2.10}$$

and

$$x_{ij}^* = \frac{x_{ij}}{x_{\max,j}}. \tag{2.11}$$

However, it will follow from the statistical analysis below that (2.11) and (2.10) do not give equal contributions to appropriate similarity measures.

Aksoy & Haralick (2001) reviewed five normalization methods for numerical data, namely linear scaling to unit range, linear scaling to unit variance, transformation to a uniform $[0,1]$ random variable, rank normalization and normalization by fitting distributions. All these approaches intended to normalize each feature component to the $[0,1]$ range. It was noted that providing all attributes are normally distributed, the probability of the attribute value normalized by (2.8) is in the $[-1,1]$ range is equal to 68 per cent. If one applies an additional shift and rescaling as

$$x_{ij}^* = \frac{0.5(x_{ij} - \mu_j)}{(3\sigma_j) + 1}, \tag{2.12}$$

then this guarantees 99 per cent of the values to be in the $[0,1]$ range Aksoy & Haralick (2001). However, any shifting of the whole attribute column does not affect the distance metrics (2.1)–(2.2) and, therefore, it has no practical importance for clustering of datasets.

The idea to use a weighted Euclidean distance that may take into account the dispersion of samples within a cluster (e.g. Chen 1973) was generalized for the Minkowski distance by Pham *et al.* (2006), who were not aware that a very similar idea was formulated earlier for a more general case by Hastie *et al.* (2001). However, one has to realize that their estimators were biased. Further, the question concerning the consistency of the proposed estimators was not discussed. Hastie *et al.* (2001) considered as example an only the same case as in Chen (1973), namely the weighted Euclidean distance. On the other hand, there are metrics for which the above approach is not valid. Indeed, if one considers the Tchebycheff metric (2.4) for numerical attributes then the approach suggested by Hastie *et al.* (2001) is not directly applicable. How can one normalize this metric? Finally, if one studies a mixed metric that is a sum of two different metrics (one metric is used for numerical data, whereas another metric is used for categorical data), then their approach is also not applicable.

3. Normalization of the feature vectors

This study presents normalized attributes whose average contributions to the measures, regardless of type of attributes, are equal to each other from a statistical point of view.

(*a*) *Normalization for numerical datasets*

(i) *Normalization of the Minkowski metric*

To obtain a new normalized metric in the general case of the Minkowski metric (2.3), one should calculate the mean contribution of each jth attribute to the metric $E|X_{1j} - X_{2j}|^p$ (here E means the expectation of a variable) and divide the attribute in all records by this mean (if the mean is equal to zero then this attribute should be removed from the feature vector). Hence, the normalized Minkowski metric can be introduced in the following way

$$\rho_p^*(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_{j=1}^{m} \alpha_j |x_{1j} - x_{2j}|^p \right)^{1/p}, \tag{3.1}$$

where $\alpha_j = 1/E|X_{1j} - X_{2j}|^p$, X_{1j} and X_{2j} are independent random variables whose values are distributed in accordance with the distribution of the jth attribute.

In the general case, the distribution of the jth attribute is not known in advance, therefore, to estimate the expectation $|X_{1j} - X_{2j}|^p$ one can use the sample mean

$$\hat{E}|X_{1j} - X_{2j}|^p = \frac{1}{N^2} \sum_{r,s=1}^{N} |x_{rj} - x_{sj}|^p. \tag{3.2}$$

The estimation (3.2) is a biased estimator of $E|X_{1j} - X_{2j}|^p$, hence for small datasets it is better to use the following estimation

$$\hat{E}|X_{1j} - X_{2j}|^p = \frac{2}{N(N-1)} \sum_{1 \le r < s \le N} |x_{rj} - x_{sj}|^p, \tag{3.3}$$

that is an unbiased estimator. It can be proved (Suarez-Alvarez 2010) that (3.2) and (3.3) are consistent estimators.

(ii) *Normalization of the Euclidean metric*

For $p = 2$, the above-mentioned Minkowski metric (2.3) is the Euclidean metric. Normalization in this case is well known (Pham *et al.* 2006). However, for the sake of completeness, this case is considered here. Because X_{1j} and X_{2j} are independent random variables having the same distribution, for $p = 2$ one obtains $E|X_{1j} - X_{2j}|^2 = EX_{1j}^2 - 2EX_{1j}EX_{2j} + EX_{2j}^2 = 2(EX_{1j}^2 - (EX_{1j})^2) = 2\sigma_j^2$, where σ_j is

the standard deviation of the jth attribute. Thus, the normalized Euclidean metric has the following form

$$\rho^*_{\mathrm{E}}(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_{j=1}^{m} \frac{(x_{1j} - x_{2j})^2}{2\sigma_j^2} \right)^{1/2}. \tag{3.4}$$

For estimating σ_j^2 in (3.4), it is possible to use the following unbiased estimator of the sample variance

$$s_j^2 = \frac{1}{N-1} \sum_{r=1}^{N} (x_{rj} - \bar{x}_j)^2,$$

where $\bar{x}_{ij} = (1/N) \sum_{r=1}^{N} x_{rj}$ is the sample mean for the jth attribute.

