New similarity index based on the aggregation of membership functions through OWA operator

Amine AÏT YOUNES, Frédéric BLANCHARD, Michel HERBIN
Université de Reims Champagne Ardenne, France
CReSTIC,
Email: {amine.ait-younes, frederic.blanchard, michel.herbin}@univ-reims.fr

Abstract—In the field of data analysis, the use of metrics is a classical way to assess pairwise similarity. Unfortunately the popular distances are often inoperative because of the noise, the multidimensionality and the heterogeneous nature of data. These drawbacks lead us to propose a similarity index based on fuzzy set theory. Each object of the dataset is described with the vector of its fuzzy attributes. Thanks to aggregation operators, the object is fuzzified by using the fuzzy attributes. Thus each object becomes a fuzzy subset within the dataset. The similarity of a reference object compared to another one is assessed through the membership function of the fuzzified reference object and an aggregation method using OWA operator.

I. INTRODUCTION

Assessing the similarity between samples is a key of success in data analysis process. Many methods rely on similarity indices. Most of clustering ones use pairwise comparisons when aggregating or separating samples [11], [12]. In the framework of case-based reasoning, solving problem needs for searching similar cases and assessing their similarities [10]. Recommender systems also deal with similarity between objects [18]. Thus the search for pairwise similarity indices remains an active field of research [2], [14]. The choice of similarity measures depends on the representation of objects we compare [5] [6]. In this paper, we restrict the scope of this study to the comparison of vector data.

When data is described with multidimensional vectors, the use of metrics remains the classical way to assess pairwise similarity [19], [6]. Unfortunately, database objects could have qualitative features making difficult to obtain standardized quantitative vectors of attributes from the objects. Thus the popular distances (Euclidean distance, Mahalanobis distance, Minkowski metric, Cosine distance, Correlation distance,...) become often inoperative. Moreover noise or vagueness can corrupt data and the curse of dimensionality is also an obstacle in processing queries in high-dimensional space [3]. These drawbacks lead us to propose a similarity index which is not based on a distance function or a metric in the data space.

To overcome these difficulties, the fuzzy set theory gives a framework to design similarity indices [17], [20]. In this paper, the dataset forms the context for our pairwise comparisons between objects.

Each object \( X_i \) of the dataset is fuzzified. Let \( \tilde{X}_i \) be the fuzzy set obtained from the crisp object \( X_i \). \( \tilde{X}_i \) is the fuzzy version of \( X_i \). The membership degree of the crisp object \( X_j \) to the fuzzy set \( \tilde{X}_i \) is considered as the similarity value from \( X_j \) to the reference object \( X_i \). Therefore the similarity indices we propose are only fuzzy membership functions. Note that such similarity indices based on membership functions do not necessarily define symmetric relations.

The challenge using our approach becomes to obtain a fuzzification of an object \( X_i \) within the dataset [13]. To achieve this goal, the attributes of the object \( X_i \) are considered as fuzzy numbers or fuzzy quantities. Then each crisp object \( X_j \) is described with a vector of membership degrees relative to the fuzzy attributes of \( X_i \). Thanks to fuzzy logic operators, we aggregate these membership degrees to obtain the aggregated membership value of \( X_j \) to the fuzzy set \( \tilde{X}_i \). The critical issue of this approach is the aggregation method we use. This communication proposes to adapt OWA operators [16] to define our aggregation method.

The paper is organized as follows.

In the sections 2 and 3 we present our approach for the fuzzification of the attributes. The section 4 exposes the methodology we use to evaluate the sensibility and specificity of each attributes. After the fuzzification of the data, we present in the section 5 the aggregation procedure used to build a new similarity index using an Ordered Weighted Aggregation operator. Before concluding, we present in the section 6 a comparison of our new pairwise similarity indice with the popular metrics.

II. DOMAIN OF FUZZIFICATION

Let \( E \) be a set of \( n \) objects defined by:

\[
E = \{ X_i / 1 \leq i \leq n \}
\]

(1)

where \( X_i \) are the \( n \) objects of \( E \).

