A Variant of the K-Means Clustering Algorithm for Continuous-Nominal Data

Aleksander Denisiuk and Michał Grabowski

Abstract The core idea of the proposed algorithm is to embed the considered dataset into a metric space. Two spaces for embedding of nominal part with the Hamming metric are considered: Euclidean space (the classical approach) and the standard unit sphere $\mathbb{S}$ (our new approach). We proved that the distortion of embedding into the unit sphere is at least 75 % better than that of the classical approach. In our model, combinations of continuous and nominal data are interpreted as points of a cylinder $\mathbb{R}^p \times \mathbb{S}$, where p is the dimension of continuous data. We use a version of the gradient algorithm to compute centroids of finite sets on a cylinder. Experimental results show certain advances of the new algorithm. Specifically, it produces better clusters in tests with predefined groups.

1 Introduction

From the very beginning we define a dissimilarity function or a metric on combinations of continuous and nominal (categorical) data. There is a huge collection of dissimilarity functions on vectors of nominal data, used in data exploration. For example, the Hamming distance, the Jaccard distance, the distance defined after the Bayesian numerical codding of nominal values, and other concepts [6, 8]. In this paper we follow the approach with the Hamming distance. Let (x, n) be a record of continuous (x) and nominal (n) data, where $x \in \mathbb{R}^p$. We define metric on the space of such records as $\operatorname{dist}\big((x_1, n_1), (x_2, n_2)\big) = K\left(d(x_1, x_2), H(n_1, n_2)\right)$, where

A. Denisiuk (✉)
University of Warmia and Mazury in Olsztyn, Olsztyn, Poland
e-mail: denisiuk@matman.uwm.edu.pl

M. Grabowski
Warsaw School of Computer Science, Warsaw, Poland
e-mail: mgrabowski@poczta.wwsi.edu.pl

R. Burduk et al. (eds.), *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015*, Advances in Intelligent Systems and Computing 403, DOI 10.1007/978-3-319-26227-7_2

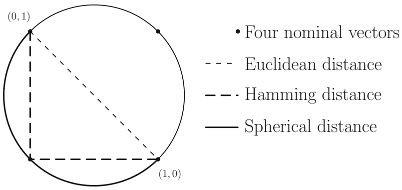

Fig. 1 The classical embedding of the Hamming metric of the Cartesian product of two-elements nominal domain $\{0, 1\} \times \{0, 1\}$ into the Euclidean plane. One can see that normalized spherical distance coincides with the Hamming distance

$d(x_1, x_2)$ is the standard Euclidean distance, $H(n_1, n_2)$ is the Hamming distance, and $K : \mathbb{R}^2 \to \mathbb{R}^+$ is an appropriate function. It can be observed [5] that such an approach produces classification results that are not worse than those of Bayesian numerical codding of nominal values. Thus, the Hamming metric may be considered as a sufficiently strong option. We accept a metric of this form and suggest a suitable function K in Sect. 2. To perform the k-means algorithm, one should be able to measure a distance between two records of data and to compute a centroid of a finite set of data records. The embedding of the considered dataset into a metric space equipped with a method of computing centroids is the core idea of this paper. We search for a relevant space by embedding the Hamming metric space of nominal data into a Riemannian manifold with possibly small distortion. The classical approach, representing nominal values as equidistant vertexes of a simplex, can be considered as embedding of the Hamming metric space into the Euclidean space. In general, isometric embedding of the Hamming metric into the Euclidean space is not possible [7, 9]. In Sect. 3 we analyze the distortion of this embedding. Two- and three-dimensional examples suggest that embedding of nominal values into a sphere has a distortion that is less than distortion of classical embedding into Euclidean space (Fig. 1). In Sect. 4 we prove that this is a general phenomenon: every Hamming metric space $(A_1 \times \cdots \times A_s, d_H)$ can be embedded into a suitable multidimensional sphere $\mathbb{S}$ with distortion that is better than the distortion of embedding it into the Euclidean space. We give a quantitative measure of this distortion improvement (Theorems 3 and 4), which is at least 75 % better than the distortion of embedding into the Euclidean space. We represent combinations of continuous and nominal data (x, n) as points on a cylinder $\mathbb{R}^p \times \mathbb{S}$, where p is the dimension of continuous data. To compute centroids of a finite subset of the cylinder, a version of the gradient algorithm with respect to the cylinder metrics is proposed. The experimental results (Sect. 5) show that the within index of clusters, produced by both algorithms are comparable. On the other hand, our algorithm has an advantage with respect to the degree of covering of predefined groups.The authors thank the participants of the CORES2015 conference for interesting discussions, related to the article.

