Chapter 3
Similarity Measures

3.1 Introduction

The concepts of similarity and distance are crucial in many scientific applications. Similarity and distance measurements are mostly needed to compute the similarities/distances between different objects, an essential requirement in almost all pattern recognition applications including clustering, classification, feature selection, outlier detection, regression, and search. There exist a large number of similarity measurements in the literature; thus, selecting one most appropriate similarity measurement for a particular task is a crucial issue. The success or failure of any pattern recognition technique largely depends on the choice of the similarity measurement. The similarity measurements vary depending on the data types used. This chapter provides a detailed discussion on a large number of similarity measurements existing in the literature. An excellent tutorial on similarity measurement is available in Ref. [278]. A large portion of this chapter is based on the discussion presented in [278] and is included here for the sake of completeness.

Similarity and dissimilarity between objects can be measured based on several features. Thus, feature selection is required before computing the similarity/distance between objects. The similarity/distance between two variables varies depending on the values of the selected features. A general approach is to find the similarity/distance for each feature, and thereafter the total similarity/distance is determined by aggregating over these individual similarities/distances. Similarity measures can be broadly divided into several categories [46, 97, 186, 278, 293]:

1. Similarity measures for binary variables.
2. Similarity measures for categorical variables.
3. Similarity measures for ordinal variables.
4. Similarity measures for quantitative variables.
5. Dissimilarity between two groups of variables.

S. Bandyopadhyay, S. Saha, *Unsupervised Classification*,
DOI 10.1007/978-3-642-32451-2_3,

3.2 Definitions

In order to define some similarity/dissimilarity measures, first some features from the data sets have to be extracted. Thereafter, the relationships between different features of different objects are used to compute the similarity among them. Note that the measure of similarity will largely depend on the scale of the measurements and also on the data types.

Similarity primarily quantifies the relationship between different features of different objects. Suppose there are two objects i and j. Let the similarity between these two objects i and j be denoted by s_{ij}. The value of s_{ij} largely depends on the scale of the measurements and also on the data types. The *distance* or *dissimilarity*, on the other hand, measures the difference between two points based on their attribute values. Distance/*dissimilarity* can also be viewed as the disorder between two objects. If the normalized distance and similarity between two objects i and j are d_{ij} and s_{ij}, respectively, then they are related as below:

$$s_{ij} = 1 - d_{ij}.$$

Distance is a quantitative value, and generally it satisfies the following conditions [278]:

- $d_{ij} \geq 0$; the distance between two objects should be always positive or zero.
- $d_{ii} = 0$; distance is zero if and only if it is measured with respect to itself.
- $d_{ij} = d_{ji}$; distance is symmetric.
- $d_{ij} \leq d_{ik} + d_{jk}$; distance should satisfy the triangular inequality.

A distance measure is called a *metric* if it satisfies all the above four conditions. Hence, not all distances are metrics but all metrics are distances.

3.2.1 Need for Measuring Similarity

Similarity measurement between two data points/objects is very important in order to distinguish between different objects [278].

The main motivations for measuring similarity are the following:

- Objects can be distinguished from one to another.
- Any clustering algorithm can be applied to group them based on this similarity measurement (for example, using K-means clustering).
- After clustering the objects form several groups. Thus it is easier to understand the characteristics of each group.
- The characteristics/behavior of the clusters can be explained.
- Clustering of data objects may also help in organizing and retrieving the information in a more organized way.
- Similarity measurement is essential for performing classification. A new object can be classified into the group by using a similarity measurement.

- Similarity measurement can also be used to predict the behavior of a new object based on this grouping information.
- The structure within the data set can be explored by using this similarity measurement.
- Many other data mining techniques such as clustering, classification, and feature selection can be applied to particular data using this similarity measurement.
- The data can be simplified by using the relationship extracted after application of similarity measurement. These simplified data can be used by different data mining techniques in a smooth manner.
- Thus, decision-making and planning become easier after knowing the structure and prediction of the data.

The following sections describe ways of measuring similarity when the variables are of different types.

3.3 Similarity/Dissimilarity for Binary Variables

Binary variables can only take values 0 or 1/yes or no/true or false/positive or negative, etc. In order to measure the similarity or dissimilarity (distance) between two objects/points, each represented by a vector of binary variables, first the total number of occurrences of each value is counted. Below is an example where the distance between two binary variables is calculated [278].

