The Impact of Triangular Inequality Violations on Medoid-Based Clustering

Saaid Baraty* Dan A. Simovici* Catalin Zara*

Abstract

We evaluate the extent to which a dissimilarity space differs from a metric space by introducing the notion of metric point and metricity in a dissimilarity space. The the effect of triangular inequality violations on medoid-based clustering of objects in a dissimilarity space is examined and the notion of rectifier is introduce to transform a dissimilarity space into a metric space.

1 Introduction

Clustering is the process of partitioning sets of objects into mutually exclusive subsets (clusters), so that objects in one cluster are similar to each other in some sense and dissimilar to members of other clusters.

The input data of a clustering technique is the dissimilarity measure between objects. Typically, such dissimilarities are actual metrics defined on the sets of objects; however, often instead of metrics, clustering uses dissimilarities that violate the usual triangular inequality (TI) (see 2). Our objectives in this paper are to analyze the extent to which a non-metric dissimilarity differs from a metric, to analyze the impact that violations of the triangular inequality have on the quality of clusterings, and to introduce the notion of rectifiers as a solution for eliminating TI .

The role of the triangular inequality in designing efficient clustering algorithm has been noted in [1], where it is used to accelerate significantly the k-means algorithm, and in [3], where it is used to improve the efficiency of searching the neighborhood space in the TI-DBSCAN variant of DBSCAN. Another area where violations of the triangular inequality are relevant is the estimation of delays between Internet hosts without direct measurements [4, 5]. These violations, caused by routing policies or path inflation impact the accuracy of Internet coordinate systems.

The role of compliance with the triangular inequality in improving the performance of vector quantization has been observed in [7]. Finally, the triangular inequality plays a fundamental role in the anchors hierarchy, a data structure that allows fast data localization and generates an efficient algorithm for data clustering [6].

*University of Massachusetts Boston, e-mails addresses: {sbaraty,dsim}@cs.umb.edu,czara@math.umb.edu

If a triangle $\{x, y, z\}$ violates the triangular inequality for a dissimilarity d we may have $d(x, y) > d(x, z) + d(z, y)$. Thus, it becomes possible to have two objects x, y that are very similar to a third object z, but very dissimilar among themselves. If x and y are placed in a cluster C whose centroid is z (because of the similarity of x and y with z), the cohesion of C may be seriously impacted by the large dissimilarity between x and y.

We examine the effect of triangular inequality violations and modalities for rectifying these violation in the context of medoid-based algorithms, specifically PAM (Partition Around Medoids), as described in [2]. The algorithm consists of two phases. In the first phase, BUILD, a collection of k objects (where k is the prescribed number of clusters) called *medoids* that are centrally located in clusters are selected. In the second phase, SWAP, the algorithm tries to improve the quality of the clustering by exchanging medoids with non-medoid objects. We selected PAM over k-means, since it works with a dissimilarity matrix without having the actual coordinates of the points. This allows us to freely generate dissimilarities in our experiments without having the actual objects at hand.

In Section 2 we introduce dissimilarity spaces and evaluate the extent a dissimilarity is distinct from a metric by using the number of TI violations of a dissimilarity space. We introduce the notion of rectifier in Section 3 as a device for transforming a dissimilarity into a metric such that the relative order of object dissimilarities is preserved. A specific measure for quantifying the impacts of TI violations of a dissimilarity on clustering is discussed in Section 4. A measure of quality or coherence of clusters which is the topic of Section 5. Finally, we present the results of our experiments in Section 6.

2 Dissimilarity Spaces and Metricity

A *dissimilarity space* is a pair $\mathcal{S} = (S, d)$, where S is a set and $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ is a function such that

(i) $d(x, x) = 0$ for every $x \in S$;

(ii) $d(x, y) = d(y, x)$ for $x, y \in S$.

If $d(x, y) = 0$ implies $x = y$, then we say that d is a *definite dissimilarity*. If $T \subseteq S$, then the pair (T, d_T) is

a *subspace* of (S, d), where d_T is the restriction of d to $T \times T$. To simplify notations we refer to the subspace (T, d_T) simply as T and we denote the restriction d_T by d.

If a dissimilarity satisfies the triangular inequality

$$d(x, y) \leq d(x, z) + d(z, y)$$

for $x, y, z \in S$, then we say that d is a *semi-metric*. A definite semi-metric is said to be a *metric* and the pair (S, d) is referred to as a *metric space*.

Unless stated otherwise all dissimilarity spaces considered here are finite and all dissimilarities are definite. The range of the dissimilarity d of a dissimilarity space $\mathcal{S} = (S, d)$ is the finite set $R(d) = \{d(x, y) \mid x \in S, y \in S\}$. Clustering is often applied to dissimilarity spaces rather than to metric spaces. As mentioned in the introduction, we aim to analyze the impact on the quality of the clustering when using dissimilarities rather than metrics.

Let (S, d) be a dissimilarity space and let $z \in S$. Denote by $M(z)$ the set of pairs

$$M(z) = \{(x, y) \in (S \times S) \mid x, y, z \text{ are pairwise distinct and } d(x, y) \leq d(x, z) + d(z, y)\}.$$

The *metricity of* z is the number $\mu(z) = \frac{|M(z)|}{(|S|-1)(|S|-2)}$. An object z is *metric* if $\mu(z) = 1$ and is *anti-metric* if $\mu(z) = 0$.