2 Technical Preliminaries

In this section the basic notations, definitions, and algorithms are introduced. Specifically, we formulate a general scheme of the k-means clustering algorithm in an abstract metric space. This scheme is instanced in Sects. 3 and 4 by embedding the Hamming metric into the Euclidean space and the unit sphere respectively.

Definition 1 Let $A_1, \ldots, A_s$ be finite sets of nominal values, $H = A_1 \times \cdots \times A_s$. The Hamming metric on H is defined as follows: for each two vectors $n, n' \in H$,

$$d_H(n, n') = s^{-1}\Big|\{ i = 1, \ldots, s | n_i \neq n'_i \}\Big|, \tag{1}$$

where $n = (n_1, \ldots, n_s)$, $n' = (n'_1, \ldots, n'_s)$.

Definition 2 (*cf.* [9]) Let (U, d) and (U', d') be two metric spaces. A mapping $f : U \to U'$ is said to have a distortion of at most c, if there exists a constant $\nu \in (0, +\infty)$ such that for all $x, y \in U$

$$\nu \cdot d(x, y) \leq d'(f(x), f(y)) \leq c \cdot \nu \cdot d(x, y). \tag{2}$$

In other words, the mapping f can magnify the relative distances between points not greater than c times. In this paper we consider datasets combining both continuous and nominal data. To represent continuous data we use points from the standard p-dimensional Euclidean metric space $(\mathbb{R}^p, d_E)$. The nominal part of record is represented as a point at the Hamming metric space $H = A_1 \times \cdots \times A_s$. The set $X = \mathbb{R}^p \times H$ is considered as a set of all records. Every record $r \in X$ is a pair of continuous and nominal data, $r = (x, n)$. We admit the following metric on X:

$$\mathrm{dist}^2\big((x, n), (x', n')\big) = d_E^2(x, x') + d_H^2(n, n')$$

We end this section with a general scheme of the k-means clustering algorithm. Let (M, d_M) be a metric space.

Definition 3 A point $m \in M$ is said to be a *centroid* of a finite set $\{ x_1, \ldots, x_N \} \subset M$, if $\sum_{j=1}^{N} d_M^2(m, x_j) = \min_{x \in M} \sum_{j=1}^{N} d_M^2(x, x_j)$.

We assume that we have some algorithm for computation of centroids of finite subsets of M. Let $\imath : X \to M$ be an embedding map. The $(M, \imath)$ algorithm is defined in algorithm 1. In what follows we use for $y, y' \in \mathbb{R}^p$ the following: notations $\langle y, y' \rangle = \sum_{j=1}^{p} y_j y'_j$, $\|y\|^2 = \sum_{j=1}^{p} y_j^2$.

Algorithm 1 A general k-means algorithm

Require: k is the number of clusters, $T \subset X$ is a finite set of records
Ensure: The resulting clusters C_1, ..., C_k in the data set T
  $MT = i(T) \subset M$
  Choose randomly initial centroids $m_1, \ldots, m_k \in M$
  {Compute initial clusters $MC_1, \ldots MC_k \subset M$}
  for all $x \in MT$ **do**
    x is classified to the cluster MC_j, if $d_M(x, m_j) = \min(d_M(x, m_1), \ldots, d_M(x, m_k))$
  end for
  while The current set of clusters $MC_1, \ldots, MC_k$ is not stabilized **do**
    for all Current clusters $MC_1, \ldots, MC_k$ **do**
      Compute new centroids $m_1, \ldots, m_k \in M$
      {Compute new clusters $MC_1, \ldots MC_k \subset M$}
      for all $x \in MT$ **do**
        x is classified to MC_j, if $d_M(x, m_j) = \min(d_M(x, m_1), \ldots, d_M(x, m_k))$
      end for
    end for
  end while
Return: $C_1 = \imath^{-1}(MC_1), \ldots, C_k = \imath^{-1}(MC_k)$

