Transactions on Machine
Learning and Data Mining
Vol. 14, No. 1 (2021) 31 - 48

P-ISSN: 1865-6781, E-ISSN 2509-9337
ISBN: 978-3-942952-86-6

ibai Publishing
www.ibai-publishing.org

Comparison of Methods for Computing Similarity Based on Clusters -Utilizing Different Membership Functions

Arthur Yosef[1], Eli Shnaider[2] and Moti Schneider[3]

[1] Tel Aviv-Yaffo Academic College, Israel, yusupoa@yahoo.com
[2]Independent, eli-sh@012.net.il
[3]Netanya Academic College, Israel, profmoti@gmail.com

Abstract. In this paper, we compare various methods for computing similarity between numerical vectors based on their division into clusters. The advantage of utilizing clusters is apparent mostly in the cases where the data are very unreliable and distorted, so that the cluster represents approximate value of its elements in a very broad term. Measuring similarity between numerical vectors following their division into clusters provides additional method for similarity measurement, which might be a preferable method when lack of confidence in the measurements of individual data elements is high. In addition, we compare the influence of applying the various types of membership functions on the results of similarity measurements.

Keywords: Data Mining, Soft Computing, fuzzy logic, similarity measure, clusters.

1 Introduction

When computing similarity between two numerical vectors, the general practice is to perform a measurement of distance between the corresponding elements of these vectors. However, when the data are characterized by severely unreliable and distorted measurements, utilizing individual data elements means, that we are implicitly assuming that the measurements are sufficiently precise so that the results are still expected to be reliable. This is often self-illusion in the cases of severe distortion (or intentional disinformation). However, there are numerous cases of problematic data, where the distortions are of limited magnitude for most of the data elements, such that when domain expert goes over the numbers, the vast majority of them seem to be in the right order of magnitude. In cases like that, it is possible to approach the computation of similarity by dividing the numerical vectors into clusters and treating the whole cluster, where a given data element is located as an approximate measure representing its value.

Obviously, such approach takes into account our lack of confidence in the precision of the individual data elements. Using the values of the whole cluster as a substitute for individual elements contained in it, allows more reasonable evaluation of similarity of vectors. The broader approach compensates for the imprecision and data distortions as long as they are not unreasonable. However, when interpreting the results, the broad nature of treating the measurements must be kept in mind.

There are many different approaches to perform clustering. The literature is very extensive, and only a small sample is presented here. There are algorithms such as K-means [4] and Clustering Large Applications based on Randomized Search [5], which are based on a partitioning approach; Gaussian mixture models ([6],[7]) are associated with a model-based approach; Divisive Analysis [8] and Balanced Iterative Reducing and Clustering using Hierarchies [9] are based on a hierarchical approach; Statistical Information Grid [14] and Clustering in Quest [10] are based on a grid-based approach. Density-Based Spatial Clustering of Applications with Noise [11] and Ordering Points to Identify the Clustering Structure [12] are examples of a density-based approach. Density-based clustering creates clusters of arbitrary shape, is robust to noise, and does not require prior knowledge regarding the number of clusters [13]. In our study we decided to utilize K-means Clustering method as elaborated in the later sections.

In various studies utilizing Fuzzy Logic, different membership functions are applied. There are studies specifically dealing with comparisons of such functions. For example, Omar et al. [18] compare the impact of various membership functions on the performance of fuzzy controller. They apply three most used types of Membership Functions: Triangular, Trapezoidal (which are linear membership functions) and the Gaussian membership function (Non- linear). Zhao and Bose [19] evaluate the impact of various membership functions on the performance of fuzzy logic-based induction motor drive. They consider Triangular, Trapezoidal, Gaussian, Bell, Sigmoidal and Polynomial membership functions. In our study we compare results generated by Linear Membership Function to the results of the Sigmoid Membership Function.

Attig and Perner [21] emphasize the need for normalizing data before computing similarity to bring the relevant numerical vectors into the same scale in the context of Case Based Reasoning. In particular, the study addresses the process of determining the lower and upper bounds and the problems that arise when these values are not correctly estimated.

2 Data preparation and similarity measures

2.1 Data preparation

The first step to utilize cluster similarity method requires specific data preparation to make the numerical vectors comparable. Each numerical vector is normalized to bring various numerical vectors into the same scale.

The normalization process is as follows: Assume we have a vector

$$X_{raw} = (x_1^{raw}, x_2^{raw}, \dots, x_n^{raw})$$

containing n values (before being normalized).

There are several possible mathematical functions for performing the normalization stage. We selected the following two membership functions for comparison, Equations (1) and (2):

- Linear membership function (LMF): Assume that we were given 2 values, $\min_{cut}$ and $\max_{cut}$ such that

$$x_k = \begin{cases} 0 & , x_k^{raw} < \min_{cut} \\ \frac{x_k^{raw} - \min_{cut}}{\max_{cut} - \min_{cut}} & , \min_{cut} \le x_k^{raw} \le \max_{cut} \\ 1 & , x_k^{raw} > \max_{cut} \end{cases} \quad (1)$$

  where raw means data before being normalized. In our examples, $\max_{cut}$ represents a cut-off point above which every data element is a full member in the fuzzy set, and $\min_{cut}$ represents a cut-off point below which every data element has a zero membership in the fuzzy set.