There exists dissimilarity spaces without any metric point. Indeed, consider $\mathcal{S}_0 = (\{x_1, \ldots, x_n\}, d)$, where $n \geq 4$ and

$$d(x_i, x_j) = \begin{cases} 0 & \text{if } i = j, \\ 1 & \text{if } |i - j| \in \{1, n-1\}, \\ 3 & \text{otherwise.} \end{cases}$$

Then, every point x_i is non-metric because we have $d(x_{i-1}, x_i) = d(x_i, x_{i+1}) = 1$ and $d(x_{i-1}, x_{i+1}) = 3$ (here $x_{n+1} = x_1$).

On the other hand, we can construct dissimilarity spaces with a prescribed proportion of metric points. Consider the set $S_{pq} = \{x_1, \ldots, x_p, y_1, \ldots, y_q\}$, where $q \geq 2$ and the dissimilarity

$$d_{pq}(u, v) = \begin{cases} 0 & \text{if } u = v, \\ a & \text{if } u \neq v \text{ and } u, v \in \{x_1, \ldots, x_p\} \\ b & \text{if } u \neq v \text{ and } u, v \in \{y_1, \ldots, y_q\} \\ c & \text{if } u \in \{x_1, \ldots, x_p\}, v \in \{y_1, \ldots, y_q\}, \end{cases}$$

where $u, v \in S_{pq}$. If $b > 2c \geq a$, then every x_i is non-metric and every y_k is metric. Indeed, any x_i is non-metric because $b = d_{pq}(y_j, y_h) > d_{pq}(y_j, x_i) + d_{pq}(x_i, y_h) = 2c$ if $j \neq h$. In the same time, every y_k is metric because for any choice of u and v we have $d(u, v) \leq d(u, y_k) + d(y_k, v)$, as it can be easily seen.

A *triangle* in a dissimilarity space (S, d) is a subset T of S such that $|T| = 3$. A triangle $T = \{x_i, x_j, x_k\}$ is said to be *metric* if $d(x_{p_1}, x_{p_2}) \leq d(x_{p_1}, x_{p_3}) + d(x_{p_3}, x_{p_2})$ for every permutation (p_1, p_2, p_3) of the set $\{i, j, k\}$. Thus, a triangle $\{x_i, x_j, x_k\}$ is metric if the subspace $\{x_i, x_j, x_k\}$ is metric. The collection of metric triangles and the collection of set of non-metric triangles of the similarity space $\mathcal{S}$ are denoted by $M(\mathcal{S})$ and $N(\mathcal{S})$, respectively.

Observe that for every triangle T of a dissimilarity space (S, d) there is at most one TI violation. Indeed, suppose that $T = \{x, y, z\}$ is a triangle and there are two violations of the TI involving the elements of T, say

$$\begin{aligned} d(x, y) &> d(x, z) + d(z, y), \\ d(y, z) &> d(y, x) + d(x, z). \end{aligned}$$

Adding these inequalities and taking into account the symmetry of d we obtain $d(x, z) < 0$, which is a contradiction. A non-metric triangle $\{x, y, z\}$ such that $d(x, z) > d(x, y) + d(y, z)$ is denoted by $T_{(y,\{x,z\})}$.

THEOREM 2.1. *Let (S, d) be a dissimilarity space such that $|S| = n$. The number of TI violations has a tight upper bound of $\binom{n}{3}$.*

Proof. For a collection of n distinct points we have $\binom{n}{3}$ distinct triangles and, by the observation that precedes this theorem, there is at most one TI violation associated with each triangle which establishes the upper bound. To prove that the upper bound is tight, we need to show that there exists a dissimilarity d such that the number of TI violations is exactly $\binom{n}{3}$. That is, each distinct triangle has one TI violation. The dissimilarity d is constructed inductively.

Base Construction Step: We have three points x, y and z. We choose $d(x, y), d(x, z)$ and $d(y, z)$ such that there is a TI violation. For example, this can be achieved by defining d such that $d(y, z) = d(x, y) + d(x, z) + 1$.

Inductive Construction Step: Given a collection of n distinct points S with exactly $\binom{n}{3}$ TI violations, we want to add a point $u \notin S$ to S. If we set

$$d(u, x) = d(x, u) = \frac{\min_{(y,z \in S, y \neq z)} d(y, z)}{2 + \epsilon},$$

where $\epsilon > 0$, then, for each newly added triangle $\{u, y, z\}$ we have

$$d(y, z) > d(y, u) + d(u, z).$$

COROLLARY 2.1. *Let (S, d) be a dissimilarity space. Then the average of the metricity of points of S has a tight lower bound of $\frac{2}{3}$.*

3 Rectifiers

We introduce the notion of rectifier as an instrument for modifying non-metric dissimilarities into metrics, with the preservation of the relative order of the dissimilarities between objects.