Each object is described by a vector of \( p \) attributes. The object \( X \) is represented by the \( p \)-tuple \( (x_{ik}) \) with \( 1 \leq k \leq p \) where \( x_{ik} \) is the value of the \( k \)-th attribute of the object \( X_i \). These \( p \) attributes are either quantitative or qualitative. If the \( k \)-th attribute is quantitative, then its values lie within an interval \([a_k, b_k]\). If the \( k \)-th attribute is qualitative, then its values are within a set \( \{v_1, v_2, v_3, \ldots, v_l\} \) of \( l \) values. In both cases, we call \( D_k \) the set we use to define the \( k \)-th attribute.

The domain of definition \( D \) of \( E \) is defined by:
\[ D = \prod_{1 \leq k \leq p} D_k \] (2)

Then we have: \( E \subseteq D \) with: \( \# E = n \). In the following each object becomes a fuzzy subset of \( E \) relatively to its attributes. Thus \( E \) is called domain of fuzzification.

III. FUZZIFICATION OF THE ATTRIBUTES

This section is devoted to the fuzzification of an object \( X_i \) within the dataset \( E \). Although the fuzzification of an attribute is itself beyond the scope of this document, we firstly describe the way we used to fuzzify each attribute value of the object \( X_i \). Thus we obtain \( k \) fuzzy attributes for \( X_i \). Then we merge these fuzzy attributes to build the fuzzy object \( \tilde{X}_i \) defined in \( E \).

Let \( X_i \) be an arbitrary reference object of the data set \( E \). Let \( x_{ik} \) be the value of the \( k \)-th attribute of \( X_i \). The values of attributes are often imprecise and the meaning could be vague. Therefore it is convenient to represent such imprecise or vague values by fuzzy sets. Thus \( x_{ik} \) is represented by a fuzzy subset of \( D_k \). The membership function \( m^i_k \) of this fuzzy subset is defined by:

\[ m^i_k: \quad D_k \rightarrow [0,1] \]
\[ x \rightarrow m^i_k(x) \] (3)

In this paper, these fuzzy sets are normalized with \( m^i_k(x_{ik}) \leq 1 \).

In this paper, we propose a simple and empirical approach of the data fuzzification. Each numeric value is represented by a conventional trapezoidal membership function defined by \((a, b, c, d)\) with (cf. fig. 1):

\[ m^i_k(x) = \begin{cases} 
0 & \text{if } x < a_i \\
\frac{x-a_i}{b_i-a_i} & \text{if } a_i \leq x < b_i \\
1 & \text{if } b_i \leq x < c_i \\
\frac{d_i-x}{d_i-c_i} & \text{if } c_i \leq x < d_i \\
0 & \text{if } d_i \leq x 
\end{cases} \] (4)

Let \( \overline{x}_k \) and \( \sigma_k \) be respectively the mean and the standard deviation of the \( k \)-th attribute within \( D_k \). If \( \text{dev}_k^i \) is the deviation between \( x_{ik} \) and \( \overline{x}_k \) (i.e. \( \text{dev}_k^i = |x_{ik} - \overline{x}_k| \)), an empirical study leads us to propose:

\[ \begin{aligned}
a_i &= x_{ik} - \sigma_k - 0.5 \text{dev}_k^i \\
b_i &= x_{ik} - 0.5 \sigma_k - 0.1 \text{dev}_k^i \\
c_i &= x_{ik} + 0.5 \sigma_k + 0.1 \text{dev}_k^i \\
d_i &= x_{ik} + \sigma_k + 0.5 \text{dev}_k^i 
\end{aligned} \] (5)

If the \( k \)-th attribute is qualitative, \( x_{ik} \) is fuzzified using a degree of membership for each possible value of the attribute. Then \( m^i_k \) is defined by \( l \) values \((m^i_k(v_1), m^i_k(v_2), ..., m^i_k(v_l))\) within \( D_k \).