From (3.4) one obtains the known form of normalization of features

$$x^*_{i1} = \frac{x_{i1} - \mu_1}{\sigma_1}, \ldots, x^*_{im} = \frac{x_{im} - \mu_m}{\sigma_m},$$

where μ_j is the mean of the jth attribute.

(iii) *Minkowski metrics with normally distributed numerical features*

Let us consider the case when one knows in advance that the values of the jth attribute are normally distributed. In this case, there is a quite attractive property of ρ_p. Let $f_{A_j^n}(x)$ be a normal distribution with mean μ_j and variance σ_j^2

$$f_{A_j^n}(x) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x - \mu_j)^2}{2\sigma_j^2} \right).$$

One can calculate now the expectation of the contribution of the jth attribute to the Minkowski metric

$$E|X_{1j} - X_{2j}|^p = \left(\frac{1}{\sqrt{2\pi}\sigma_j} \right)^2 \times I,$$

where

$$I = \iint_{-\infty}^{+\infty} |x_1 - x_2|^p \psi(x_1)\psi(x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2, \quad \psi(t) = \exp\left(-\frac{(t - \mu_j)^2}{2\sigma_j^2} \right).$$

If variables t_1 and t_2 are introduced

$$t_1 = \frac{x_1 - \mu_j}{\sigma_j} \quad \text{and} \quad t_2 = \frac{x_2 - \mu_j}{\sigma_j}$$

then, one has

$$E|X_{1j} - X_{2j}|^p = \sigma_j^p \times c_p,$$

where

$$c_p = \frac{1}{2\pi} \iint_{-\infty}^{+\infty} |t_1 - t_2|^p \exp\left(-\frac{t_1^2}{2} \right) \exp\left(-\frac{t_2^2}{2} \right) \mathrm{d}t_1\, \mathrm{d}t_2 = \frac{2^p}{\sqrt{\pi}} \Gamma\left(\frac{p+1}{2} \right),$$

and Γ is the Euler gamma function.

Thus, it follows from (3.1) that a new metric ρ^* for a dataset with Minkowski metric and numerical features having a normal distribution with mean μ_j and variance σ_j is

$$\rho_p^*(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_{j=1}^{m} \alpha_j |x_{1j} - x_{2j}|^p \right)^{1/p}, \tag{3.5}$$

where $\alpha_j = 1/(\sigma_j^p c_p)$. If $p = 2$, then (3.5) is reduced to (3.4).

(iv) *Normalization of the Tchebycheff metric*

The above approach can also be applied to normalize the Tchebycheff metric (2.4) for numerical attributes. Let X be a random variable and $p \geq 1$ then put $\|X\|_p = (E|X|^p)^{1/p}$ and denote $\|X\|_\infty$ the essential supremum of X, i.e.

$$\|X\|_\infty \equiv \operatorname{ess\,sup} |X| \equiv \inf\{0 \leq C < \infty : P(|X| > C) = 0\}.$$

It will be assumed further that the essential supremum of X is finite for all X under consideration.

In the case under consideration, $\mathbf{X}_{1j}$ and $\mathbf{X}_{2j}$ have the same distribution, and if the distribution function is unknown, then it is estimated using sampling distribution. In the latter case,

$$\|\mathbf{X}_{1j} - \mathbf{X}_{2j}\|_\infty = \max_i x_{ij} - \min_i x_{ij}. \tag{3.6}$$

For the normal distribution, the value of $\|X\|_\infty$ equals to ∞ and the above-described normalization procedure cannot be used. It can be proved (Suarez-Alvarez 2010) that the normalized Tchebycheff metric is given by

$$\rho_{\max}^*(\mathbf{x}_1, \mathbf{x}_2) = \max(\alpha_1 |x_{11} - x_{21}|, \ldots, \alpha_m |x_{1m} - x_{2m}|), \tag{3.7}$$

where $\alpha_j = 1/\|X_{1j} - X_{2j}\|_\infty$.

(*b*) *Datasets with mixed attributes*

For datasets with categorical attributes, it is possible to introduce different metrics (Ralambondrainy 1995; Gibert & Cortés 1997; Huang 1998). One of the most cited variants of metrics is studied here (Huang 1998; Gan *et al.* 2007), namely the distance between two categorical feature vectors $\mathbf{y}_1 = (y_{11}, \ldots, y_{1l})$ and $\mathbf{y}_2 = (y_{21}, \ldots, y_{2l})$ is defined as

$$\rho_{\text{cat}}(\mathbf{y}_1, \mathbf{y}_2) = \omega(y_{11}, y_{21}) + \cdots + \omega(y_{1l}, y_{2l}), \tag{3.8}$$

where

$$\omega(y_{1j}, y_{2j}) = \begin{cases} 0, & \text{for } y_{1j} = y_{2j} \\ 1, & \text{for } y_{1j} \neq y_{2j}. \end{cases}$$