- Sigmoid membership function (SMF):

$$x_k = \begin{cases} 0 & , x_k^{raw} \le \min_{cut} \\ 2\left(\frac{x_k^{raw} - \min_{cut}}{\max_{cut} - \min_{cut}}\right)^2 & , \min_{cut} < x_k^{raw} \le \alpha \\ 1 - 2\left(\frac{x_k^{raw} - \min_{cut}}{\max_{cut} - \min_{cut}}\right)^2 & , \alpha < x_k^{raw} < \max_{cut} \\ 1 & , x_k^{raw} \ge \max_{cut} \end{cases} \quad (2)$$

$$where\ \alpha = \frac{\min_{cut} + \max_{cut}}{2}$$

For computing similarity, there are two types of broad relations between two numerical vectors: direct relations or inverse relations.

- Direct relation: In general, high values (large numbers) elements of one numerical vector are associated with high values of the corresponding elements in second numerical vector; and there is a similar general correspondence between the data elements having low values in both vectors.

- Inverse relation: In general, high values (large numbers) elements of one numerical vector are associated with low values of the corresponding elements in second numerical vector; and the low value elements of first numerical vector are associated with high values of the corresponding elements of second numerical vector.

In order to determine, whether the relation is direct or inverse, we perform the following steps:

1. Sort one numerical vector by the value of the data elements.
2. Select a group "A" of elements having high values (relatively large numbers for that vector)
3. Select a group "B" of elements having low values (relatively small numbers for that vector)
4. Find corresponding groups ("A" and "B") of data elements in the second numerical vector.
5. Compute median values for "A" and "B" for the second numerical vector.
6. If Median of "A" > Median of "B" then the relation is direct. If Median of "A" < Median of "B" then the relation is inverse.

If the relation is direct, then normalization equations as presented above can be utilized. If the relation is inverse, then the normalization process is inversed as shown in Equation 3.

$$x_k^{inverse} = 1 - x_k \tag{3}$$

As an example, let us look at a pair of variables: GDP and EXP. The data are presented in Figure 1a. It is easy to see, at this point, that the numbers of these vectors are not in the same scale and therefore not comparable.

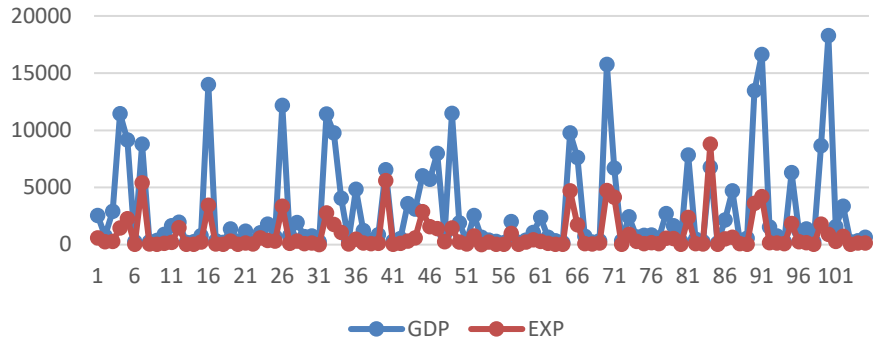

Fig. 1. Raw Data for GDP and EXP variables

Now, we go through the process of normalization. Let V_{LMF}^{GDP} (V_{LMF}^{EXP}) be the numerical vector of GDP (Export) created following the application of the LMF process and let V_{SMF}^{GDP} (V_{SMF}^{EXP}) be the numerical vector of GDP (Export) containing the results based on the SMF process. Now let

$$\Delta v_{LMF}^{GDP,EXP} = | v_{LMF}^{GDP} - v_{LMF}^{EXP}| \quad (4)$$

In other words, $\Delta v_{LMF}^{GDP,EXP}$ (Equation 4), measures the distance between v_{LMF}^{GDP} and v_{LMF}^{EXP}. This is shown in Figure 2a:

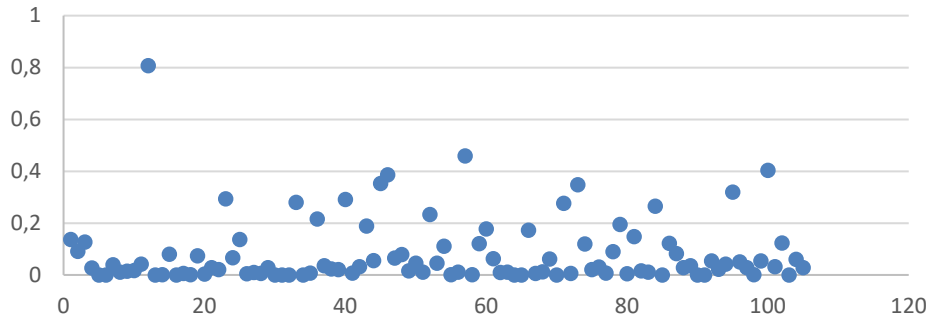

Fig. 2a.: Display of the difference (distance) between GDP and EXP based on the LMF approach (Equation 4).