This paper proposes to use \( m^i_k \) to fuzzify the object \( X_i \) within \( E \) in respect with its \( k \)-th attribute. The membership function of the object \( X_i \) is defined by:

\[ \mu^i_k: \quad E \rightarrow [0,1] \]
\[ X_j \rightarrow \mu^i_k(X_j) = m^i_k(x_{jk}) = \mu^i_{jk} \] (6)

with \( 1 \leq j \leq n \) and \( 1 \leq k \leq p \).

If the value of \( x_{ik} \) is not set, then we propose to define \( \mu^i_k \) by simply \( \mu^i_{jk} \), i.e. \( \frac{1}{2} \) in order to ensure the robustness of the proposed approach.

At this stage of the communication, several points should be noted. Each object \( X_i \) gives rise to \( p \) fuzzy subsets of \( E \). Each subset is associated with an attribute. These fuzzy subsets are defined with reference to the object \( X_i \). They are normalized because \( \mu^i_{jk} \leq 1 \).

We propose to consider the membership degrees \( \mu^i_{jk} \) (with \( X_j \in E \)) as similarity values from \( X_j \) to the reference \( X_i \) with respect to the \( k \)-th attribute. If \( \mu^i_{jk} = 1 \), then \( X_i \) and \( X_j \) are considered as similar with respect to the \( k \)-th attribute. In contrast, if \( \mu^i_{jk} = 0 \), then \( X_i \) and \( X_j \) are considered as dissimilar with respect to the attribute. The more \( \mu^i_{jk} \) is close to 1, the larger the similarity from \( X_j \) to \( X_i \) for the \( k \)-th attribute. Thus the membership function \( \mu_{ik} \) (with \( 1 \leq k \leq p \)) is considered as a similarity index to \( X_i \) with respect to its attribute \( x_{ik} \).

We can see in "fig. 1" that \( x_{jk} \) is considered as similar to \( x_{ik} \) but \( x_{jk} \) is not comparable to \( x_{ik} \). This similarity value is asymmetric. In "fig. 2" \( x_{jk} \) is not comparable to \( x_{ik} \) : \( \mu^i_k(x_{jk}) = 0 \) but \( \mu^i_{jk}(x_{ik}) > 0 \).

The membership functions \( \mu_{ik} \) give \( p \) indices of similarity to \( X_i \) within the set \( E \). Let us define two characteristics of these indices that we call the sensibility and the specificity to the similarity with \( X \).

Let \( \text{sens}_{ik} \) be the mean of \( \mu^i_{jk} \) when \( X_j \in E \):

\[ \text{sens}_{ik} = \frac{1}{n} \sum_{X_j \in E} \mu^i_{jk} \] (7)

The value \( \text{sens}_{ik} \) lies between 0 and 1. It assesses an average similarity between the reference object \( X_i \) and the whole dataset \( E \) in respect with the \( k \)-th attribute.

If \( \text{sens}_{ik} \) is close to 1, then the \( n \) similarity values \( \mu^i_{jk} \) are also rather close to 1. Then the \( n \) objects \( X_j \) of \( E \) are rather similar to \( X_i \). In this case, the values \( \mu^i_{jk} \) are sensitive indicators of the similarity to \( X_i \). Since these values are rather equal to 1 or close to 1, then the value \( \mu^i_{jk} \) becomes highly symptomatic of a dissimilarity (non-similarity) with \( X_i \) when the indicator of similarity \( \mu^i_{jk} \) is close to 0. Thus the \( k \)-th attribute is considered as an attribute sensitive to the similarity with \( X_i \).
In contrast, if \( \text{sens}_{ik} \) is close to 0, the \( n \) similarity values \( \mu_{ijk} \) are also rather close to 0. Then the \( n \) objects \( X_j \) are rather dissimilar (non-similar) to \( X_i \). In this case, the values \( \mu_{ijk} \) are specific indicators of the similarity to \( X_i \). Since these values are rather equal to 0 or close to 0, then the value \( \mu_{ijk} \) becomes highly symptomatic of a similarity to \( X_i \). Thus the \( k \)-th attribute is considered as an attribute specific of the similarity with \( X_i \).