Evidently, the square of the metric (3.8) is

$$\rho_{\text{cat}}^2(\mathbf{y}_1, \mathbf{y}_2) = \omega^2(y_{11}, y_{21}) + \cdots + \omega^2(y_{1l}, y_{2l}). \tag{3.9}$$

(i) *The Euclidean mixed metric*

Combining ρ_{E} and ρ_{cat} for mixed data, one obtains that the square distance between two mixed feature vectors $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ is

$$\rho^2_{\mathrm{EM}}((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) = \rho^2_{\mathrm{E}}(\mathbf{x}_1, \mathbf{x}_2) + \rho^2_{\mathrm{cat}}(\mathbf{y}_1, \mathbf{y}_2) \tag{3.10}$$

where $\rho^2_{\mathrm{E}}(\mathbf{x}_1, \mathbf{x}_2)$ is defined by (2.1) and $\rho^2_{\mathrm{cat}}(\mathbf{y}_1, \mathbf{y}_2)$ is defined by (3.9).

The same approach that has been applied to numerical features, will be applied here to categorical ones, namely the contribution of each attribute to the distance measure will be divided by the contribution mean. Hence, the normalized mixed metric is defined similarly to (3.1)

$$\rho^*_{\mathrm{EM}}((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) = \left(\sum_{j=1}^{m} \alpha_j (x_{1j} x_{2j})^2 + \sum_{j=1}^{l} \beta_j \omega^2(y_{1j}, y_{2j}) \right)^{1/2} \tag{3.11}$$

where $\alpha_j = 1/E(X_{1j} - X_{2j})^2$, $\beta_j = 1/E\omega^2(Y_{1j}, Y_{2j})$ and Y_{1j}, Y_{2j} are independent random variables whose values are distributed in accordance with the distribution of the A^c_jth attribute. If the attribute A^c_j can take q_j values $\{y_{j1}, y_{j2}, \ldots, y_{jq_j}\}$ and the probabilities $\{p_{j1}, p_{j2}, \ldots, p_{jq_j}\}$ of these values are known then

$$E\omega^2(Y_{1j}, Y_{2j}) = E\omega(Y_{1j}, Y_{2j}) = \sum_{\substack{r,s=1 \\ r \neq s}}^{q_j} 1 \cdot p_{jr} p_{js} = \sum_{r,s=1}^{q_j} p_{jr} p_{js} - (p^2_{j1} + \cdots + p^2_{jq_j})$$

or

$$E\omega^2(Y_{1j}, Y_{2j}) = (p_{j1} + \cdots + p_{jq_j})^2 - (p^2_{j1} + \cdots + p^2_{jq_j}) = 1 - (p^2_{j1} + \cdots + p^2_{jq_j}).$$

Thus, it follows from the above equality and (3.4) that $\alpha_j = 1/2\sigma^2_j$ and

$$\beta_j = \frac{1}{(1 - (p^2_{j1} + \cdots + p^2_{jq_j}))}. \tag{3.12}$$

(ii) *The Minkowski mixed p-metric*

Let us extend the above approach to the general case of the Minkowski metric. It follows from the Minkowski inequality that the following function is a metric (Suarez-Alvarez 2010)

$$\begin{aligned} \rho((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) = (&|x_{11} - x_{21}|^p + \cdots + |x_{1m} - x_{2m}|^p + \omega^p(y_{11}, y_{21}) \\ &+ \cdots + \omega^p(y_{1l}, y_{2l}))^{1/p}. \end{aligned} \tag{3.13}$$

It will be called the Minkowski mixed p-metric.

The normalization of the p-metric (3.13) is fulfilled in the same way as the normalization of the Euclidean mixed metric

$$\rho^*_{\mathrm{P}}((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) = \left[\sum_{j=1}^{m} \alpha_j |x_{1j} - x_{2j}|^p + \sum_{j=1}^{l} \beta_j \omega^p(y_{1j}, y_{2j}) \right]^{1/p} \tag{3.14}$$

where $\alpha_j = 1/E|X_{1j} - X_{2j}|^p$ and $\beta_j = 1/E\omega^p(Y_{1j}, Y_{2j})$. Because $E\omega^p(Y_{1j}, Y_{2j}) = E\omega(Y_{1j}, Y_{2j})$, β_j are calculated in the same way as in (3.12).

If the distribution of the attributes is unknown, then to calculate α_j one can use the estimation (3.3), and to estimate $E\omega(Y_{1j}, Y_{2j})$ one can use the sampling mean

$$\hat{E}\omega^p(Y_{1j}, Y_{2j}) = \frac{1}{N^2}\sum_{r,s=1}^{N}\omega(y_{rj}, y_{sj}). \tag{3.15}$$

The estimation (3.15) is a biased estimator of $E\omega^p(Y_{1j}, Y_{2j})$, hence for small databases it is better to use the following estimation

$$\hat{E}\omega^p(Y_{1j}, Y_{2j}) = \frac{2}{N(N-1)}\sum_{1\le r<s\le N}\omega(y_{rj}, y_{sj}), \tag{3.16}$$

which is an unbiased estimator. It can be proved (Suarez-Alvarez 2010) that (3.15) and (3.16) are consistent estimators.