Let there be two instances, H_1 and H_2, from the class "Human". Suppose there are three features "Height $>$ 5.2 ft", "Weight $>$ 60 kg" and "Male or Female". Then, if H_1 has the height of 5.8 ft, weight of 58 kg, and gender is Male, then H_1 is represented as "$H_1 = (1, 0, 1)$". If H_2 has height of 5.0 ft, weight of 50 kg, and gender is Female, then H_2 is represented as " $H_2 = (0, 0, 0)$". Here, both H_1 and H_2 are three-dimensional objects because each object is represented by three variables. Suppose:

$$
\begin{aligned}
m_{00} &= \text{total number of features having 0s in both objects,} \\
m_{01} &= \text{total number of features which have 0s for the } i\text{th object} \\
&\quad \text{and 1s for the } j\text{th object,} \\
m_{10} &= \text{total number of features which have 1s for the } i\text{th object} \\
&\quad \text{and 0s for the } j\text{th object,} \\
m_{11} &= \text{total number of features having 1s in both objects.}
\end{aligned}
$$

Then, the total number of features$(F) = m_{00} + m_{01} + m_{10} + m_{11}$.

Table 3.1 illustrates the concepts of m_{00}, m_{01}, m_{10}, and m_{11}.

For the above examples of H_1 and H_2, $m_{00} = 1$, $m_{01} = 0$, $m_{10} = 2$, and $m_{11} = 0$.

There exist a large number of binary distance measurements, among which the following are the most popular [278]:

Table 3.1 Confusion matrix

H_1	H_2	
	0	1
0	m_{00}	m_{01}
1	m_{10}	m_{11}

- Simple matching distance (d_{ij}) [14, 278]: This distance function is used to compute the dissimilarity between binary variables when the frequency of occurrence of 0 and 1 values are the same in the two variables; For example, the simple matching distance can be used to measure the dissimilarity between two 'gender' variables which have equal information for male and female classes. It is defined as follows:

$$d_{ij} = \frac{m_{01} + m_{10}}{F}.$$

  Here, F denotes the total number of features; For example, the simple matching distance between H_1 and H_2 is $\frac{0+2}{3} = \frac{2}{3} = 0.67$. Here, $m_{01} + m_{10}$ also provides the Hamming distance between two objects.
- Jaccard's distance [142, 278]: Jaccard's coefficient is a similarity measurement, whereas Jaccard's distance is a distance measurement. These measures are mostly used for those binary variables where the values 0 and 1 do not have equal frequency of occurrence. Let us take the example of a library. In a library there is a large collection of books. Thus, there are a greater number of books than a student can take. Suppose the task is to count the total number of similar books taken by two students. Then, it would be a time-consuming task to calculate the total number of not common books taken by two students, as done in case of computing the simple matching distance. Rather, it would be easier to calculate the total number of common books taken by two students. Jaccard's coefficient takes care of this. It is defined as follows:

$$s_{ij} = \frac{m_{11}}{m_{11} + m_{01} + m_{10}}.$$

  The Jaccard distance is computed as:

$$d_{ij} = 1 - s_{ij} = \frac{m_{01} + m_{10}}{m_{01} + m_{10} + m_{00}}.$$

  For the above example the Jaccard coefficient between H_1 and H_2 is 0, and the Jaccard distance is $\frac{2}{3}$. Jaccard's coefficient can also be generalized to apply to nonbinary variables. In that case, it is computed based on set theory. Suppose there are two sets A and B. Then, Jaccard's coefficient between these two sets is computed as follows:

$$s_{AB} = \left| \frac{A \cap B}{A \cup B} \right|.$$

Let us consider the following two sets: $A = \{2, 3, 4, 5\}$ and $B = \{3, 4, 5, 6\}$; so here $A \cup B = \{2, 3, 4, 5, 6\}$ and $A \cap B = \{3, 4, 5\}$. Then, $s_{AB} = \frac{3}{5} = 0.6$.

- Hamming distance [120, 277, 278]: The Hamming distance between two binary vectors is equal to the number of positions where they have distinct digits. This distance is defined as

$$d_{ij} = m_{01} + m_{10}.$$

Again consider the two binary variables $H1$ and $H2$. The Hamming distance between them is 2. Note that, the simple matching distance is equal to the Hamming distance divided by the word length (= total number of variables).

3.4 Distance for Nominal/Categorical Variable

Nominal/categorical variables are those which cannot be measured in a quantitative way [278]; rather, they represent some data categories. In case of nominal or categorical variables, numbers are only used to represent different categories; For example, gender is a nominal variable with value of 1 = male and 2 = female. The main characteristic of nominal or categorical variable is that categories should be labeled "consistently". The order of the labels is not important, but consistency is very important. As an example, gender labeling can be changed to 10 = female, 15 = male, without affecting the representation logic in any way.