We can now do the same thing with the SMF (Equation 5):

$$\Delta v_{SMF}^{GDP,EXP} = | v_{SMF}^{GDP} - v_{SMF}^{EXP}| \quad (5)$$

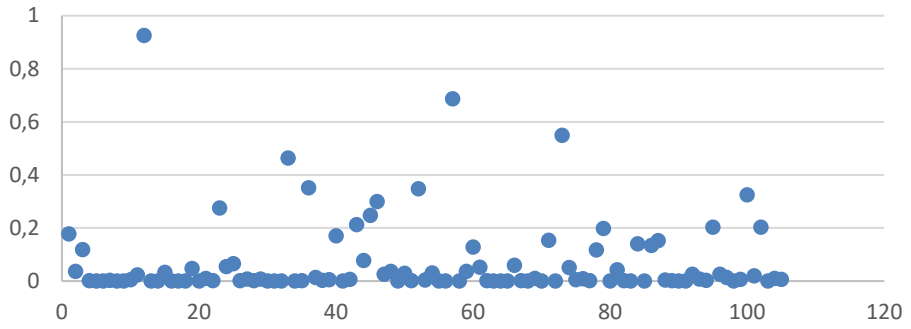

Fig. 3b.: Display of the difference (distance) between GDP and EXP based on the SMF approach (Equation 5)

2.2 Similarity between vectors

For comparison, we involve the following methods of similarity measurements:

- Minkowski Distance [17]: The Minkowski distance measures the distance between two vectors as follows:

$$d_m(A,B) = \sqrt[m]{\sum_{k=1}^{n} |a_k - b_k|^m}$$

where $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$ are the parallel elements of different numerical vectors, n is the size of the vectors and m is some coefficient. When $m = 1$, the distance becomes a Manhattan distance. When $m = 2$, the Minkowski distance is often called the Euclidean distance. The Minkowski distance with $m = 1{,}2$ is used very often in constructing clusters. The similarity (S_m) using the Minkowski Distance is defined as:

$$S_m(A,B) = 1 - \sqrt[m]{\sum_{k=1}^{n} |a_k - b_k|^m}$$

- Cosine Similarity [15], [16]:
Let $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$ be two vectors, then

$$S_{cos}(A,B) = \frac{A \cdot B}{\|A\|\|B\|}$$

where $\cdot$ indicates vector dot product and $\|A\| = \sqrt{\sum_{k=1}^{n} {a_k}^2}$ and $\|B\| = \sqrt{\sum_{k=1}^{n} b_k^2}$

- Fuzzy Linear Similarity Measure (FLSM):

$$S_{FLSM}(A,B) = 1 - \frac{1}{n}\sum_{k=1}^{n} |a_k - b_k| \quad (6)$$

where $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$ are the parallel elements of different numerical vectors and n is the size of the vectors.

- As can be seen, Equation 6 uses Equations 4 and 5 to compute the distance between the numerical vectors, whereas the Fuzzy Linear Similarity Measure (FLSM) is used to compute the similarity between two numerical vectors.

3 K-mean clustering

3.1 Clustering algorithm

The second step in performing cluster-based measure of similarity, is to divide both numerical vectors into clusters. In this paper we apply K-means Clustering method [1], [2], [3], which is a well-known, simple and widely used method. There are 2 variations to the K-means approach: (i) Fixed size clustering method (ii) Varied size clustering.

(i) Fixed size clustering method requires that the user decides in advance regarding the number of clusters in the numerical vector. The formal description of the algorithm is as follows:

Algorithm 1: Construction of a fixed size cluster set

Input: $A = (a_1, a_2, \dots, a_n)$
Output: K clusters of A

1. Select K elements as the initial means
2. Repeat
   a. Form K clusters by assigning all elements to the closest mean
   b. Recompute the mean of each cluster

   Until the means don't change

(ii) The varied size clustering differs from the fixed size methods in one important aspect: there is no need for the user to determine in advance the number of clusters that are needed. However, there is a requirement to determine in advance a maximum threshold distance between the value of the element and the value of the center (the mean) of the cluster. It essentially means that all the data elements having a distance below the given threshold from the center of the cluster – are similar to each other. The formal description of the algorithm is as follows:

Let r_{max} be the maximum threshold distance between 2 points such that they are considered similar

Algorithm 2: Construction of a varied size cluster set.