When we consider the \( k \)-th attribute, \( \text{sens}_{ik} \) is a coefficient of the sensibility of this attribute to the similarity with \( X_i \) and \( 1-\text{sens}_{ik} \) is a coefficient of the specificity of the attribute for the similarity with \( X_i \). These two coefficients characterize the \( k \)-th attribute with reference to the object \( X_i \) within \( E \).

Let us consider an example (see Table I) to explain these coefficients. The dataset \( E \) has six objects \( X_1, X_2, X_3, X_4, X_5 \) and \( X_6 \). Each object is described with four attributes. The reference object is \( X_1 \). The four attribute values of \( X_1 \) are fuzzified. The four membership functions \( \mu_{1k}, \mu_{2k}, \mu_{3k} \) and \( \mu_{4k} \) indicate the degrees of similarity to \( X_1 \). In this example, the sensitivities of the four attributes are respectively 0.767, 0.400, 0.583 and 0.680. The 1st attribute is the most sensitive one. Only \( X_6 \) has 1st attribute value dissimilar from the one of \( X_1 \). Thus the 1st attribute reveals the dissimilarity with \( X_1 \). The specificities of the four attributes are respectively 0.233, 0.600, 0.417 and 0.320. The 2nd attribute is the most specific one. Only \( X_2 \) has 2nd attribute value similar to the one of \( X_1 \). Thus the 2nd attribute reveals the similarity with \( X_1 \).

<table>
<thead>
<tr>
<th>Objects</th>
<th>( \mu_1 )</th>
<th>( \mu_2 )</th>
<th>( \mu_3 )</th>
<th>( \mu_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>0.9</td>
<td>1</td>
<td>0.3</td>
<td>0.9</td>
</tr>
<tr>
<td>( X_3 )</td>
<td>0.9</td>
<td>0.1</td>
<td>0.4</td>
<td>0.7</td>
</tr>
<tr>
<td>( X_4 )</td>
<td>0.9</td>
<td>0.1</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>( X_5 )</td>
<td>0.9</td>
<td>0.1</td>
<td>0.6</td>
<td>0.3</td>
</tr>
<tr>
<td>( X_6 )</td>
<td>0</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**TABLE I**

Sensitivity and Specificity of the Attributes in Respect with the Reference Object \( X_1 \): Example of 6 Objects with 4 Fuzzy Attributes. \( \mu_k \) are the Degrees of Membership to \( X_1 \) Relative to the \( k \)-th Attribute with \( 1 \leq k \leq 4 \)

IV. AGGREGATION WITH OWA OPERATORS

Let us consider the reference object \( X_i \) in \( E \). The fuzzy subset defined by the membership function \( \mu_{ik} \) depends on the value \( x_{ik} \) of the \( k \)-th attribute of \( X_i \). We propose to aggregate these \( p \) fuzzy subsets taking into account all the attributes. The goal is to fuzzify the reference object \( X_i \) within \( E \) defining a new membership function \( \mu^i \) fusing the functions \( \mu_{ik} \).
The aggregation operators give a classical way to merge the fuzzy subsets in $E$. Let $aggref$ be an aggregation operator. The function $\mu^i$ is defined by:

$$
\mu^i : E \rightarrow \mu^i(X_j) = \frac{1}{w_k}\sum_{1 \leq k \leq p} w_k \times \mu^i_k(X_j)
$$

(9)

The aggregation operators are well studied in literature [7], [8]. The minimum is the reference operator to obtain a conjunction and the maximum is the one for a disjunction. The operators used in this paper are a tradeoff between the conjunction (AND) and the disjunction (OR).