However, this labeling should be used consistently while using some categorical variables. One can generate his/her own labeling as long as consistency is preserved. To calculate the distance between two objects having categorical features, first the number of possible categories for each feature has to be counted. In case of two categories, distance measures for binary variables such as simple matching, Jaccard's or Hamming distance can be used. If the number of category is more than two, transformation of these categories into a set of binary variables is needed. There are two methods to transform a categorical variable (with number of categories greater than 2) into binary variables:

1. Each category is represented by a binary variable.
2. Each category is represented by several binary variables.

Use of the above two methods produces different distance measures. Calculation of the distance function is based on the original variables. First the total number of binary variables needed to represent values of a categorical variable is determined. The distance between two categorical objects is then calculated as the ratio of the number of unmatched and the total number of binary variables used for the representation. If p = number of variables having 0s for the ith object and 1s for the jth object and q = number of variables having 1s for the ith object and 0s for the jth object, then

$$d_{ij} = \frac{p + q}{\text{total No. of binary variables to represent categorical variables}}.$$

3.4.1 Method 1: Each Category Is Represented by a Single Binary Variable [278]

In this procedure each category can be represented by a binary variable [278]. The distance between two objects is then calculated as the ratio of the number of unmatched to the total number of binary variables used to represent these categorical variables.

Suppose there are two variables: Color and Gender. Gender has two values: $0 =$ male and $1 =$ female. Color has three choices: white, blue, and red. Suppose there are three subjects: Ram (male) wears a red shirt, Rekha (female) wears a white shirt, and Mithi (female) wears a blue shirt. Each value of Color is assigned a binary variable. Let us set the first coordinate as Gender, and the second coordinate as Color (red, white, blue). Then, the feature vectors corresponding to the three objects, Ram, Rekha, and Mithi, are: Ram $= (0, (1, 0, 0))$, Rekha $= (1, (0, 1, 0))$, and Mithi $= (1, (0, 0, 1))$.

To compute the distance between objects, it is first computed for each coordinate. Suppose the Hamming distance is used (= length of different digits). Then:

- Distance(Ram, Rekha) is $(1, 2)$, and the overall distance for the two variables is $1 + 2 = 3$.
- Distance(Ram, Mithi) is $(1, 2)$, and the overall distance for the two variables is $1 + 2 = 3$.
- Distance(Rekha, Mithi) is $(0, 2)$, and the overall distance for the two variables is $0 + 2 = 2$.

The distance between two categorical objects is then calculated as the ratio of the number of unmatched and total number of binary variables used for the representation:

- Distance(Ram, Rekha) is $(1/1, 2/3)$, and the average distance for the two variables is $(1 + 2/3)/2 = 5/6 = 0.83$.
- Distance(Ram, Mithi) is $(1/1, 2/3)$, and the average distance for the two variables is $(1 + 2/3)/2 = 0.83$.
- Distance(Rekha, Mithi) is $(0/1, 2/3)$, and the average distance for the two variables is $(0 + 2/3)/2 = 0.33$.

3.4.2 Method 2: Each Category Is Represented by Several Binary Variables [278]

Suppose there are a total of c categories, then, a total number dv of binary variables are used to represent these categories, where $c < 2^{dv}$. Thus, $dv = \lceil \frac{\log c}{\log 2} \rceil$; For example, in the previous example in order to represent the color of shirts, two binary variables are needed because $\lceil \frac{\log 3}{\log 2} \rceil = \lceil 1.5 \rceil = 2$. Then, the three colors may be

represented as follows: red color by 00, white by 01, and blue by 10. Here again the assignment to dummy variables is somewhat arbitrary but consistent.

Let us consider the previous example, where there are two categorical variables Gender and Color; Gender has two values: 0 = male and 1 = female. Color has three choices: red, white, and blue. In order to represent Gender, a binary dummy variable is used, whereas in order to represent color two dummy binary variables are used. Then, Ram, Rekha, and Mithi are represented as follows: Ram = (0, (0, 0)), Rekha = (1, (0, 1)), Mithi = (1, (1, 0)).

To compute the distance between objects, it must be calculated for each original variable.

Suppose the Hamming distance is used. Then:

- Distance(Ram, Rekha) is (1, 1), and the overall distance is $1 + 1 = 2$.
- Distance(Ram, Mithi) is (1, 1), and the overall distance is $1 + 1 = 2$.
- Distance(Mithi, Rekha) is (0, 2), and the overall distance is $0 + 2 = 2$.