(*c*) *A general algorithm for normalization of mixed metrics*

Let a metric ρ be a sum of two metrics ρ_1 and ρ_2

$$\rho((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) = \rho_1(\mathbf{x}_1, \mathbf{x}_2) + \rho_2(\mathbf{y}_1, \mathbf{y}_2).$$

First, one needs to normalize metrics ρ_1 and ρ_2, i.e. one needs to find ρ_1^* and ρ_2^*. Then, the metric ρ is normalized by

$$\rho^*((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) = \alpha_1\rho_1^*(\mathbf{x}_1, \mathbf{x}_2) + \alpha_2\rho_2^*(\mathbf{y}_1, \mathbf{y}_2).$$

Here, $\alpha_1 = m/E[\rho_1^*(\mathbf{X}_1, \mathbf{X}_2)]$ and $\alpha_2 = l/E[\rho_2^*(\mathbf{Y}_1, \mathbf{Y}_2)]$, m and l are the number of attributes in vectors $\mathbf{X}_1$ and $\mathbf{Y}_1$, respectively. The normalization of the sum of an arbitrary number of metrics can be fulfilled in the same way.

As an example, let us consider the case of a mixed metric that is the sum of the Tchebycheff metric $\rho_1 = \rho_{\max}$ applied to the numerical attributes and the metric $\rho_2 = \rho_{\text{cat}}(\mathbf{y}_1, \mathbf{y}_2)$ applied to the categorical attributes.

One can normalize this mixed metric using (3.7) and (3.11), namely

$$\rho_{\text{cat}}^*(\mathbf{y}_1, \mathbf{y}_2) = \sum_{j=1}^{l}\beta_j\omega(y_{1j}, y_{2j}),$$

where $\beta_j = 1/E\omega(Y_{1j}, Y_{2j})$.

Thus, in accordance with the described algorithm, the normalized mixed metric ρ^* is

$$\rho^*((\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2)) = \alpha\rho_{\max}^*(\mathbf{x}_1, \mathbf{x}_2) + \beta\rho_{\text{cat}}^*(\mathbf{y}_1, \mathbf{y}_2),$$

where $\alpha = m/E[\rho_{\max}^*(\mathbf{X}_1, \mathbf{X}_2)]$ and $\beta = l/E[\rho_{\text{cat}}^*(\mathbf{Y}_1, \mathbf{Y}_2)]$.

4. Applications to datasets

The above-mentioned methods will be applied to several datasets from the University of California, Irvine repository (Asuncion & Newman 2007). All

records in those datasets have class labels and, hence, 'true clustering' can be checked.

(*a*) *Accuracy of the clustering algorithm*

The traditional statistical approaches to data clustering validation are based on the use of Rand index or its modifications that give measures of classification agreement between two partitions of the same set of objects. The classic definition is the following (Rand 1971).

Let us consider a set S of N elements, and two partitions $\mathbf{C} = \{C_1, \ldots, C_k\}$ and $\mathbf{D} = \{D_1, \ldots, D_k\}$ of the dataset. To calculate the Rand index, one needs first to calculate the following numbers: a is the number of pairs of elements in S that are in the same set in $\mathbf{C}$ and in the same set in $\mathbf{D}$; b is the number of pairs of elements in S that are in different sets in $\mathbf{C}$ and in different sets in $\mathbf{D}$; c is the number of pairs of elements in S that are in the same set in $\mathbf{C}$ and in different sets in $\mathbf{D}$; and d is the number of pairs of elements in S that are in different sets in $\mathbf{C}$ and in the same set in $\mathbf{D}$. Then, the Rand index, (R), is calculated as

$$R = \frac{a + b}{a + b + c + d}.$$

The above formula is used by many researchers. It is accepted that the larger the Rand index, the higher the accuracy of the clustering. However, for datasets whose inherent structures are known in advance, other methods for checking the accuracy of the proposed clustering algorithm with normalization may be used.

Let us consider a dataset having a categorical attribute A that may have k different states $\{a_1, a_2, \ldots, a_k\}$ that may be associated with labels of the clusters. The inherent structure of the dataset is associated with the states (labels) of this attribute. The clustering algorithm will map the records to a discrete set of labels (classes).

It is proposed to perform the normalization procedure of the dataset as described above and then to apply the clustering algorithm. After clustering, each record will belong to a cluster with a corresponding number i. Let us assign a state $a_{\varphi(i)}$ of the attribute $A = \{a_1, a_2, \ldots, a_k\}$ to the ith cluster. Evidently, different clusters should have different states of the attribute A. Let us denote by $n_{i,j}$ the number of records with the attribute $A = a_j$ that belongs to the ith cluster.

For a given φ, one can estimate the accuracy $(\mathrm{Acc}(\varphi))$ of the clustering as

$$\mathrm{Acc}(\varphi) = \frac{\sum_{i=1}^{k} n_{i,\varphi(i)}}{N},$$

where $n_{i,\varphi(i)}$ is the number of records of the ith cluster whose state of the attribute A is the same as the assigned $a_{\varphi(i)}$ (Suarez-Alvarez 2010).