If the similarity index $\mu^i_k$ is very sensitive ($sens_{i,k}$ close to 1), then the similarity index $\mu^i_k$ should contribute to $\mu^i$ using a conjunction operator. Indeed, a conjunction seems desirable because significant information is obtained when $\mu^i_k$ is close to 0. In contrast, if the similarity index $\mu^i_k$ is very specific ($sens_{i,k}$ close to 0), a disjunction operator seems preferable because significant information is obtained when $\mu^i_k$ is close to 1.

In this paper, the tradeoff between conjunction and disjunction is defined using an Ordered Weighted Aggregation (OWA operators proposed by R. Yager [16]).

Let us describe the function $\mu^i$ obtained when using such an OWA operator.

For an object $X_j$ in $E$, the membership degrees $\mu^i_k(X_j)$ are ordered by decreasing order. We obtain:

$$
\mu^i(1)(X_j) \geq \mu^i(2)(X_j) \geq \mu^i(3)(X_j) \geq \ldots \geq \mu^i(p)(X_j)
$$

The aggregation is defined by:

$$
\mu^i(X_j) = \frac{1}{w_k}\sum_{1 \leq k \leq p} w_k \times \mu^i_k(X_j)
$$

(9)

We denote $W$ the weighting vector:

$$
W = [w_1, w_2, \ldots, w_p]
$$

(10)

with $\sum_{1 \leq k \leq p}(w_k) = 1$ and $w_k \in [0, 1]$

The weights are not associated with attributes but with their ordered positions. The challenge is to determine the weights.

The conjunction operator (i.e. the minimum) is obtained if:

$$
W_* = [0, 0, \ldots, 1]
$$

(11)

The disjunction operator (i.e. the maximum) is obtained if:

$$
W^* = [1, 0, \ldots, 0]
$$

(12)

The ordinary average is recovered if:

$$
\bar{W} = \left[\frac{1}{p}, \frac{1}{p}, \ldots, \frac{1}{p}\right]
$$

(13)

Pérez and Lamata [15] discuss the weights determination by means of linear functions that we use in this paper.

To emphasize the conjunction, we propose to use decreasing weights $w_{\min}$ defined by using the linear orders of Borda [4] where:

$$
w_{\min}(k) = \frac{2k}{p(p+1)}
$$

(14)

where $1 \leq k \leq p$.

To emphasize the disjunction, we propose to use increasing weights $w_{\max}$ defined by increasing linear orders with:

$$
w_{\max}(k) = \frac{2(p+1-k)}{p(p+1)}
$$

(15)

where $1 \leq k \leq p$.

Then we have:

$$
w_{\min}(k) + w_{\max}(k) = \frac{2}{p}
$$

(16)

For example, if $p = 4$

$$
W_{\min} = [\frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20}]
$$

$$
W_{\max} = [\frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20}]
$$

The membership function $\mu^i_k$ is a similarity index with $X_i$ in respect with the $k$-th attribute. $sens_{i,k}$ describes the sensitivity of the index. In this paper, $sens_{i,k}$ is considered as the ANDness of the attribute and $1-sens_{i,k}$ defines the specificity i.e. the ORness of the attribute. Then the weights of OWA operator we use are defined by:

$$
w_k = C \left( (sens_{i,k})w_{\min}(k) + (1-sens_{i,k})w_{\max}(k) \right)
$$

(18)

where $C$ is the coefficient for obtaining $\sum_{1 \leq k \leq p}(w_k) = 1$.

The membership degrees $\mu^i(X_j)$ with $X_j \in E$ are obtained through these weights. Such an OWA operator in $E$ permits us to define a similarity index $sim_i$ from the reference $X_i$ to the other objects $X_j$ of $E$ by:

$$
sim(X_i, X_j) = \mu^i(X_j)
$$

(19)

Note that the similarity index we propose is not necessarily symmetrical (cf. fig: fuzzy2). In fact $sim(X_i, X_j)$ is not always equal to $sim(X_j, X_i)$.