DEFINITION 3.1. *A* rectifier *is a function $f : \mathbb{R}_{\geq 0} \times U \longrightarrow \mathbb{R}_{\geq 0}$ that satisfies the following conditions:*

(i) $U \subseteq \mathbb{R}_{>0}$ *and* $\inf U = 0$;

(ii) $\lim_{\alpha \to 0+} f(t, \alpha) = y_0$ *for every* $t > 0$, *where* $y_0 > 0$;

(iii) $f(0, \alpha) = 0$ *for every* $\alpha \in U$;

(iv) f *is strictly increasing in its first argument;*

(v) f *is sub-additive in its first argument, that is* $f(t_1 + t_2, \alpha) \leq f(t_1, \alpha) + f(t_2, \alpha)$ *for* $t_1, t_2 \in \mathbb{R}_{\geq 0}$ *and* $\alpha \in U$.

EXAMPLE 3.1. *Let $f(t, \alpha) = t^\alpha$ for $t \in \mathbb{R}_{\geq 0}$ and $\alpha \in (0, 1]$. The function f is a rectifier. Indeed, we have $\lim_{\alpha \to 0+} t^\alpha = 1$ for every $t > 0$ and $f(0, \alpha) = 0$ for every $\alpha \in (0, 1]$.*

For a fixed α the function f is obviously monotonically increasing in its first argument. Furthermore, for any $t_1, t_2 > 0$ the function

$$\varphi(\alpha) = \left(\frac{t_1}{t_1 + t_2}\right)^\alpha + \left(\frac{t_2}{t_1 + t_2}\right)^\alpha$$

is decreasing on $[0, 1]$ and $\varphi(1) = 1$. Therefore,

$$\left(\frac{t_1}{t_1 + t_2}\right)^\alpha + \left(\frac{t_2}{t_1 + t_2}\right)^\alpha \geq 1,$$

which yields the sub-additivity.

EXAMPLE 3.2. *Let $g(t, \alpha) = 1 - e^{-\frac{t}{\alpha}}$ for $t \in \mathbb{R}_{\geq 0}$ and $\alpha \in (0, \infty)$. We claim that g is a rectifier.*

Indeed, we have $\lim_{\alpha \to 0+} g(t, \alpha) = 1$ for every $t > 0$. Also, $g(0, \alpha) = 0$ and g is obviously increasing in t. The sub-additivity of g in its first argument amounts to

$$1 - e^{-\frac{(t_1 + t_2)}{\alpha}} \leq 2 - e^{-\frac{t_1}{\alpha}} - e^{-\frac{t_2}{\alpha}}, \tag{3.1}$$

or equivalently

$$1 - u - v + uv \geq 0,$$

where $u = e^{-\frac{t_1}{\alpha}}$ and $v = e^{-\frac{t_2}{\alpha}}$. In turn, this is equivalent to

$$(1 - u)(1 - v) \geq 0.$$

Since $t \geq 0$, $u, v \leq 1$ which proves the sub-additivity of g.

Note that for a rectifier $f(t, \alpha)$ and a metric d, the function $d_\alpha : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ defined by

$$d_\alpha(x, y) = f(d(x, y), \alpha)$$

is also a metric on S. Indeed, $d(x, y) \leq d(x, z) + d(z, y)$ implies $d_\alpha(x, y) = f(d(x, y), \alpha) \leq f(d(x, z) + d(z, y), \alpha)$ because f is increasing and $f(d(x, z) + d(z, y), \alpha) \leq f(d(x, z), \alpha) + f(d(z, y), \alpha)$ because f is sub-additive. Together, they yield the triangular inequality. However, our interest in the notion of rectifier stems mainly from the following result.

THEOREM 3.1. *Let $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ be a (in)definite dissimilarity and let f be a rectifier. There exists $\delta > 0$ such that the function d_α is a (semi-)metric on S if $\alpha < \delta$. Furthermore, if $d(s_1, s_1') \leq d(s_2, s_2')$, then $d_\alpha(s_1, s_1') \leq d_\alpha(s_2, s_2')$ for $s_1, s_1'.s_2, s_2' \in S$.*

Proof. Note that $d_\alpha(x, x) = f(d(x, x), \alpha) = f(0, \alpha) = 0$ due to the second property of f. Also, it is immediate that $d_\alpha(x, y) = d_\alpha(y, x)$, so we need to show only that there exists δ with the desired properties.

Since $\lim_{\alpha \to 0+} f(t, \alpha) = y_0$ for any t, it follows that for every $\epsilon > 0$ there is $\delta(\epsilon, t) > 0$ such that $\alpha < \delta(\epsilon, t)$ implies $y_0 - \epsilon < f(t, \alpha) < y_0 + \epsilon$ *for every* t. If we choose

$$\delta_0(\epsilon) = \min\{\delta(\epsilon, t) \mid t \in R(d)\},$$

then $\alpha < \delta_0(\epsilon)$ implies

$$d_\alpha(x, y) = f(d(x, y), \alpha) < y_0 + \epsilon$$

and

$$d_\alpha(x, z) + d_\alpha(z, y) = f(d(x, z), \alpha) + f(d(z, y), \alpha) \geq 2y_0 - 2\epsilon.$$

If ϵ is sufficiently small we have $y_0 + \epsilon < 2y_0 - 2\epsilon$, which implies

$$d_\alpha(x, y) \leq d_\alpha(x, z) + d_\alpha(z, y),$$

which concludes the argument for the first part of the statement. The second part follows immediately from .