3 Euclidean Embedding of the Hamming Metric

The Hamming metric space cannot be isometrically mapped into the Euclidean space (v. [7]). However, it can be embedded with a particular distortion. Distortion analysis of embeddings of finite metric spaces into Euclidean spaces is actually hard [7, 9]. Bourgaine in [1] proved that every m-point metric space can be embedded into the $O(\log^2 m)$-dimensional Euclidean space with distortion $O(\log m)$. Consider the following well-known, folklore-type idea: map each of the m values of the nominal domain into a vertex of a simplex of equidistant points in $\mathbb{R}^{m-1}$. We expand this mapping to the space of nominal records:

$$\phi_{sim} : A_1 \times \cdots \times A_s \to \mathbb{R}^{a_1-1} \times \cdots \times \mathbb{R}^{a_s-1},$$

where $|A_j| = a_j$, $j = 1, \ldots, s$.

Theorem 1 *The distortion of ϕ_{sim} is not greater than $\sqrt{\sum(a_j - 1)}$.*

The simple proof is omitted. The above estimation of distortion is optimal. We can deduce Theorem 1 as well as optimality of this estimation from the following theorem by Enflo [2].

Theorem 2 *The standard embedding of s-dimensional unit cube with graph metric has optimal distortion, which is equal to $\sqrt{s}$.*

Let $q = \sum_{j=1}^{s}(a_j - 1)$. We map a data record $r = (x, n) \in X$ into the Euclidean space $\mathbb{R}^p \times \mathbb{R}^q$ as follows: $\phi_E(x, n) = (x, \phi_{sim}(n))$. Considering this embedding as $\imath$ and $\mathbb{R}^p \times \mathbb{R}^q$ with the standard Euclidean metric as M (with obvious computation of centroids as means), we instance algorithm 1. Thus we arrive at Euclidean k-means clustering of combinations of continuous and nominal data by embedding the Hamming metric into the Euclidean space. The main purpose of this paper is to compare this approach with embedding of the Hamming metric into a sphere, with better distortion.

4 Spherical Embedding of the Hamming Metric

In this section we consider another metric space M, the cylinder $\mathbb{R}^p \times \mathbb{S}$, where $\mathbb{S}$ is the unit sphere. We define embedding of the Hamming metric space into the sphere and analyze the distortion of this embedding (Theorems 3 and 4). Theorems 2 and 3 imply that the distortion of the spherical embedding is much better than the distortion of Euclidean embedding. We end this section with an algorithm of calculation of centroids. We show that binary Hamming s-dimensional set can be embedded into $s-1$ dimensional sphere with distortion less than that of Enflo's theorem.

Theorem 3 *For any $s > 1$ an s-dimensional cube $A = \{-1, 1\}^s$ with Hamming distance can be embedded into the standard unit sphere $\mathbb{S}^{s-1}$ with distortion less than $\lambda^{-1}\sqrt{s}$, where $\lambda^{-1} < 3/4$ is the constant, defined in (5).*