The distance between two categorical objects is the ratio of the number of unmatched and total number of binary variables used for their representation:

- Distance(Ram, Rekha) is $(1/1, 1/2)$, and the average distance for the two variables is $(1 + 1/2)/2 = 3/4$.
- Distance(Ram, Mithi) is $(1/1, 1/2)$, and the average distance for the two variables is $(1 + 1/2)/2 = 3/4$.
- Distance(Mithi, Rekha) is $(0/1, 2/2)$, and the average distance for the two variables is $(0 + 1)/2 = 1/2$.

3.5 Distance for Ordinal Variables

Ordinal variables generally provide an ordering of a given data set in terms of degrees. These variables are primarily used to indicate the relative ordering/ranking of some data points [127, 278]. Thus, the quantitative difference between these variables is not important, but their order is important; For example, consider the 'letter grades' which are used to provide marks to students. AA would be ranked higher than AB, and BB is higher than CC. Here again ranking is important, not the precise distance between AA and AB. Sometimes, numbers can also be used to represent ordinal variables. Below are some examples of ordinal scale:

- Relative grading: AA = excellent, AB = good, BB = average, DD = poor.
- Rank of priority: 1 = best, higher value indicates low priority.
- Rating of satisfaction: 1 = very dissatisfied, 100 = very satisfied.

In order to compute the distances/dissimilarities between ordinal variables, the following methods are mostly used: normalized rank transformation, Spearman distance, footrule distance, Kendall distance, Cayley distance, Hamming distance, Chebyshev/Maximum distance, and Minkowski distance. Below, some of these distances are described.

3.5.1 Normalized Rank Transformation

Here, using some normalization techniques [278], the given ordinal variables are first converted into quantitative variables [127, 237]. Thereafter, the distance between quantitative variables can be calculated by using any method described in Sect. 3.6. In order to determine the distance between two ordinal variables, these variables are first transformed into a ratio scale by applying the following steps:

- The ordinal value is converted into rank ($r = 1$ to R).
- The rank is normalized to a standardized value of zero to one [0, 1] by using the following formula:

$$x = \frac{r-1}{R-1}.$$

- The distances between these ordinal variables can be computed by considering the ordinal value as a quantitative variable. Any distance measure for quantitative variables can be used.

This distance can only be applied if an ordinal variable can be normalized as a quantitative variable. If such types of conversion need to be avoided, then other distance measures such as the Spearman distance, Kendall distance, Cayley distance, and Hamming distance for ordinal variables or the Ulam distance, Chebyshev/Maximum distance can be used.

Let us consider a conference paper review system; each reviewer has to review some papers; for each paper there is a questionnaire asking level of acceptance in terms of appropriateness, clarity, originality, soundness, and meaningful comparison. Each five acceptance criterion has five values: $-2 =$ very dissatisfied, $-1 =$ dissatisfied, $0 =$ average, $1 =$ good, $2 =$ excellent. Suppose the scores of two reviewers for a given paper are as follows:

	Appropriateness	Clarity	Originality	Soundness	Meaningful comparison
Rev1	0	−1	0	1	0
Rev2	1	0	1	0	0

Suppose that the aim is to measure the distance/dissimilarity between these two reviewers according to the answers. First, the ordinal scale is transformed into a ratio scale. The original index ($i = -2$ to 2) is ordered and converted into a rank ($r = 1$ to 5). The highest rank is $R = 5$. Then, the rank is normalized to a value [0, 1]. For instance, in position 1 of Rev1, there is $i = 0$, which converted to rank becomes $r = 3$, and the normalized rank is $\frac{3-1}{5-1} = 2/4 = 0.5$. The following conversions hold:

- Original index: -2 -1 0 1 2.
- Converted to rank: 1 2 3 4 5.
- Normalized rank: 0 $\frac{1}{4}$ $\frac{2}{4}$ $\frac{3}{4}$ 1.

Using the normalized rank as the new values, the coordinates of Rev1 become $(\frac{2}{4}\ \frac{1}{4}\ \frac{2}{4}\ \frac{3}{4}\ \frac{2}{4})$ and Rev2 become $= (\frac{3}{4}\ \frac{2}{4}\ \frac{3}{4}\ \frac{2}{4}\ \frac{2}{4})$. The Euclidean distance between Rev1 and Rev2 is

$$d_{12} = \sqrt{\left(\frac{2}{4} - \frac{3}{4}\right)^2 + \left(\frac{1}{4} - \frac{2}{4}\right)^2 + \left(\frac{2}{4} - \frac{3}{4}\right)^2 + \left(\frac{3}{4} - \frac{2}{4}\right)^2 + \left(\frac{2}{4} - \frac{2}{4}\right)^2} = 0.5.$$

3.5.2 Spearman Distance

The Spearman distance [127, 237, 278] is the square of the Euclidean distance between two ordinal vectors, computed as:

$$d_{ij} = \sum_{p=1}^{n} (x_{ip} - x_{jp})^2,$$

where n denotes the total number of ordinal features.