Input: $A = (a_1, a_2, \dots, a_n)$, r_{max}- the maximum threshold distance (radius) between 2 points
Output: The clusters of A (denoted by $\{C_j^A\}$)

1. Create the first cluster with the first element as its mean and $k = 1$
2. For $i \leftarrow 1$ to n do
   a. Find a cluster j ($j \in \{1, \dots, k\}$) such that $dist_j(a_i)$ is minimum
      where $dist_m(a_i) = |mean_m - a_i|$

and $mean_m$ is a mean of cluster m, for $m \in \{1, \dots, k\}$
(i.e., find a cluster j such that
$$\left|mean_j - a_i\right| = min_{m=1}^{k} |mean_m - a_i|).$$

b. If $dist_j(a_i) \leq r_{max}$ then add a_i to cluster j
Else, create a new cluster ($k \leftarrow k + 1$) and set its mean to be the value of a_i.

3. Recompute all the cluster means and repeat from stage 2, until the cluster means don't change

3.2 Case study

In this study, we utilize the model of factors facilitating economic performance to demonstrate the computation of similarities based on clusters. The model of factors facilitating economic performance was first introduced in [20]. We use six numerical vectors in our case study. The data were downloaded from the Data Base of the World Bank. The following variables (numerical vectors) were downloaded:

- GDP per capita (denoted by GDP)
- High Tech Exports per capita (denoted by High-Tech)
- Secondary Education Enrollment (denoted by Secondary)
- Birth Rate
- Tertiary Education Enrollment (denoted by Tertiary)
- Exports per capita (denoted by Exports)

Each numerical vector consists of cross-national data (for the year 1985), by country. The data of all the variables were normalized, and hence brought to the same scale. All the values in the numerical vectors are between zero and one, which allows us to use the same maximum threshold distances for construction the clusters in all six data series.

In order to understand the measurement reliability issues pertaining to the model, we explain in more details, what are the individual variables and what are the underlying difficulties in measuring such variables. The variables comprising our model are:

1. High technology per capita (High-Tech) - refers to exports (per capita) of products associated with advanced technologies. These variable measures total amount of income earned by exporting advanced-technology-intensive products and services; It is directly related to the dependent variable.
2. Secondary education enrollment (Secondary)- Percentage of the relevant population group that attends secondary education institutions; directly related to the dependent variable.
3. Birth Rate – Measures average amount of births per 1000 people during a given year; Inversely related to the dependent variable.

4. Tertiary education enrollment (Tertiary)- Percentage of the relevant population group that attends tertiary education institutions; directly related to the dependent variable.
5. Exports per capita (Exports) - Measures total amount of income earned by exporting products and services to other countries; directly related to the dependent variable.

It is reasonable to expect that the measurements of Birth Rate (measuring the amount of births per 1000 people) and of Secondary (percentage of the relevant age group enrolled in the Secondary Education Institutions) are reasonably accurate across the various countries. However, the other time series are much more problematic and less reliable, especially due to the international characteristics of the data, as clarified below.

Tertiary: In the broad terms, the definition of this variable is clear. It refers to enrollment to educational institutions, where students pursue their academic studies following the completion of the secondary (high school) education. However, there is no uniform, worldwide-accepted standard, regarding the academic accreditation of tertiary institutions. Some institutions are easily accredited in some countries, while in other countries they would not be accredited. In addition, accreditation rules are not constant and change continuously in various countries. Thus, when we see substantial increase of Tertiary enrollment rate in a given country, we usually do not know if it is truly due to a larger percentage of the population enrolling to study in academic institutions, or it is due to change in accreditation rules, such that the same students that were not accounted for under the previous rules, now are included. Of course, from economic impact perspective, there is a great difference between increasing Tertiary enrollment (increasing the number of students), vs adding already enrolled students by just changing the status of their institutions.

High Tech: This variable is also very problematic as far as reliability of measurements and their interpretations. As stated above, the variable High Tech supposedly reflects competitive advantage of some economies in sectors requiring top skills in advanced scientific and technological domains. However, in practice this is not always the case. Some countries develop high-tech core components, while other countries are capable to produce such components cheaper. Mass production usually requires lower skills in comparison to Research and Development. In addition, numerous products consist of high-tech components and low-tech components. Some countries import various (high-tech and low-tech) components, possibly produce some additional low-tech parts, and then assemble all the components into the final product. However, very often, for the purpose of prestige, such exports are classified by governments as High-Tech exports, even when the high-tech core components contained within the product are imported.

Export is another problematic variable as far as reliability of its measurements. Exports variable is a proxy for the degree of international competitiveness. The term “international competitiveness” reflects the ability of a given country to produce products

and services in a competitive manner within international markets. If all the components of the exported products were produced by the exporting economy, this would reflect its global competitive capabilities. However, in reality, very high proportion of the exported products contain some percentage of imported components, which the exporting country imports and later re-exports as a component of another product. Additional measurement problem: If Export would consist, only of commodities exported worldwide, then its measurements would be fairly accurate, because customs services in each country monitor and record all the products imported into the country. However, exports involve more than just commodities. They include payments for services, flow of dividends and remittances. These money flows are not under control of customs services and are greatly affected by the tax laws in various counties. Some international money flows are taxable and some are not, and this greatly affects how various money flows are categorized when reporting to authorities. Therefore, the steps to minimize tax liability substantially affect the magnitude of the variable "Exports" (by definition, exports include only flows of money categorized as earned income, and excludes financial transfers).