The clustering accuracy is defined as the maximum of $\mathrm{Acc}(\varphi)$ for all possible φ

$$\mathrm{Acc} = \max_{\varphi} \mathrm{Acc}(\varphi).$$

Evidently, the closer is Acc to 1 the less is the difference between the partitioning of the data after clustering and the partitioning of the data associated with the attribute A. If $\mathrm{Acc} = 1$, then both partitioning into classes are the same.

Thus, one need to solve the assignment problem with an efficiency matrix $n_{i,j}$, $(i, j = 1, \ldots, k)$ in order to find the clustering accuracy Acc.

(*b*) *Objective function based on the Euclidean distance*

Here, a particular case of objective function (2.5) has been used in application to particular datasets, where ρ is defined by (3.10).

First, the clustering procedure has been applied to the dataset under consideration without normalization of the data. Then, the clustering procedure with normalization of all attributes has been applied to the dataset. For each run, the initial cluster prototypes have been generated randomly. Both procedures with and without normalization have been applied 100 times to the dataset.

The clustering accuracy and the Rand index values have to be taken, not as the best values of the accuracy or the Rand index, but in accordance with the obtained best objective function value because this is the criterion of the unsupervised objective function-based clustering.

(i) *Soybean disease dataset*

This set has 47 records ($N = 47$) with 35 attributes. Each record is attributed to one of four diseases. This is a standard categorical dataset that was studied a number of times to test clustering algorithms (Michalski & Stepp 1983; Huang 1998).

Table 1 presents the cost function, the clustering accuracy and the number of attempts (N_*) whose outcome had the corresponding accuracy. Because the dataset is quite small, the 'true clustering' (Acc = 1) has been obtained quite often in both cases. Acc = 1 and $R = 1$ have been obtained in 22 per cent after clustering without normalization and in 58 per cent after clustering with normalization. The average accuracy in both cases has been 0.8636 and 0.9077, respectively, for the former and the latter cases.

(ii) *Wine dataset*

This set has 178 records ($N = 178$) with 14 attributes. The class attribute takes three categorical values, whereas the rest of the attributes are numerical. Although the dataset is quite small, the 'true clustering' (Acc = 1 or $R = 1$) has not been obtained (table 2). The values of the clustering accuracy corresponding to the obtained best objective function values are 0.702 (without normalization) and 0.966 (with normalization). After application of the normalization procedure to the dataset, the Rand index has increased from 0.7187 to 0.9543.

(iii) *Heart diseases dataset*

This set has 270 records ($N = 270$) with 14 attributes. There are seven categorical attributes, six numerical and the class attribute. The 'true clustering' (Acc = 1) has not been obtained (table 3). The values of the clustering accuracy corresponding to the obtained best objective function values is 0.593 (without normalization) and 0.804 (with normalization). After application of the normalization procedure to the dataset the Rand index has increased from 0.5154 to 0.6833.

Table 1. Clustering of the Soybean disease dataset. Here, N_* is the number of attempts whose outcome had the corresponding accuracy.

	without normalization		with normalization	
accuracy	N_*	cost function	N_*	cost function
1.00	22	199.00	58	359.87
0.98	18	199.17	14	361.36
0.96	22	199.00	0	—
0.81	2	208.00	0	—
0.77	1	214.00	1	407.33
0.75	2	215.50	2	402.64
0.73	5	219.80	5	401.13
0.71	3	237.33	9	471.64
0.69	3	237.33	0	—
0.66	3	256.33	1	494.43
0.64	6	250.17	1	499.86
0.62	12	252.75	2	492.92
0.60	1	252.00	7	492.33

Table 2. Clustering of the Wine dataset.

	without normalization		with normalization	
accuracy	N_*	cost function	N_*	cost function
0.972	0	—	16	635.788
0.966	0	—	24	635.375
0.961	0	—	20	636.388
0.955	0	—	14	636.271
0.949	0	—	23	637.629
0.702	88	2 370 689.7	0	—
0.624	0	—	1	789.640
0.596	1	2 631 657.1	0	—
0.590	0	—	1	804.144
0.579	1	2 625 223.2	0	—
0.573	10	2 633 555.3	0	—
0.528	0	—	1	793.421

(iv) *Credit approval dataset*

This set has 690 records ($N = 690$) with 16 attributes: six attributes are numerical, whereas the rest of the attributes are categorical. The results of all 100 runs of the procedure without normalization have been the same and therefore equal to the average accuracy 0.55 (table 4). After application of the normalization procedure to the dataset, the clustering accuracy corresponding to the obtained best objective function value is 0.80, and the Rand index has increased from 0.5048 to 0.6806.