Proof Place elements of A into corresponding vertexes of the standard cube $[-1, 1]^s$ in $\mathbb{R}^s$. Consider the circumscribed sphere and project it to the unit sphere $\mathbb{S}^{s-1}$, centered at the origin. The spherical distance between embedded vertexes is the angle measure between corresponding cube vertexes. Using the Euclidean scalar product, we obtain for $\theta = \widehat{e, f}$

$$\cos\theta = \frac{\sum e_i f_i}{s} = \frac{s - 2|\{i | e_i \neq f_i\}|}{s} = 1 - 2d_H(e, f),$$

since all the coordinates of e and f are ± 1. Therefore,

$$d_s(e, f) = \arccos(1 - 2d_H(e, f)), \tag{3}$$

where d_s is the spherical distance and d_H is the Hamming distance. To estimate the distortion of this embedding, note that possible values for the Hamming distances are: $\frac{1}{s}, \frac{2}{s}, \ldots, \frac{s-1}{s}, 1$. Therefore, possible values for the spherical distances are: $\arccos(1 - 2 \cdot \frac{1}{s})$, $\arccos(1 - 2 \cdot \frac{2}{s})$, …, $\arccos(1 - 2 \cdot \frac{s-1}{s})$, π. So, the distortion σ satisfies inequality $\sigma < M/m$, where

$$M = \max_{k=1,\ldots,s} \frac{\arccos(1 - 2 \cdot \frac{k}{s})}{k/s}, \quad m = \min_{k=1,\ldots,s} \frac{\arccos(1 - 2 \cdot \frac{k}{s})}{k/s}. \tag{4}$$

To estimate M and m in (4), consider a function $f(t) = \arccos(1 - t)/t$, $t \in (0, 2)$. By analyzing its first and second derivatives, one can prove that f is convex on $(0, 2)$

and decreases for $t \in (0, 1)$. Since $f(1) = f(2) = \frac{\pi}{2}$, $f(t)$ has the only minimum, belonging to $(1, 2)$. Hence $M = s \arccos(1 - 2/s)$, $m > 2\lambda$, where

$$\lambda = \min_{t \in [1,2]} \arccos(1 - t)/t. \tag{5}$$

To estimate M, we make use of the inequality

$$\theta < 2 \tan \frac{\theta}{2} = 2\sqrt{\frac{1 - \cos\theta}{1 + \cos\theta}},$$

where $\theta = \arccos(1 - 2/s)$, and therefore $\cos\theta = 1 - 2/s$. So, $M < 2\frac{s}{\sqrt{s+1}} < 2\sqrt{s}$. Therefore $\sigma < \lambda^{-1}\sqrt{s}$, where λ is defined in (5). Since $f(t)$, defining λ, is strictly convex, one can numerically estimate its value: $\lambda \approx 1.380$, therefore $\lambda^{-1} < 3/4$.

Now let us consider a general case for any finite set.

Theorem 4 *Let $|A_j| = a_j$ for $j = 1, \ldots, s$. Then $A = A_1 \times \cdots \times A_s$ with the standard Hamming metrics d_h (1) can be embedded into the standard unit sphere $\mathbb{S}^{q-1}$ ($q = \sum_{j=1}^{s} a_j - s$) with distortion less than*

$$\frac{1}{\lambda}\sqrt{2\frac{a_{\min}}{a_{\min} - 1}\frac{a_{\max} - 1}{a_{\max}}}\sqrt{s}, \tag{6}$$

where $a_{\min} = \min_{j=1,\ldots,s} a_j$, $a_{\max} = \max_{j=1,\ldots,s} a_j$, λ is the constant (5), $\lambda^{-1} < 3/4$.

Proof The set A_j can be embedded into $\mathbb{R}^{a_j - 1}$ by placing its elements into vertexes of a simplex, centered at the origin. Let $(e_1, \ldots, e_{a_j - 1})$ be the coordinates of certain vertex of the simplex. We will assume that $\sum e_i^2 = 1$,. The desired embedding of $A_1 \times \cdots \times A_s$ to $\mathbb{R}^q$ is defined as follows:

$$\phi_s : n = (n_1, \ldots, n_s) \mapsto (e_1^{(1)}, \ldots, e_{k_1}^{(1)}, \ldots, e_1^{(s)}, \ldots, e_{k_s}^{(s)}), \tag{7}$$

where $k_j = a_j - 1$ and $e_1^{(j)}, \ldots, e_{k_j}^{(j)}$ are the coordinates of n_j on the corresponding simplex ($j = 1, \ldots, s$). One can see that the image of n belongs to the sphere of radius $\sqrt{s}$, i.e. $\|y\|^2 = s$. The distance between two points y and y' on the sphere is the angular measure between corresponding vertexes. For the Euclidean scalar product we obtain: $\langle y, y' \rangle = \lambda_1 + \cdots + \lambda_s$, where

$$\lambda_j = \begin{cases} 1, & \text{if } y_j = y'_j, \\ \frac{1}{1 - a_j}, & \text{if } y_j \neq y'_j. \end{cases}$$