Theorem 3.1 shows that by using a rectifier we can transform a dissimilarity into a semi-metric. In some instances we can avoid computing $\delta_0(\epsilon)$ using the technique shown in the next example.

EXAMPLE 3.3. *Let $f(t, \alpha) = t^\alpha$ be the rectifier considered in Example 3.1. Suppose that the triple $(u, v, w) \in S^3$ violates the triangular inequality, that is, $d(u, v) > d(u, w) + d(w, v)$.*

Since $d(u, v)^0 \leq d(u, w)^0 + d(w, v)^0$, the set

$$E_{u,v,w} = \{\alpha \in \mathbb{R}_{\geq 0} \mid d(u, v)^\alpha \leq d(u, w)^\alpha + d(w, v)^\alpha\}$$

is non-empty, so $\sup E_{u,v,w} \geq 0$. *If* $\alpha_S = \inf\{\sup E_{u,v,w} \mid u,v,w \in S\} > 0$, *then* d_{α_S} *is a non-trivial semi-metric on* S.

Thus, we need to solve the inequality $1 \leq a^\alpha + b^\alpha$, *where*

$$a = \frac{d(u,w)}{d(u,v)} \text{ and } b = \frac{d(w,v)}{d(u,v)}.$$

Because of the assumption made about (u,v,w) *we have* $a + b < 1$, *so we have* $a, b < 1$.

The solution of this inequality cannot be expressed using elementary functions. However, a lower bound of the set of solution can be obtained as follows.

Let $f : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{\geq 0}$ *be the function defined by* $f(x) = a^x + b^x$. *It is clear that* $f(0) = 2$ *and that* f *is a decreasing function because both* a *and* b *belong to* $[0,1)$. *The tangent to the graph of* f *in* $(0,2)$ *is located under the curve and its equation is*

$$y - 2 = x \ln ab.$$

Therefore, an upper bound of the solution of the equation $1 = a^x + b^x$ *is obtained by intersecting the tangent with* $y = 1$, *which yields*

$$x = -\frac{1}{\ln ab} = \left(\ln \frac{d^2(u,v)}{d(u,w)d(w,v)}\right)^{-1}.$$

Thus, if

$$\alpha \leq \inf\left\{\left(\ln \frac{d^2(u,v)}{d(u,w)d(w,v)}\right)^{-1} \Big| (u,v,w) \in N(\mathcal{S})^3\right\},$$

we can transform d *into semi-metric* d_α.

EXAMPLE 3.4. *It is easy to see that for the dissimilarity space* (S_{pq}, d_{pq}) *introduced in Section 2, the function* $f(t,\alpha) = t^\alpha$ *is a rectifier if and only if* $\alpha \leq \frac{\log 2}{\log \frac{b}{c}}$.

4 Impact of TI Violations on Clustering

We evaluate the impact of using a triangular inequality violating dissimilarity d on clustering. Let (X,d) be a dissimilarity space, where $X = \{x_1, \ldots, x_n\}$. Without loss of generality, we may assume that the range of a dissimilarity has the form

$$R(d) \subseteq \{n \in \mathbb{N} \mid 0 \leq n \leq m\},$$

where $m \in \mathbb{N}$ is the maximum value for dissimilarity d. This is a safe assumption, since we can multiply all the dissimilarities among a finite set of objects by a positive constant without affecting their ratios. Define,

$$\mathbf{AVG}(d) = \sum_{1 \leq i < j \leq n} \frac{2d(x_i,x_j)}{n^2 - n}.$$

Then, if $d(x_i,x_j) \leq \mathbf{AVG}(d)$ we say that x_i, x_j are *almost-similar*, otherwise they are *almost-dissimilar*.

In a non-metric triangle $T_{(x_j,\{x_i,x_k\})}$ the objects x_i, x_k may be similar to x_j but very dissimilar to each other. Clearly, this impacts negatively the quality of the clustering. Yet, the degree of impact differs depending on which of the following cases may occur:

1. If x_i, x_k are almost-similar, that is, $d(x_i,x_k) \leq \mathbf{AVG}(d)$, then $d(x_i,x_j) \leq \mathbf{AVG}(d)$ and $d(x_j,x_k) \leq \mathbf{AVG}(d)$. Thus, all three objects are almost-similar to each other and the clustering algorithm most likely will put all the three objects in one cluster and this will limit the negative impact of this instance of TI violation.

2. If $d(x_i,x_j) > \mathbf{AVG}(d)$ and $d(x_j,x_k) > \mathbf{AVG}(d)$. then $d(x_i,x_k) > \mathbf{AVG}(d)$. No pair of objects are almost-similar and the clustering algorithm will likely place each object in a separate cluster, which cancels the effects of this triangular inequality violation.