V. COMPARISON WITH POPULAR METRICS

The way we describe to design pairwise similarity between multidimensional data leads us to propose a new pairwise similarity index based on fuzzy logic operators. This section is devoted to the assessment of the new similarity index we propose. First we define a criterion to compare the similarity indices. Second we apply this criterion for comparing our new
pairwise similarity index with more classical indices based on the popular metrics.

A. Assessment of a similarity index

Such a pairwise similarity indices are often used in the context of data clustering [9]. In this paper, we take the problem upside down using the clusters for assessing the similarity indices. The clusters define a partition of $E$, we call $C_X$, the cluster to which the object $X$ belongs and we call $\text{sim}$ a similarity index between two objects of $E$.

Let us consider all pairs of objects $(X_i, X_j)$ within the sample data $E$. If the two objects $X_i$ and $X_j$ belong to the same cluster, then an optimal similarity index from $X_i$ to $X_j$ should be equal to one (i.e. $X_i$ and $X_j$ are similar). On the contrary, if the two objects $X_i$ and $X_j$ belong to two different clusters, then an optimal similarity index from $X_i$ to $X_j$ should be equal to zero (i.e. $X_i$ and $X_j$ are dissimilar). Thus we define the intra-cluster similarity of $\text{sim}$ with:

$$\text{intra(sim)} = \frac{1}{n_1} \sum_{C_X,=C_X_i} \text{sim}(X_i, X_j) \quad (20)$$

where $n_1$ is the number of couples $(X, Y)$ where $X$ and $Y$ belong to the same cluster. The inter-cluster similarity is defined with:

$$\text{inter(sim)} = \frac{1}{n_2} \sum_{C_X,\neq C_X_j} \text{sim}(X_i, X_j) \quad (21)$$

where $n_2$ is the number of couples $(X_i, X_j)$ where $X_i$ and $X_j$ belong to two different clusters.

The similarity index $\text{sim}$ is optimal for the clusters when $\text{intra(sim)} = 1$ and $\text{inter(sim)} = 0$. Therefore we define a criterion to evaluate $\text{sim}$ with:

$$\text{crit(sim)} = \text{intra(sim)} - \text{inter(sim)} \quad (22)$$

The value $\text{crit(sim)}$ lies always between -1 and 1. The higher $\text{crit(sim)}$, the more optimal $\text{sim}$ with respect to the clusters.

We propose to use this criterion to assess our new pairwise similarity index.

B. Applications

This paper proposes a new way to evaluate the similarity between multidimensional vector data. The most classical way consists in using the popular metrics when data are quantitative. In this paper we consider Euclidean distance, Manhattan distance, Chebyshev distance, Canberra distance and Mahalanobis distance (see Table II). In fact, these distances are dissimilarity indices that we transform into similarity indices with:

$$\text{simil}(X, Y) = 1 - \frac{\text{dist}(X, Y)}{\max_{A, B, E} \text{dist}(A, B)} \quad (23)$$

where $\text{dist}$ is the distance that we use.

Then we have five similarity indices we call Euclidean, Manhattan, Chebyshev, Canberra, and Mahalanobis which are based on their three respective popular metrics.

We compare these similarity indices based on distances with two indices we propose based on the aggregation operators of membership functions. The first one is called $\text{simOWA}$ that is the index which is described in this paper. It is based on OWA operators. The second one replaces the OWA operator with the arithmetic mean of the membership functions. This second one is called Arithmetic.

The similarity indices are computed using the databases from Machine Learning Repository of UCI [1].

In this paper we propose to use six numerical multivariate clustering databases that are iris, wine, ecoli, glass, seeds and haberman. The number of attributes lies between 3 and 15. The number of objects lies between 100 and 500. The number of clusters lies between 3 and 10. iris is the classical database that has 150 iris plants with 4 attributes and three clusters. The wine recognition database has 178 objects with 13 attributes and three clusters. ecoli is the database of sites of protein localization, it has 336 objects with 7 attributes and eight clusters. The glass identification database has 214 objects with 9 attributes and seven clusters. The seeds database of wheat varieties has 210 objects with 7 attributes and three clusters. Haberman’s survival database has 306 objects with 3 attributes and two clusters.