A dependent variable, representing successful long-term economic performance, could be selected as one of the following measures of income per capita or value of output per capita, such as GDP per capita, GNP per capita and GNI per capita, all of them are well known measurements of the value of economic activity. For convenience, from now on, we refer to all of them as GDP. All these measurements are indices and have several built-in deficiencies, greatly affecting the reliability of their measurements.

The most important source of deficiency in measuring GDP is the well-known under-reporting of income. The reason for under-reporting is the willingness to conceal some portion of income in order to reduce the payment of income tax. Despite the fact, that statistical authorities (measuring country's GDP) and tax authorities are usually different entities, the tax evaders are unwilling to have any evidence of their tax evasion, due to penalty involved.

The measurements of GDP are the aggregates of several components. Some of these components are problematic and lead to potential distortions due to their definitions, which simplify measurements while misrepresenting reality. Following are some examples:

1. GDP is defined as an aggregate value of all products and services produced and sold in a given time period. However, when firms cannot sell their products at a face value price and are forced to lower prices to clear their warehouses, the value of sales for computing GDP is still counted at face value prices.
2. When there is a severe economic crisis and lack of demand by buyers, such that large percentage of unsold products accumulate in warehouses, there is a very neat accounting procedure that presents a more favorable picture: all the unsold products are redefined as investment in inventories, and as such added to the GDP at their face value.
3. In all modern economies, public sector constitutes high percentage of GDP. However, for most of government services and investments, there is no market value,

because they are not sold in the markets. Hence, government expenditures are treated as if representing market value of such services and investments. Therefore, governments can pay large amounts of money to officials supposedly providing services, or to straw companies supposedly involved in projects, and all that money will be added to GDP even if it was in fact wasted, transferred or stolen money. In addition, there could be substantial percentage of bureaucracy, receiving nice salaries, and doing almost nothing useful. Nevertheless, all this money will be added to GDP as value of public services. Relatively efficient and transparent governments might be more successful in minimizing such negative effects. However, in reckless, lacking transparency governments, the proportion of money flows in public sector, that do not produce any useful services or products, could be substantial.

Since in our model the dependent variable is GDP, it means that we measure similarity of GDP to all other variables. Due to the measurement problems elaborated above, all such computations are subject to inaccuracies, which justifies similarity measures by clusters as the appropriate tool.

3.3 Dividing numerical vectors into clusters

Table 1 displays the results of dividing the numerical vectors into clusters, using different distance thresholds, ranging from 0.1 to 0.3. It can be observed, that as maximum threshold increases, the number of clusters decreases. In such cases, domain experts will have to decide, what maximum threshold (and the resulting amount) of clusters is the most appropriate for a specific model under consideration.

Table 1. Possibilities of dividing the numerical vectors into clusters

		Number of clusters					
		GDP	High-Tech	Secondary	Birth Rate	Tertiary	Export
r_{max}	0.10	8	7	8	9	8	7
	0.15	6	5	6	6	5	5
	0.20	5	5	5	4	5	5
	0.25	5	4	4	4	5	4
	0.30	4	4	4	4	4	4

4 Using clusters to compute similarity

To demonstrate the generality of the method presented here, three different models of computing similarity based on clusters are presented. The general idea of similarity by clusters is as follows: we start with a given data element from one of the numerical vectors, and measure whether that element could belong to the cluster containing the parallel data element from another numerical vector. If it could belong to that cluster,

then we assign it a score of one (the highest score, since the two elements can be associated with the same cluster). If on the other hand, the element could not belong to the relevant cluster, then score is less than 1, depending on the model used.

For example, assume we have two vectors A and B (see Figure 3). Also, assume that we divide the vectors A and B into clusters such that each vector is associated with different clusters. Now, also assume that we have two data points $a_i \in A$ and $b_i \in B$ and i is some index. As can be seen, a_i is in cluster C_k. If b_i is within the borders of C_k, then we conclude that the two data points are similar. The degree of similarity is determined by the model we use.

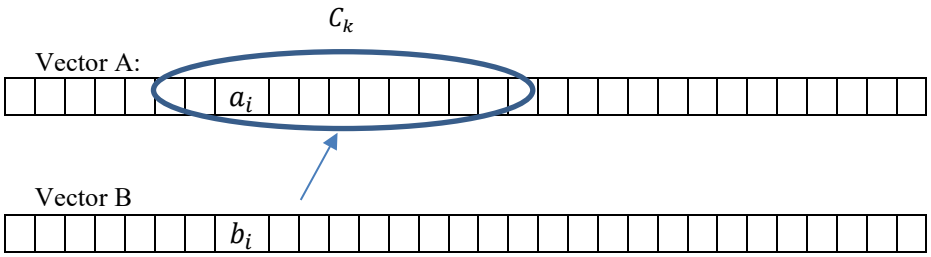

Fig. 4. The idea of similarity measure.