Here $\frac{1}{a_j - 1}$ is the cosine of the angle between corresponding simplex vertexes. If y and y' differ in k places, $d_H(y, y') = k/s$, then the possible values for $d_s(y, y')$ are

$$d_s(y, y') = \arccos\left(1 - \frac{1}{s}\left(k + \sum\nolimits_{i=1}^{k} \frac{1}{a_{j_i} - 1}\right)\right). \tag{8}$$

To estimate the distortion, we should estimate the minimum and maximum of $d_s(y, y')/k$ for all $k = 1, \dots, s$ and all k-tuples $j_1, \dots, j_k$. Since $\arccos(1 - t)$ increases, one can substitute a_j in (8) by a minimal value for estimation of $M = \max d_s(y, y')/k$, and by a maximal value for estimation of $m = \min d_s(y, y')/k$. Then, repeating the estimates from Theorem 3, we get

$$M \le \sqrt{2s \frac{a_{\min}}{a_{\min} - 1}} \quad m \ge \lambda \frac{a_{\max}}{a_{\max} - 1},$$

These estimates immediately give (6).

The embedding of a record $r = (x, n)$ into the cylinder $\mathbb{R}^p \times \mathbb{S}^{q-1}$ is defined as follows: $(x, n) \mapsto (x, y) = (x, \phi_s(n)) = (x, y)$, where ϕ_s is defined in (7), $\|y\| = 1$. The distance between two records on the cylinder is defined by the following formula: $\operatorname{dist}^2\big((x, y), (x', y')\big) = \sum_{j=1}^{p} \big(x_j - x'_j\big)^2 + \kappa \arccos^2\big(\langle y, y'\rangle\big)$, where κ is a scaling coefficient. Let a finite set X of records, embedded into the cylinder be given as $(x^{(j)}, y^{(j)}) = (x_1^{(j)}, \dots, x_p^{(j)}, y_1^{(j)}, \dots, y_q^{(j)})$, where $\|y^{(j)}\| = 1$, and $j = 1, \dots, N$. To compute a centroid $(\bar{x}, \bar{y})$ we should minimize the following expression:

$$\sum_{j=1}^{N} \sum_{k=1}^{p} \big(x_k - x_k^{(j)}\big)^2 + \kappa \sum_{j=1}^{N} \arccos^2\big(\langle y, y^{(j)}\rangle\big).$$

So, the minimization splits into two independent problems. Minimization with respect to x gives the standard mean. For the second term we suggest to use a gradient method. This is summarized in Algorithm 2.

Algorithm 2 Calculation of centroid of the finite set on a sphere

Require: $(x^{(j)}, y^{(j)}) = (x_1^{(j)}, \dots, x_p^{(j)}, y_1^{(j)}, \dots, y_q^{(j)})$, where $\|y^{(j)}\| = 1$, $j = 1, \dots, N$
Ensure: $(\bar{x}, \bar{y})$ is a centroid with the accuracy ε
  $\bar{x} = N^{-1} \sum_{j=1}^{N} x^{(j)}$
  Zero approximation: $\bar{y} = \left(\sum_{j=1}^{N} y^{(j)}\right) / \left\|\sum_{j=1}^{N} y^{(j)}\right\|$.
  if $\sum_{j=1}^{N} y^{(j)} = 0$ **then**
    choose one of $y^{(j)}$ (of maximal multiplicity).
  end if
  repeat {The next approximation}
    $y' = \frac{y - \omega \Delta y}{\|y - \omega \Delta y\|}$, where ω is a relaxation parameter, $\Delta y = \nabla \left(\sum \arccos^2\big(\langle y, y^{(j)}\rangle\big)\right)$
  until $\|y' - y\| > \varepsilon$