The results obtained are in Table III. We can see that the similarity indices proposed in this paper (SimOWA and Arithmetic) are better than the others in 5 cases and ranked second for one of them (glass Database). In 5 cases of 6, the similarity indice based on the OWA operator gives us better results than the one based on the arithmetic mean.

VI. CONCLUSION

The approach we propose has a significant advantage, it allows us to deal with imperfection that is a general case with real data. In medicine and biology, data is often imprecise mainly due to the inherent variability of biological data. In physics, data is also imperfect and it is usual to assign a value from a sensor with the accuracy of the measurement. Qualitative data is also imprecise or vague. Thus the use of the fuzzy set theory is relevant in this context of imperfect data.

In this paper we propose a simple method of fuzzification for imperfect multidimensional data. With this fuzzification we define a new similarity indice that will allow us in future works to identify the main features of the dataset and build a robust classification. We can also use this approach to compare a new object with the existing data, for example by finding the nearest objects.

REFERENCES


### Definition of the Most Popular Metrics Between Quantitative Data

<table>
<thead>
<tr>
<th>Popular metrics</th>
<th>Distances Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>$\text{dist}(X,Y) = \sqrt{\sum_{1 \leq k \leq p} (x_k - y_k)^2}$</td>
</tr>
<tr>
<td>Manhattan (city block)</td>
<td>$\text{dist}(X,Y) = \sum_{1 \leq k \leq p}</td>
</tr>
<tr>
<td>Chebyshev</td>
<td>$\text{dist}(X,Y) = \max_{1 \leq k \leq p}</td>
</tr>
<tr>
<td>Canberra</td>
<td>$\text{dist}(X,Y) = \sum_{1 \leq k \leq p} \frac{</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>$\text{dist}(X,Y) = \sqrt{(X-Y)^T C^{-1} (X-Y)}$</td>
</tr>
</tbody>
</table>

### Comparison of Similarity Indices

<table>
<thead>
<tr>
<th>Database</th>
<th>iris</th>
<th>uscme</th>
<th>ecoli</th>
<th>glass</th>
<th>seeds</th>
<th>haberman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of objects</td>
<td>130</td>
<td>150</td>
<td>356</td>
<td>214</td>
<td>306</td>
<td>136</td>
</tr>
<tr>
<td>Number of attributes</td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>9</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Number of clusters</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Index of similarity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Euclidean</td>
<td>0.336</td>
<td>0.175</td>
<td>0.230</td>
<td>0.098</td>
<td>0.275</td>
<td>0.020</td>
</tr>
<tr>
<td>Manhattan</td>
<td>0.331</td>
<td>0.176</td>
<td>0.210</td>
<td>0.097</td>
<td>0.272</td>
<td>0.026</td>
</tr>
<tr>
<td>Chebyshev</td>
<td>0.344</td>
<td>0.175</td>
<td>0.211</td>
<td>0.085</td>
<td>0.258</td>
<td>0.017</td>
</tr>
<tr>
<td>Canberra</td>
<td>0.422</td>
<td>0.222</td>
<td>0.142</td>
<td>0.166</td>
<td>0.238</td>
<td>0.048</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>0.113</td>
<td>0.047</td>
<td>0.078</td>
<td>0.064</td>
<td>0.080</td>
<td>0.027</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>0.506</td>
<td>0.264</td>
<td>0.245</td>
<td>0.142</td>
<td>0.447</td>
<td>0.052</td>
</tr>
<tr>
<td>SimOWA</td>
<td>0.510</td>
<td>0.270</td>
<td>0.289</td>
<td>0.151</td>
<td>0.451</td>
<td>0.044</td>
</tr>
</tbody>
</table>