3. If $d(x_i,x_k) > \mathbf{AVG}(d)$, $d(x_i,x_j) > \mathbf{AVG}(d)$ and $d(x_j,x_k) \leq \mathbf{AVG}(d)$, then x_i is almost-dissimilar from the two other objects. The clustering algorithm will likely put the two similar objects x_j and x_k in one cluster and x_i in another. This diminishes the negative influence of this triangular inequality violation.

4. The last case occurs when $d(x_i,x_k) > \mathbf{AVG}(d)$, $d(x_i,x_j) \leq \mathbf{AVG}(d)$ and $d(x_j,x_k) \leq \mathbf{AVG}(d)$. In this situation, if the clustering algorithm assigns all three objects to one cluster, we end up with two almost-dissimilar objects x_i and x_k inside a cluster which is not desirable. On the other hand, if the clustering algorithm puts the two dissimilar objects x_i and x_k in two different clusters and x_j in one of the two clusters, for instance in the cluster which contains x_k then, two almost-similar objects x_i and x_j are in two different clusters which is also undesirable. Thus, in this case the impact of triangular violation is substantial.

We penalize the dissimilarity for any triangular inequality violation, but this penalty must be heavier on instances of the last case. Define

$$\theta_{ijk} = \mathbf{AVG}(d) \max\left(\frac{d(x_i,x_k) - \mathbf{AVG}(d)}{d(x_i,x_j) + d(x_j,x_k)}, 0\right).$$

Let $T_{(x_j,\{x_i,x_k\})}$ be a non-metric triangle. If the TI violation falls in to the first category $\theta_{ijk} = 0$. If the violation falls into second and third categories θ_{ijk} will be a positive number. For the last case, θ_{ijk}

will be a larger positive number which exhibits the negative impacts of this violation on clustering. We can normalize θ_{ijk} to make its magnitude consistent across different values of m, the upper bound of dissimilarity, as follows,

$$\begin{aligned}\hat{\theta}_{ijk} &= \frac{2\theta_{ijk}}{\mathbf{AVG}(d)(m-\mathbf{AVG}(d))} \\ &= \max\left(\frac{2\ d(x_i,x_j) - 2\ \mathbf{AVG}(d)}{(d(x_i,x_q)+d(x_j,x_q))(m-\mathbf{AVG}(d))}, 0\right).\end{aligned}$$

The total score for the impact of TI violations of d on clustering is defined as

$$\Phi(X,d) = \sum\{\hat{\theta}_{ijk} \mid T_{(x_j,\{x_i,x_k\})} \in M(X)\}.$$

$\Phi(X,d)$ is normalized by dividing it by the maximum possible number of triangular inequality violations in order to make the measure consistent across different values of n, the number of objects:

$$\hat{\Phi}(X,d) = \frac{6\Phi(X,d)}{n(n-1)(n-2)}.$$

5 A quality measure for clusterings

Let $\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$ be a clustering of the set of objects S and assume that m_i is the medoid of the cluster C_i for $1 \le i \le k$. To assess the quality of the clustering $\mathcal{C}$ we define a measure which we refer as *incoherence degree* of the clustering and is denoted by $\gamma(\mathcal{C})$. This measure is computed from a $k \times k$ matrix $\mathcal{I}$ referred to as the *incoherence matrix* defined as follows:

$$\mathcal{I}_{ij} = \begin{cases} \frac{\max_{v\in C_i}\sum_{u\in C_i} d(u,v)}{\sum_{u\in C_i} d(u,m_i)} & \text{if } i=j \text{ and } |C_j| \ge 1, \\ 1 & \text{if } i=j \text{ and } |C_j| = 1, \\ \frac{|C_j|\cdot d(m_i,m_j)}{\min_{u\in C_i}\sum_{v\in C_j} d(u,v)} & \text{otherwise.} \end{cases}$$

In a "good" clustering, objects within a cluster are similar to each other, and objects that belong to distinct clusters are dissimilar. We construct clusters based on the similarity of objects to medoids. Thus, a clustering is considered as coherent if

1. the average dissimilarity between an object and the other members of the cluster is about the same as the average dissimilarity between the medoid of the cluster and the non-medoid object of the cluster;

2. the sum of dissimilarities of an object u in cluster C_i from all objects $v \in C_j$ is about the same as the product $|C_j| \cdot d(m_i, m_j)$.

The *incoherence degree of a clustering* $\mathcal{C}$ is the average of the maximum diagonal and maximum off-diagonal elements of $\mathcal{I}$. That is,

$$\begin{aligned}\gamma(\mathcal{C}) = &\max_{1\le i\le k} \frac{\max_{v\in C_i}\sum_{u\in C_i} d(u,v)}{2\cdot\sum_{u\in C_i} d(u,m_i)} \\ &+ \max_{\substack{1\le i,j\le k \\ i\ne j}} \frac{|C_j|\cdot d(m_i,m_j)}{2\cdot\min_{u\in C_i}\sum_{v\in C_j} d(u,v)}.\end{aligned}$$

6 Experimental Results

We performed two series of experiments. In the first type of experiments we randomly generated symmetric $n \times n$ dissimilarity matrices with the maximum dissimilarity value m. For such a matrix $\mathcal{M}$ the corresponding dissimilarity is denoted by $d_{\mathcal{M}}$. In the next step, we applied the PAM clustering algorithm to partition the set of n objects into k clusters using $d_{\mathcal{M}}$. We computed the incoherence degree $\gamma(\mathcal{C}_{\mathcal{M}})$ for the resulting clustering $\mathcal{C}_{\mathcal{M}}$ and $\hat{\Phi}(X, d_{\mathcal{M}})$ for dissimilarity $d_{\mathcal{M}}$.