5 Experimental Results

Clustering validation is a very subtle issue. There is a clear measure of success in the classification problem: cross-validation as estimation of expected loss over the joint probability distribution on data × decision. In the context of clustering, we have no such direct measure of success. As Hastie, Tibshirani, Friedman put it (v. [6]): *This uncomfortable situation has led to heavy proliferation of proposed methods, since effectiveness is a matter of opinion and cannot be verified directly.* The approach with predefined clusters seems to be credible (to some extent). Some predefined groups of data are considered as intended clusters. The mean covering degree of predefined groups by computed clusters may be considered as the clustering validity index. The degree of covering A by B is defined as $|A \cap B|/|B|$. We analyze two training sets with decision categories as intended clusters [4]: the *Heart Disease* (6 continuous, 7 nominal attributes, 370 records, 2 decision categories) and the *Australian Credit Approval* (6 continuous, 8 nominal attributes, 690 records, 2 decision categories). The following artificial datasets are analyzed as well:

ADS1: The set of records $X = \mathbb{R} \times \{0, 1\}$—one continuous attribute and one nominal attribute with two-elements domain. We define ADS1 as the union of two intended clusters $A \times \{0\} \cup B \times \{1\}$, where the elements of A and B have been randomly chosen according to the normal distributions $N(170, 10)$ and $N(155, 10)$ respectively. Each of A and B contains 300 real numbers.

ADS2: The set of records $X = \mathbb{R}^4 \times A_1 \times A_2 \times A_3 \times A_4 \times A_5$, where nominal domains are of arbitrarily chosen cardinalities: 3, 2, 4, 5 and 2 respectively. Let $\mathbb{B}(x, r) = \{ y \in X | \text{dist}(x, y) \leq r \}$ be a ball with a center x and radius r. Choose randomly two elements $x_1, x_2 \in X$, $r = \text{dist}(x_1, x_2)/2$. Let $A, B \subset X$ be two sets of 200 uniform randomly chosen elements from balls $\mathbb{B}(x_1, r)$ and $\mathbb{B}(x_2, r)$ respectively. The disjoint sets A, B are considered as two intended, not well separated, clusters. We define ADS2 as $A \cup B$. The dataset ADS2 was randomly generated 10 times in experiments.

The following two clustering validity indexes appear to make sense in the context of k-means clustering: the number of iterations stabilizing clustering and the within index W, which, for $C_1, \ldots, C_k \subset X$ is defined as $W = 1/2 \cdot \sum_j \sum_{r,r' \in C_j} \text{dist}^2(r, r')$. We compare the following variants of k-means clustering:

MC: k-medoids clustering: the centroids are records from the analyzed dataset.

EC: Euclidean k-means: with classical embedding of the Hamming metric into the Euclidean space.

CC: Cylindrical k-means: the Hamming metric is embedded into the sphere.

In order to maintain a balance between continuous and Hamming parts of the metric, all the continuous attributes were renormalized to the normal Gaussian distribution $N(0, 1)$ or to the segment $[0, 1]$. The results of numerical experiment are presented in Table 1. We do not present results for the within index W and for number of iterations, stabilizing clustering. These results do not indicate any advantage of one

Table 1 The average covering degree for tested datasets

Dataset	MC	EC	CC
ADS1	0.93	0.98	1
ADS2–I	0.78	0.84	0.91
ADS2–II	0.88	1	1
ADS2–III	0.92	0.95	0.96
ADS2–IV	1	0.99	1
ADS2–V	0.9	0.9	1
ADS2–VI	0.79	0.84	0.93
ADS2–VII	0.85	0.88	0.97
ADS2–VIII	0.85	0.94	1
ADS2–IX	0.92	0.91	0.95
ADS2–X	0.95	0.93	0.95
Heart Disease	0.72	0.80	0.86
Australian Credit Approval	0.69	0.79	0.83

Each algorithm was executed 30 times with two randomly chosen initial centroids on each dataset

clustering (EC, CC) over another, with respect to these indexes. The only observed phenomenon is that MC clustering on datasets ADS2 is visibly worse with respect to the index W. We can see that the cylindrical clustering has better covering degree in all datasets. Note that it is very stable on ADS1 with respect to choice of initial centroids; in this case the normalized spherical metric coincides with the Hamming metric.