Table 3. Clustering of the Heart diseases dataset.

	without normalization	with normalization
accuracy	N_*	N_*
1.00	—	—
0.826	—	7
0.822	—	1
0.819	—	3
0.815	—	1
0.807	—	43
0.804	—	25
0.796	—	12
0.593	32	—
0.589	68	—
0.559	—	1
0.556	—	1
0.526	—	1
0.522	—	3
0.519	—	1
0.507	—	1

Table 4. Clustering of the Credit approval dataset.

	without normalization	with normalization
accuracy	N_*	N_*
1.00	—	—
0.82	—	5
0.80	—	48
0.79	—	8
0.69	—	1
0.68	—	2
0.64	—	3
0.55	100	—
0.55	—	21
0.53	—	2
0.51	—	10

(*c*) *Clustering with objective functions based on the Minkowski distance*

Here, the cost function J defined by (2.6) with the Minkowski mixed p-metric (3.13) has been used. The algorithm is based on ideas that were suggested by Miyamoto & Agusta (1998) and Hathaway *et al.* (2000) as a generalization of the fuzzy clustering strategies using L_p norm distances. The algorithm was described by Suarez-Alvarez (2010). The novelty of the present

Table 5. Clustering of the Adult dataset without and with normalization of attributes for various values of the Minkowski power p.

	accuracy corresponding to the best value of the objective function	
Minkowski power, p	without normalization	with normalization
1.0	0.525	0.577
1.5	0.596	0.586
2.0	0.613	0.589
2.5	0.636	0.620
3.0	0.659	0.756
3.5	0.688	0.756
4.0	0.719	0.756
4.5	0.738	0.756
5.0	0.744	0.756

approach is that the algorithm has been developed for hard clustering using the Minkowski mixed p-metric. This algorithm has been used instead of the k-prototypes algorithm for the cases when $p \neq 2$. The algorithm has been applied 100 times to each dataset without and with normalization of attributes for various values of the Minkowski power p.

(i) *Adult dataset*

This set, also known as Census Income dataset, has 48 842 records and 30 162 records without missing values ($N = 30\,162$) with 14 attributes and one class attribute. Each record has six numerical attributes, whereas the rest of the attributes are categorical including a class attribute. Table 5 presents the Acc values corresponding to the best value of the objective function. The Acc values vary considerably with variation of p. In clustering without normalization, the best value has been obtained for $p = 5$ and it is $\mathrm{Acc} = 0.744$. After normalization, the accuracy is better and its best value has been obtained for $p = 3 \div 5$ and it is $\mathrm{Acc} = 0.756$.

(ii) *Shuttle dataset*

The full Shuttle dataset, also known as Statlog (Shuttle) dataset, has $N = 14\,500$ records with nine numerical attributes and one class attribute. As one can see from table 6, the Acc values vary considerably with variation of p. In clustering without normalization, the best value has been obtained for $p = 2.5$ and it is $\mathrm{Acc} = 0.829$. After the runs of the procedure with normalization, the best accuracy increased to $\mathrm{Acc} = 0.852$ for $p = 3$. One can see that it is more advantageous to apply the general Minkowski metrics to the Shuttle dataset than a particular case $p = 2$ (the Euclidean metric).

Thus, the new normalized metrics has been used for clustering mixed data. It has been shown that when clustering has been performed using normalized metrics, the accuracy has increased in all considered examples. These examples

Table 6. Clustering of the Shuttle dataset without and with normalization of attributes for various values of the Minkowski power p.

	accuracy corresponding to the best value of the objective function	
Minkowski power, p	without normalization	with normalization
1.0	0.451	0.461
1.5	0.471	0.472
2.0	0.695	0.672
2.5	0.829	0.715
3.0	0.791	0.852
3.5	0.791	0.846
4.0	0.791	0.791

have demonstrated the advantages of the introduced normalized metrics. Other examples of clustering of various databases with and without normalization of attributes were given by Suarez-Alvarez (2010).

Amorim & Mirkin (2012) have published an alternative algorithm for clustering of datasets, the so-called intelligent Minkowski metric weighted k-means (iMWK-means) algorithm that is a development of the weighted k-means (WK-means) introduced by Huang *et al.* (2005) and the 'intelligent' version of k-means (Mirkin 2005) algorithms. Similar to the WK-means algorithm that prescribes the specific weights to all features that minimize the objective function, the iMWK-means algorithm minimizes the cost function $\sum \rho_p^p$ where metric is taken between rescaled cluster points and prototypes. Because the described statistical normalization procedure can be treated as a kind of weighting, one could expect that the iMWK-means algorithm would always overperform the algorithm that uses just the statistical normalization. In fact, the iMWK-means algorithm shows better results in application to several benchmark datasets, e.g. it gives $\mathrm{Acc} = 0.8407$ for $p = 2.7$ in application to the Heart disease dataset (Asuncion & Newman 2007), whereas for the same dataset the best result of the algorithm with statistical normalization is $\mathrm{Acc} = 0.8185$ for $p = 1$ and $p = 3.5$. Both algorithms gave the same result $\mathrm{Acc} = 0.8037$ for $p = 2$. However, for the Pima Indian Diabetes dataset (Asuncion & Newman 2007), the algorithm with statistical normalization shows slightly better result $\mathrm{Acc} = 0.7083$ for $p = 3$ than $\mathrm{Acc} = 0.6940$ for $p = 4.9$ shown by the iMWK-means algorithm. For the above-mentioned Wine dataset, the iMWK-means algorithm shows $\mathrm{Acc} = 0.949$ for $p = 1.2$, $\mathrm{Acc} = 0.921$ for $p = 2$ and $\mathrm{Acc} = 0.938$ for $p = 3$, whereas the algorithm with statistical normalization shows $\mathrm{Acc} = 0.955$ for $p = 1.5$, $\mathrm{Acc} = 0.9663$ for $p = 2$ and $\mathrm{Acc} = 0.927$ for $p = 3$ (table 7). The detailed comparison of performances of these algorithms and detailed analysis of the results are out of the scope of this study. It looks quite natural to apply the weighting procedure to metrics that have already been normalized by the above-described procedure. If one knows *a priori* that some attributes have larger contributions to similarity measures than the rest of the attributes then this can be taken into account by appropriate weighting.