This process was repeated 200 times for randomly generated dissimilarity matrices such that the number $\hat{\Phi}(X, d_{\mathcal{M}})$ lies within a given subinterval. Figures 1, 2, 3, 4, 5 and 6 depict the results of this experiment. The x-coordinate of each point is the $\hat{\Phi}(X, d_{\mathcal{M}})$ average and the y-coordinate is the average incoherence degrees $\gamma(\mathcal{C}_{\mathcal{M}})$ for the clusterings of the form $\mathcal{C}_{\mathcal{M}}$. These figures show a clear ascending trend in the incoherence degree of resultant clusterings (indicating a deterioration of the quality of these clusterings) as the TI violation measure $\hat{\Phi}(X, d_{\mathcal{M}})$ for underlying dissimilarities $d_{\mathcal{M}}$ increases.

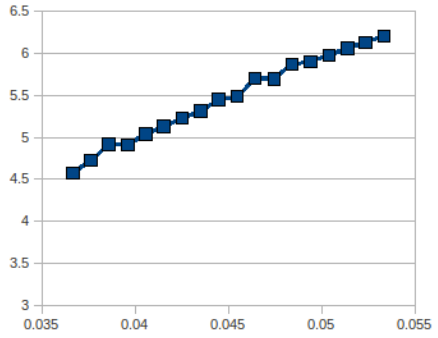

Figure 1: Plot of average $\gamma(\mathcal{C}_{\mathcal{M}})$ to average $\hat{\Phi}(X, d_{\mathcal{M}})$ over 200 randomly generated $\mathcal{M}$ for $k = 5$, $n = 60$ and $m = 50$.

In the second experiment, we used the adjustable dissimilarity d_{pq} described in Section 2, where p specifies the number of non-metric points and $q = n - p$ the number of metric points.

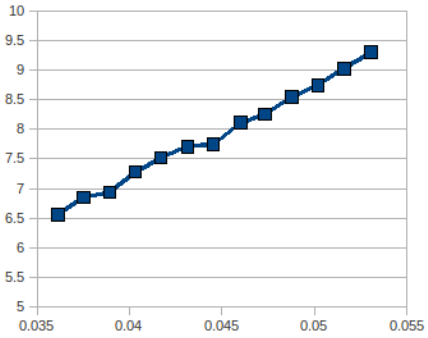

Figure 2: Plot of average $\gamma(\mathcal{C}_{\mathcal{M}})$ to average $\hat{\Phi}(X, d_{\mathcal{M}})$ over 200 randomly generated $\mathcal{M}$ for $k = 7$, $n = 60$ and $m = 50$.

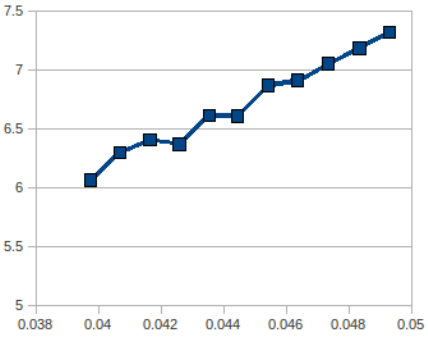

Figure 3: Plot of average $\gamma(\mathcal{C}_{\mathcal{M}})$ to average $\hat{\Phi}(X, d_{\mathcal{M}})$ over 200 randomly generated $\mathcal{M}$ for $k = 7$, $n = 100$ and $m = 50$

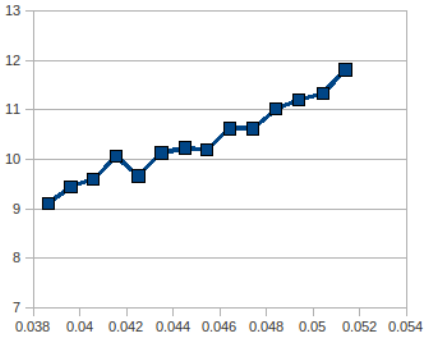

Figure 4: Plot of average $\gamma(\mathcal{C}_{\mathcal{M}})$ to average $\hat{\Phi}(X, d_{\mathcal{M}})$ over 200 randomly generated $\mathcal{M}$ for $k = 9$, $n = 60$ and $m = 50$

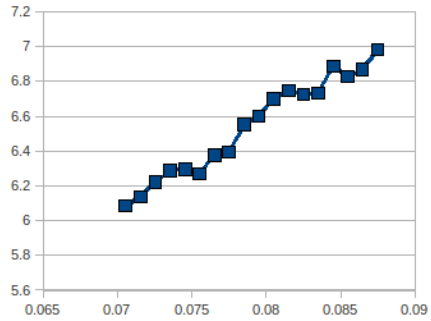

Figure 5: Plot of average $\gamma(\mathcal{C}_{\mathcal{M}})$ to average $\hat{\Phi}(X, d_{\mathcal{M}})$ over 200 randomly generated $\mathcal{M}$ for $k = 7$, $n = 60$ and $m = 25$