6 Conclusions and Final Remarks

The explored datasets showed the advantage of cylindrical clustering over two other analyzed methods with respect to the degree of covering of predefined groups. However, we do not see a hard relation between the distortion improvement and the resulting values of the within index W. None of the considered clustering algorithms (Euclidean, cylindrical) has a definite advantage over the other with respect to the within index W. The k-medoids algorithm produces visibly worse results compared to the Euclidean and the cylindrical k-means algorithms. At least on the analyzed datasets and with respect to the here accepted clustering validation indexes. In our opinion, the cylindrical k-means clustering is worth considering as one of the alternatives, when several algorithms are tested in order to find relevant, interpretable clusters of continuous-nominal data. Our approach is applicable to those datasets where information hidden in nominal data influences relevant clusters by the Hamming metric. Many other ways of such nominal information influence are possible, for instance, via leading approaches by probabilistic models like attribute information role in the classical version of COBWEB algorithm [3]. If real clusters are not (roughly) spherical with respect to the metric of the data space, then the resulting clusters, produced by our approach, may not be credible. The same objection is valid for the Euclidean embedding. Perhaps it is worth to analyze the embedding of Hamming metric into the Euclidean space using the idea of Bourgaine's theorem [1], since the dimension of this space is significantly less than the dimension of Euclidean spaces and of spheres used by our approach. This is a possible future work as well as a deeper experimental study. We have analyzed only two geometrical structures: Euclidean space and sphere. There could be other possibilities, with distortion that is significantly better than that of spherical embedding. In general, the following hard question is opened up: does there exist a Riemannian manifold providing really good k-means clustering of continuous-nominal data with respect to the within index W? Note that Riemannian metric works well in some data exploration contexts, for instance, a version of adaptive k-nearest neighbors algorithm based on Riemannian metric [10].

References

1. Bourgaine, J.: On lipschitz embeddings of finite metric spaces in Hilbert space. Isr. J. Math. **52**, 46–52 (1985)
2. Enflo, P.: On the nonexistence of uniform homeomorphisms between Lp spaces. Ark. Mat. **8**, 5–103 (1969)
3. Fisher, D.H.: Knowledge acquisition via incremental conceptual clustering. Mach. Learn. **2**, 139–172 (1987)
4. Frank, A., Asuncion, A.: UCI Machine Learning Repository http://archive.ics.uci.edu/ml (2010). University of California, Irvine, School of Information and Computer Science
5. Grabowski, M., Korpusik, M.: Metrics and similarities in modeling dependencies between continuous-nominal data. Zeszyty Naukowe WWSI. 10/7 (2013)
6. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, Berlin (2001)
7. Indyk, P., Matoušek, J.: Low-distortion embeddings of finite metric spaces. Handbook of Discrete and Computational Geometry. CRC Press LLC, Boca Raton (2004)
8. Krzanowski, W.J.: Principles of Multivariate Analysis: A User's Perspective. Clarendon Press, Oxford (1998)
9. Linial, N.: Finite metric spaces: combinatorics, geometry and algorithms. In: Symposium on Computational Geometry. ACM, Barcelona (2002)
10. Peng, J., Heisterkamp, D., Dai, H.: Adaptative quasiconformal kernel nearest neighbor classification. IEEE Trans. Pattern Anal. Mach. Intell. **26**(5), 656–661 (2004)

http://www.springer.com/978-3-319-26225-3

Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015

Burduk, R.; Jackowski, K.; Kurzynski, M.; Woźniak, M.; Żołnierek, A. (Eds.)

2016, XIV, 855 p. 279 illus., 145 illus. in color., Softcover
ISBN: 978-3-319-26225-3