Table 7. Comparison of accuracy of the iMWK-means and the present algorithm with normalization for several benchmark datasets.

	accuracy									
	iMWK-means					present algorithm				
	p					p				
dataset	1.2	2	2.7	3	4.9	1	1.5	2	3	3.5
Heart disease	—	0.8037	0.8407	—	—	0.819	0.807	0.804	0.815	0.819
Pima Indian Diabetes	—	—	—	—	0.694	0.577	0.663	0.676	0.708	0.704
Wine	0.949	0.921	—	0.938	—	0.961	0.955	0.966	0.927	0.921

5. Conclusion

In this study, a mixed database is treated as a random sample of an object under consideration. In the overwhelming majority of the earlier approaches to normalization, scaling was used for the Euclidean metric and/or the normal distribution of the variables in order to assure the values to be in the $[0,1]$ range. However, it has been shown that, in general, this does not provide equal contributions of the features to the metrics.

A unified statistical approach has been applied to normalization of all attributes of the feature vectors of mixed datasets that assures that the means of the different normalized attributes are equal to each other and therefore, these variables give equal contributions to the similarity measures. The presented approach extends the ideas considered earlier (Hastie *et al.* 2001; Pham *et al.* 2006). The normalization is achieved by scaling the numerical attributes, whereas the categorical attributes are normalized by appropriate choice of their weights according to described statistical procedure. Normalized Minkowski metrics and metrics for mixed datasets have been introduced in an explicit way.

Following the general principle of equal importance of features, it has been shown that the classic min–max normalization (2.7) of numerical attributes is the proper one from a statistical point of view in the case of the Tchebycheff metric, whereas the z-score normalization should be used in the case of the Euclidean metric.

The main idea of normalization that has been developed in this study is valid not only for the above mentioned specific metrics, but also in application to any other measure. Hence, a general procedure for normalization of mixed metrics has been introduced.

The introduced normalized metrics have been applied to various datasets. Results on benchmark datasets are presented together with a comparison with other approaches. Clustering has been performed with non-normalized and introduced normalized mixed metrics using either the k-prototypes (for $p=2$) or another algorithm (for $p\neq 2$). It has been observed that the accuracy has increased in all considered examples when clustering has been performed using

normalized metrics. It has also been shown on examples that sometimes it is more advantageous to use objective functions based on the general Minkowski metrics than on the Euclidean metric.

We thank Prof. Feodor M. Borodich (Cardiff University) for his valuable comments on the study.

References

Aksoy, S. & Haralick, R. M. 2001 Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognit. Lett.* **22**, 563–582. (doi:10.1016/S0167-8655(00)00112-4)

Amorim, R. C. & Mirkin, B. 2012 Minkowski metric, feature weighting and anomalous cluster initializing in K-means clustering. *Pattern Recognit.* **45**, 1061–1075. (doi:10.1016/j.patcog.2011.08.012)

Asuncion, A. & Newman, D. J. 2007 UCI Machine learning repository. University of California, CA: School of Information and Computer Science. See http://www.ics.uci.edu/mlearn/MLRepository.html.

Bachner, J. 2000 A probabilistic clustering model for variables of mixed type. *Qual. Quant.* **34**, 223–235. (doi:10.1023/A:1004759101388)

Bezdek, J. C. 1980 A convergence theorem for the Fuzzy ISODATA clustering algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2**, 1–8. (doi:10.1109/TPAMI.1980.4766964)

Bock, H. H. 1996 Probabilistic models in cluster analysis. *Comput. Stat. Data Anal.* **23**, 5–28. (doi:10.1016/0167-9473(96)88919-5)

Chen, C.-H. 1973 *Statistical pattern recognition*. Rochelle Park, NJ: Hayden Book Company.

Cios, K. J., Pedrycz, W., Swiniarski, R. W. & Kurgan, L. A. 2007 *Data mining: a knowledge discovery approach*. New York, NY: Springer.