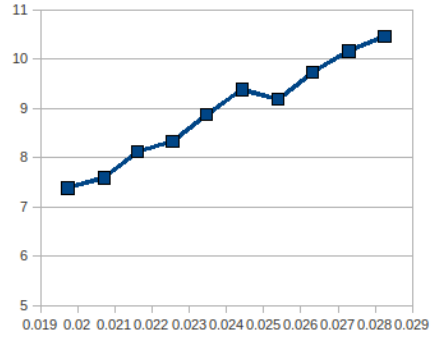

Figure 6: Plot of average $\gamma(\mathcal{C}_{\mathcal{M}})$ to average $\hat{\Phi}(X, d_{\mathcal{M}})$ over 200 randomly generated $\mathcal{M}$ for $k = 7$, $n = 60$ and $m = 100$

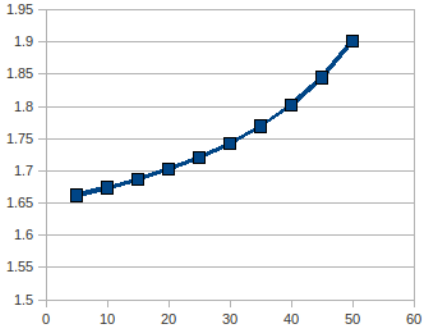

Figure 7: Plot of $\gamma(\mathcal{C}_{pq})$ to p for $n = 60$, $k = 7$.

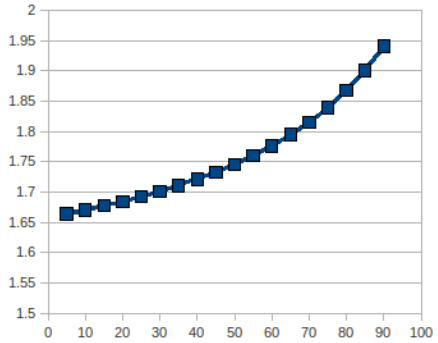

Figure 8: Plot of $\gamma(\mathcal{C}_{pq})$ to p for $n = 100$, $k = 7$.

First, we defined various dissimilarities d_{pq} by setting parameters a, b and c to $1, 7$ and 3 respectively and we assigned p with different values. Then, we rectified the dissimilarities d_{pq} by applying the rectifier described in Example 3.4. We denote the rectified dissimilarities with d^r_{pq}. In the next step we used PAM to generate clusterings based on these two dissimilarities. We denote by $\mathcal{C}_{pq}$ and $\mathcal{C}^r_{pq}$ the clusterings generated based on dissimilarities d_{pq} and d^r_{pq} respectively. Finally, we calculated the incoherence measures $\gamma(\mathcal{C}_{pq})$ and $\gamma(\mathcal{C}^r_{pq})$. Figures 7, 8, 9 and 10 show the increase in the incoherence measure $\gamma(\mathcal{C}_{pq})$ as we increase p. That is the incoherence degree of the clustering is increased as the number of triangular inequality violations of our dissimilarity increases. We repeated the experiment for various parameters. Figures 11, 12, 13 and 14 plot the difference $\gamma(\mathcal{C}_{pq}) - \gamma(\mathcal{C}^r_{pq})$ as p varies. Observe that not only using rectified dissimilarity yields a clustering with better quality according to incoherence measure, but also this improvement in the quality of the clustering due to rectification process increases even when the number of triangular inequality violations of the original dissimilarity increases.

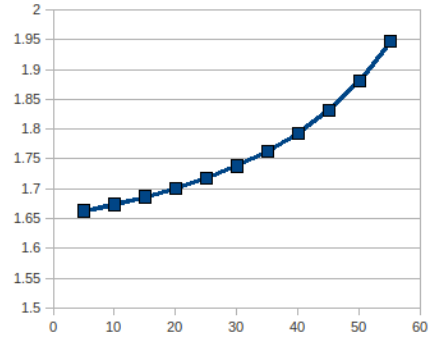

Figure 9: Plot of $\gamma(\mathcal{C}_{pq})$ to p for $n = 60$, $k = 5$.

7 Conclusions and Future Work

We investigated the impact of using non-metric dissimilarities in medoid-based clustering algorithms on the quality of clusterings and demonstrated the impact that TI violations have on clustering quality.

A similar study will be carried out on several variations of the k-means algorithm, which is centroid-based, as well as on density-based clusterings such as DBSCAN. In the later type of algorithms the notion of density is closely tied with the idea of ϵ-neighborhood of an object. Clearly, objects x and z are in the $\max[d(x, y), d(y, z)]$-neighborhood of y and we expect x and z be in $(d(x, y) + d(y, z))$-neighborhood of each

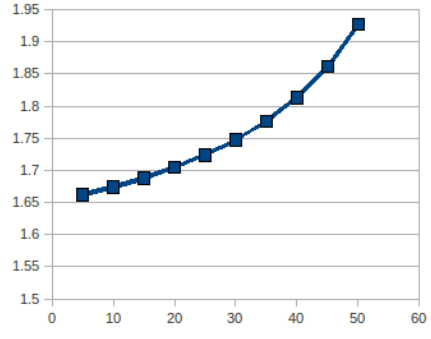

Figure 10: Plot of $\gamma(\mathcal{C}_{pq})$ to p for $n = 60$, $k = 9$.