Algorithm 3, presented below, represents all 3 different models to compute similarity. In other words, all the components of the algorithm are the same for all three different models, except the component that computes the scores.

Successful model is a model consisting of explanatory variables, such that the similarity of each one of these explanatory variables to the dependent variable is as high as possible.

Algorithm 3: Three models for computing similarity (Boolean, Exponential and Linear)

Input:$A = (a_1, a_2, \dots, a_n)$,$B = (b_1, b_2, \dots, b_n)$, r_{max}- the maximum distance threshold between 2 points
Output: $S(A, B)$

1. Based on Algorithm 2, divide A into clusters (denoted by $\{C_j^A\}_{j=1}^{l}$)
2. $Sum_A \leftarrow 0$
3. For $i = 1$ to n do
   a. Suppose that a_i belongs to the cluster C_j^A. Find a cluster k ($k \in \{1, \dots, l\}$) such that $dist_k(b_i)$ is minimum where
   $$dist_j(b_i) = |mean_j^A - b_i|$$
   and $mean_j^A$ is a mean of cluster C_j^A

b. $Sum_A \leftarrow Sum_A + \varphi(k,j)$

Similarly, in the following 3 stages, repeat steps 1-3 while interchanging between A and B:

4. Based on Algorithm 2, divide B into clusters (denoted by $\{C_j^B\}_{j=1}^r$)
5. $Sum_B \leftarrow 0$
6. For $i = 1$ to n do
   a. Suppose that b_i belongs to the cluster C_j^B.
      Find a cluster k ($k \in \{1, \dots, r\}$) such that $dist_k(a_i)$ is minimum where $dist_m(a_i) = |mean_m^B - a_i|$ and $mean_m^B$ is a mean of cluster C_m^B
   b. $Sum_B \leftarrow Sum_B + \varphi(k,j)$
7. $S(A,B) \leftarrow \frac{1}{2n}(Sum_A + Sum_B)$

The equations to compute score $\varphi(k,j)$ for each one of the models are as follows:

Boolean score function:

$$\varphi(k,j) = \begin{cases} 1 & ,k = j \\ 0 & ,k \neq j \end{cases} \tag{7}$$

Exponential score function:

$$\varphi(k,j) = \begin{cases} 1 & ,k = j \\ 2^{-|j-k|} & ,k \neq j \end{cases} \tag{8}$$

Linear score function:

$$\varphi(k,j) = \begin{cases} 1 - 0.25|k-j| & ,|k-j| = 0,1,2,3 \\ 0 & ,else \end{cases} \tag{9}$$

Note: If $k = j$ then $\varphi(k,j) = 1$. Else, $0 \leq \varphi(k,j) < 1$.

In other words, in Boolean model (Equation 7), if the distance is not 0, the value of the score function (in short, score) is zero ($\varphi = 0$). In two other models, we are basically assigning count number for each cluster in both data series. Then we compute score based on the difference between the count numbers for the clusters containing the equivalent data elements.

Table 2. Comparison of similarity measures using various models

A	B	r_{max}	S_{BOOL}	S_{EXP}	S_{LIN}	S_{cos}	S_{FLSM}
GDP	High-Tech	0.10	0.620	0.698	0.763	0.766	0.871
		0.15	0.655	0.731	0.795		
		0.20	0.730	0.792	0.840		
		0.25	0.735	0.839	0.852		
		0.30	0.830	0.883	0.905		
	Secondary	0.10	0.435	0.550	0.579	0.710	0.765
		0.15	0.467	0.647	0.648		
		0.20	0.500	0.680	0.684		
		0.25	0.528	0.716	0.717		
		0.30	0.594	0.757	0.773		
	Birth Rate	0.10	0.399	0.532	0.537	0.721	0.751
		0.15	0.449	0.610	0.613		
		0.20	0.479	0.673	0.673		
		0.25	0.530	0.707	0.714		
		0.30	0.576	0.740	0.745		
	Tertiary	0.10	0.443	0.589	0.601	0.788	0.773
		0.15	0.522	0.686	0.688		
		0.20	0.565	0.738	0.742		
		0.25	0.614	0.769	0.764		
		0.30	0.636	0.776	0.782		
	Export	0.10	0.655	0.775	0.797	0.808	0.889
		0.15	0.731	0.836	0.853		
		0.20	0.771	0.862	0.862		
		0.25	0.764	0.867	0.880		
		0.30	0.839	0.915	0.915		

Table 2 above display the results of the various models addressed in our study. In the case of all the three cluster-based models, as could be expected – when the maximum distance threshold increases, the similarity increases as well. The explanation and justification of such behavior is obvious: when the amount of clusters decreases, the size of the clusters (amount of data elements in it) increases, thus increasing the possibility that two parallel data elements will become part of two parallel clusters (clusters are having the same amount of data elements). The results of the Boolean model are lower in comparison to the other two models due to the distortions implied by its Boolean nature: when the element could not belong to the relevant cluster, then no matter how far it is from the closest edge of that cluster, the score is always 0