Dempster, A., Laird, N. & Rubin, D. 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**, 1–38.

Doherty, K., Adams, R. & Davey, N. 2004 Non-Euclidean norms and data normalization. In *Proc. 12th Euro. Symp. Artificial Neural Networks* (ed. M. Verleysen), *Bruges*, *Belgium*, *28–30 April 2004*, pp. 181–186. Bruges, Belgium: d-side publications.

Fraley, C. & Raftery, A. E. 2002 Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **97**, 611–631. (doi:10.1198/016214502760047131)

Gan, G., Ma, Ch. & Wu, J. 2007 *Data clustering: theory, algorithms, and applications*. Philadelphia, PA: SIAM.

Gibert, K. & Cortés, U. 1997 Weighing quantitative and qualitative variables in clustering methods. *Mathware Soft Comput.* **4**, 251–266.

Hathaway, R. J., Bezdek, J. C. & Hu, Y.-K. 2000 Generalized fuzzy c-means clustering strategies using L_p norm distances. *IEEE Trans. Fuzzy Syst.* **8**, 576–582. (doi:10.1109/91.873580)

Hastie, T., Tibshirani, R. & Friedman, J. 2001 *The elements of statistical learning. Data mining, inference, and prediction*. Berlin, Germany: Springer.

Huang, Z. 1997 Clustering large data sets with mixed numeric and categorical values. In *Proc. the First Pacific Asia Knowledge Discovery and Data Mining Conference*, pp. 21–34. Singapore: World Scientific.

Huang, Z. 1998 Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowl. Discov.* **2**, 283–304. (doi:10.1023/A:1009769707641)

Huang, J. Z., Ng, M. K., Rong, H. & Li, Z. 2005 Automated variable weighting in k-means type clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 657–668. (doi:10.1109/TPAMI.2005.95)

Jain, A. K. & Dubes, R. C. 1988 *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.

Kamath, C. 2009 *Scientific data mining: a practical perspective*, Philadelphia, PA: SIAM.

Kaufman, L. & Rousseeuw, P. J. 2005 *Finding groups in data: an introduction to cluster analysis*. New York, NY: Wiley.

Larose, D. T. 2005 *Discovering knowledge in data: an introduction to data mining*. Hoboken, NJ: Wiley.

MacQueen, J. B. 1967 Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Math. Stat. Probab.*, pp. 281–297. Berkeley, CA: University of California Press. See http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.bsmsp/1200512992.

Michalski, R. S. & Stepp, R. E. 1983 Automated construction of classifications: conceptual clustering versus numerical taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **5**, 396–410. (doi:10.1109/TPAMI.1983.4767409)

Milligan, G. W. & Cooper, M. C. 1988 A study of variable standardization in cluster analysis. *J. Classif.* **5**, 181–204. (doi:10.1007/BF01897163)

Mirkin, B. 1996 *Mathematical classification and clustering*. Dordrecht, The Netherlands: Kluwer Academic Press.

Mirkin, B. 1997 L_1 and L_2 approximation clustering for mixed data: scatter decompositions and algorithms L_1. *Stat. Proc. Related Top.* **31**, 473–486.

Mirkin, B. 1998 Least-squares structuring, clustering, and data processing issues. *Comput. J.* **41**, 518–536. (doi:10.1093/comjnl/41.8.518)

Mirkin, B. 2005 *Clustering for data mining: a data recovery approach*. Boca Raton FL: Chapman and Hall/CRC.

Miyamoto, S. & Agusta, Y. 1998 Algorithms for L_1 and L_p fuzzy c-means and their convergence. In *Studies in classification, data analysis, and knowledge organization; data science, classification, and related Methods* (eds C. Hayashi, N. Ohsumi, K. Yajima, Y. Tamaka, H. H. Bock & Y. Baba), pp. 295–302. Tokyo, Japan: Springer.

Pham, D. T., Prostov, Y. I. & Suarez-Alvarez, M. M. 2006 Statistical approach to numerical databases: clustering using normalised Minkowski metrics. In *Intelligent Production Machines and Systems. Proc. of 2nd I*PROMS Virtual Conference, 3–14 July 2006*, pp. 356–361. Amsterdam, The Netherlands: Elsevier.

Pham, D.-T., Suarez-Alvarez, M. M. & Prostov, Y. I. 2011 Random search with k-prototypes algorithm for clustering mixed datasets. *Proc. R. Soc. A* **467**, 2387–2403. (doi:10.1098/rspa.2010.0594)

Ralambondrainy, H. 1995 A conceptual version of the K-means algorithm. *Pattern Recognit. Lett.* **16**, 1147–1157. (doi:10.1016/0167-8655(95)00075-R)

Rand, W. M. 1971 Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850. (doi:10.2307/2284239)

Suarez-Alvarez, M. M. 2010 Design and analysis of clustering algorithms for numerical, categorical and mixed data. PhD thesis, Cardiff University, Cardiff, UK.

Zhigljavsky, A. A. & Žilinskas, A. G. 2008 *Stochastic global optimization*. Berlin, Germany: Springer.