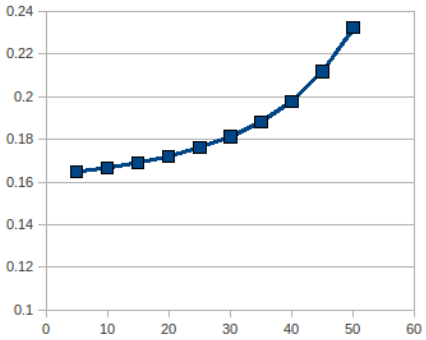

Figure 11: Plot of the difference $\gamma(\mathcal{C}_{pq}) - \gamma(\mathcal{C}^r_{pq})$ to p for $n = 60$, $k = 7$

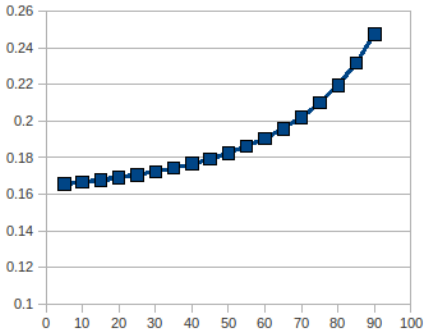

Figure 12: Plot of the difference $\gamma(\mathcal{C}_{pq}) - \gamma(\mathcal{C}^r_{pq})$ to p for $n = 100$, $k = 7$

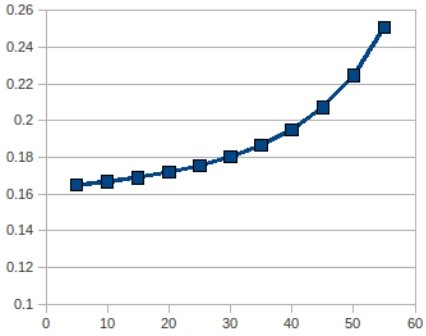

Figure 13: Plot of the difference $\gamma(\mathcal{C}_{pq}) - \gamma(\mathcal{C}^r_{pq})$ to p for $n = 60$, $k = 5$

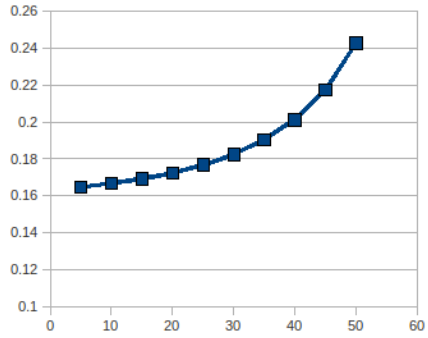

Figure 14: Plot of the difference $\gamma(\mathcal{C}_{pq}) - \gamma(\mathcal{C}^r_{pq})$ to p for $n = 60$, $k = 9$

other, which may not be the case in a TI violating triangle $T_{(y,\{x,z\})}$.

Another direction for extending this work is further analysis of rectifiers involving the relative magnitude deviations of the rectified dissimilarity d_α from the original dissimilarity d.

References

[1] C. Elkan, *Using the Triangle Inequality to Accelerate k-Means*, Proceedings of the 20th International Conference on Machine Learning, ICML-2003, pp. 147–153.

[2] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data – An Introduction to Cluster Analysis*, Wiley Interscience, New York, 1990.

[3] M. Kryszkiewicz and P. Lasek, *TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality*, Rough Sets and Current Trends in Computing, Lecture Notes in Computer Science, 2010, Volume 6086, Springer, pp. 60–69.

[4] Y. Liao, M. A. Kaafar, F. Cantin, B. Gueye, and G.Leduc, *Detecting Triangle Inequality Violations for Internet Coordinate Systems*, Lecture Notes in Computer Science, 2009, Volume 5550, pp. 352–363.

[5] C. Lumezeanu, R. Baden, N. Spring, and B. Bhattacharjee, *Triangle Inequalities and Routing Policy Violations in the Internet*, Passive and Active Network Measurement, Lecture Notes in Computer Science, 2009, Volume 5448, Springer, pp. 45–54.

[6] A. W. Moore, *The Anchors Hierarchy: Using the Triangle Inequality to Survive High Dimensional Data*, Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence, 2000, AAAI Press, pp. 397–405.

[7] J. S. Pan, F. R. McInnes, and M. A. Jack, *fast Clustering Algorithms for Vector Quantization*, Pattern Recognition, Volume 29, 1966, pp. 511–518.

[8] D. A. Simovici and C. Djeraba, *Mathematical Tools*

for Data Mining – Set Theory, Partial Order, Combinatorics, Springer, London, 2008.