Table 2 above also presents the results of Exponential Model. The idea behind determining the score in the exponential model is as follows: when an element of the first numerical vector could belong to the cluster containing the parallel data element of the second numerical vector, then the score is 1. Else, if the parallel element is not a member of the cluster having the same count number, but is a member of the neighboring cluster, the score is 0.5. If the parallel element is contained in the further cluster, the score is 0.25, if it is even further; the score is 0.125 and so on (see Equation (8))

The exponential and the linear models are based on the very similar concepts. When parallel data elements are contained in the corresponding clusters, in both models the score is 1. However, when they are not located in the corresponding clusters, the score is lower than 1. As data elements are contained in clusters which are further apart (according to their count number), the exponential model assigns greater penalty, which means lower partial score that very quickly approaches 0. In contrast, linear model is more moderate, offers higher partial scores, which decline towards 0 more gradually (0.75, 0.5, 0.25 – see Equation (9)).
In general, one can observe, that Exponential and Linear models generate high scores for larger r_{max} in comparison to methods not based on clusters (Cosine and FLSM). In addition, the fact that the similarity measures based on clusters generate comparable results to other well-known methods reflects the reliability of the clustering-based models.

Note: The reader is reminded that the clustering-based similarity measurements are intended for applications in the cases where the data are highly unreliable, so that minor differences among individual data elements are meaningless.

5 Comparison between similarity methods based on different MFs

As we elaborated in section 2, each one of the similarity measurement methods presented above can be applied, based on several possible membership functions (MF). In our case study, we utilized a Linear Membership Function and Sigmoid Membership Function. The results of the case study are summarized in Table 3, presenting a comparison of the results generated by the three clustering-based similarity models, while utilizing both – linear and sigmoid membership functions for data normalization. It can be observed that the results for the two types of membership functions are more or less compatible, which indicates the robustness of the clustering-based similarity measurements.

Table 3. Comparison of similarity results based on Linear and Sigmoid MFs.

A	B	r_{max}	LMF			SMF		
			S_{BOOL}	S_{EXP}	S_{LIN}	S_{BOOL}	S_{EXP}	S_{LIN}
GDP	High-Tech	0.10	0.620	0.698	0.763	0.780	0.784	0.851
		0.15	0.655	0.731	0.795	0.795	0.803	0.864
		0.20	0.730	0.792	0.840	0.830	0.841	0.897
		0.25	0.735	0.839	0.852	0.881	0.899	0.926
		0.30	0.830	0.883	0.905	0.901	0.901	0.935
	Secondary	0.10	0.435	0.550	0.579	0.512	0.615	0.617
		0.15	0.467	0.647	0.648	0.534	0.671	0.677
		0.20	0.500	0.680	0.684	0.534	0.671	0.677
		0.25	0.528	0.716	0.717	0.674	0.783	0.786
		0.30	0.594	0.757	0.773	0.678	0.785	0.789
		0.10	0.399	0.532	0.537	0.510	0.598	0.603
		0.15	0.449	0.610	0.613	0.581	0.687	0.691
		0.20	0.479	0.673	0.673	0.591	0.709	0.713

	Birth Rate	0.25	0.530	0.707	0.714	0.627	0.761	0.762
		0.30	0.576	0.740	0.745	0.649	0.771	0.772
	Ter-tiary	0.10	0.443	0.589	0.601	0.561	0.652	0.671
		0.15	0.522	0.686	0.688	0.614	0.723	0.734
		0.20	0.565	0.738	0.742	0.632	0.737	0.747
		0.25	0.614	0.769	0.764	0.685	0.790	0.793
		0.30	0.636	0.776	0.782	0.711	0.827	0.829
	Export	0.10	0.655	0.775	0.797	0.741	0.805	0.832
		0.15	0.731	0.836	0.853	0.759	0.823	0.849
		0.20	0.771	0.862	0.862	0.816	0.887	0.899
		0.25	0.764	0.867	0.880	0.816	0.897	0.900
		0.30	0.839	0.915	0.915	0.839	0.907	0.910

Table 3 summarizes all the similarity results generated by the 3 models. It is easy to see that the Boolean model generated the lowest results. The reason for lower results is due to the distortions which are a direct result of a Boolean approach, as was explained above. We can also observe that consistently, linear model generates higher results than the exponential model. The reason for that is (as was explained above) the more restrictive nature of the exponential model. Nevertheless, both (Exponential and Linear) models generate similar results. It is up to the users to decide which one of the two models is more appropriate for their purposes: either more restrictive score computation, or more generous one. The Sigmoid membership function is associated with higher scores in comparison to Linear Membership Function for all three models.

To summarize in broad terms:

The process of computing similarities between two numerical vectors is consistent and involves the following steps:

1. The data of all numerical vectors used in this study are normalized and brought to the same scale [0,1].
2. We utilize the dynamic K-means Clustering method based on the idea that the number of clusters can be of different in each numerical vector (Algorithm 2).
3. Three different models of similarity by clusters are presented.
4. We utilize 6 data series (numerical vectors) to demonstrate the application of the 3 models.

The study demonstrated the possibility to compute similarity between numerical vectors by clusters. In other words, when there is a strong suspicion, that the data series are severely distorted and unreliable, we can view each cluster as broadly representing the values of all its data elements.

6 Conclusions

This study demonstrates the feasibility to utilize clustering of data for the purpose of constructing similarity-measurement models. Three different models for computing similarity are presented: Boolean, Exponential and Linear. While the scores of the Boolean model are rather low, the scores of the Exponential model and the Linear Model are compatible with the scores of methods not based on clusters, such as Cosine and FLSM, which demonstrates the reliability of the clustering-based methods. Applying two different membership functions generated similar results, demonstrating the robustness of the method. The Sigmoid Membership function led to slightly higher scores in comparison to the FLSM. It is up to the user to select the appropriate membership function, which more accurately (in the opinion of the user) reflects the data and maintains the integrity of the data. Theoretically, when the reliability and the precision of data series are questionable, the clustering-based similarity measurements are expected to be more appropriate tools for modeling. The results of the case study, presented above, are in line with the theoretical foundations.

References

1. H.-P. Kriegel, P. Kröger, J. Sander and A. Zimek, "Density-based Clustering," *WIREs Data Mining and Knowledge Discovery,* p. 1 (3): 231–240., 2011.
2. R. Ng and J. Han, "Efficient and effective clustering method for spatial data mining," in *Proceedings of the 20th VLDB Conference*, Santiago, Chile, 1994.
3. T. Zhang, R. Ramakrishnan and M. Livny, "An Efficient Data Clustering Method for Very Large Databases.," in *Proc. Int'l Conf. on Management of Data*.
4. J. A. Hartigan and M. A. Wong, "Algorithm as 136: A K-means clustering algorithm," *J. of the Royal Stat. Soc. Ser. C (Applied Statistics),* pp. vol. 28, no. 1, pp. 100–108, 1979.
5. R. T. Ng and J. Han, "CLARANS: A method for clustering objects for spatial data mining," *," IEEE Trans. on Knowl. and Data Eng.,* pp. vol.14, no. 5, pp. 1003–1016, 2002.
6. C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. of the Am. Stat. Assoc.,* pp. vol. 97,no. 458, pp. 611–631, 2002.
7. D. H. Fisher, "Improving inference through conceptual clustering," in *Proc. 6th Nat. Conf. Artificial Intell. (AAAI-87)*, Seattle, WA, 1987.
8. L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, 1990.
9. T. Zhang, R. Ramakrishnan and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD'96)*, Montreal, Canada, 1996.
10. R. Agrawal, J. E. Gehrke, D. Gunopulos and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *Proc. ACM SIGMOD Int. Conf. Manage.Data (SIGMOD'98)*, Seattle, WA, 1998.
11. M. Ester, H. P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Data Min. and Knowl. Discov.,* pp. vol. 96, no. 34, pp. 226–231, 1996.

12. M. Ankerst, M. M. Breunig, H. P. Kriege and J. Sande, "OPTICS: Ordering points to identify the clustering structure," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD'99)*, Philadelphia, PA, 1999.
13. H. P. Kriegel, P. Kroger, J. Sander and A. Zimek, "Density-based clustering," *WIREs: Data Min. and Knowl. Discov.,* pp. vol. 1, no. 3, pp.231–240, 2011.
14. W. Wang, J. Yang and R. R. Muntz, "STING: A statistical information grid approach to spatial data mining," in *Proc. 23rdConf. Very Large Data Bases (VLDB)*, Athens, Greece, 1997.
15. G. Sidorov, A. Gelbukh, H. Gomez-Adomo and D. Pinto, "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Mode," Computación y Sistemas, p. 18 (3): 491–504, 2014
16. V. Novotny, "Implementation Notes for the Soft Cosine Measure.," in The 27th ACM International Conference on Information and Knowledge Management, Torun, Italy, 2018.
17. J. E. Gentle, "Matrix Algebra: Theory, Computations, and Applications in Statistics", Springer-Verlag, 2007
18. Omar Adil M. Ali, Aous Y. Ali and Balasem Salem Sumait, Comparison between the Effects of Different Types of Membership Functions on Fuzzy Logic Controller Performance, International Journal of Emerging Engineering Research and Technology Volume 3, Issue 3, PP 76-83, 2015
19. Jin Zhao and B.K. Bose, Evaluation of membership functions for fuzzy logic controlled induction motor drive, IEEE 2002 28th Annual Conference of the Industrial Electronics Society. IECON 02, 2002.
20. Shnaider E. and Haruvy N. (2008), Background Factors Facilitating Economic Growth Using Linear Regression and Soft Regression. Fuzzy Economic Review, vol. XIII, No 1, pages 41-55.
21. Attig A. and Perner P. ((2011). The Problem of Normalization and a Normalized Similarity Measure by Online Data. Transactions on Case-Based Reasoning Vol. 4, No 1, pp. 3-17. ©2011, ibai-publishing, ISSN: 1867-366X , ISBN: 978-3-942952-09-5.