Suppose two persons P1 and P2 are asked to provide their preference on color. Let the choice of P1 be [Blue, White, Red] and the choice of P2 be [Red, White, Blue]. Suppose the pattern vector is [Blue, White, Red]; then the following rank vectors are obtained $A = [1, 2, 3]$ and $B = [3, 2, 1]$. Thus, the two vectors are represented as two points in a three-dimensional space. Point P1 has coordinates (1, 2, 3), and point P2 has coordinates (3, 2, 1).

Then, the Spearman distance of preference between P1 and P2 is $d_{12} = (1 - 3)^2 + (2 - 2)^2 + (3 - 1)^2 = 4 + 0 + 4 = 8$.

3.5.3 Footrule Distance

The footrule distance [127, 237, 278] is the summation of the absolute differences between the feature values of two ordinal vectors. The equation for this distance is very similar to that of city block distance or Manhattan distance for quantitative variables. Another name for this distance is the Spearman footrule distance. This distance is defined as:

$$d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|.$$

For, the previous example the rank vectors of P1 and P2 are (1, 2, 3) and (3, 2, 1), respectively. Then, the footrule distance between these two points is

$$d_{12} = 2 + 0 + 2 = 4.$$

3.6 Distance for Quantitative Variables

Quantitative variables can be measured on a numeric scale [278]. Thus, they are different from categorical variables, which primarily represent some categories, as well as from ordinal variables, which represent ordering of variables. These quantitative variables are measured as a number. Thus, any arithmetic operations can be applied to quantitative variables. Examples of quantitative variables are height, weight, age, amount of money, GPA, salary, temperature, area, etc.

Suppose the cost, weight, and time taken need to be measured. Let us use three quantitative variables to distinguish one object from another. The three features are cost (rounded down to thousands of rupees), time (in hours) and weight (in kg). Suppose our findings are Machine 1: (Rs. 1200, 4 hours, 10 kg); Machine 2: (Rs. 1700, 3 hours, 15 kg).

The two objects can be represented as points in three dimensions. Machine 1 has coordinates (1200, 4, 10) and Machine 2 has coordinates (1700, 3, 15). The dissimilarity (or similarity) between these two objects is then based on these coordinates.

3.6.1 Euclidean Distance

The most commonly used distance measure for quantitative variables is the Euclidean distance [277, 278]. In general, the term 'distance' refers to Euclidean distance. This primarily measures the root of the square differences between the coordinates of a pair of objects.

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}.$$

For example, if there are two quantitative variables $A = (2, 3, 4)$ and $B = (5, 9, 10)$, then the Euclidean distance between them is

$$d_{AB} = \sqrt{\left\{(2-5)^2 + (3-9)^2 + (4-10)^2\right\}} = \sqrt{\{9 + 36 + 36\}} = 9.$$

Euclidean distance is a special case of Minkowski distance with $\lambda = 2$.

3.6.2 Minkowski Distance of Order λ

This is again a general distance measure. Depending on the value of the variable λ, this distance reduces to several different distance functions. When $\lambda = 1$, the Minkowski distance [277, 278] reduces to the city block distance; when $\lambda = 2$, the Minkowski distance reduces to the Euclidean distance; when $\lambda = \propto$, the Minkowski

distance reduces to the Chebyshev distance. The Minkowski distance can be used for both ordinal and quantitative variables.

It is defined as follows:

$$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{n}(x_{ik} - x_{jk})^{\lambda}}.$$

For example, if there are two quantitative variables $A = (2, 3, 4)$ and $B = (5, 9, 10)$, then the Minkowski distance of order 3 between them is

$$d_{AB} = \sqrt[3]{\{(2-5)^3 + (3-9)^3 + (4-10)^3\}} = -7.714.$$

3.6.3 City Block Distance

Other names of this distance are the Manhattan distance,/ boxcar distance, and/ absolute value distance [277, 278]. It primarily measures the absolute differences between the coordinates of a pair of objects, calculated as

$$d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|.$$

Consider the above example. Here, the city block distance between objects A and B is $d_{AB} = |2-5| + |3-9| + |4-10| = 3 + 6 + 6 = 15$. The city block distance is a special case of Minkowski distance with $\lambda = 1$.

3.6.4 Chebyshev Distance

Chebyshev/Tchebyschev distance [277, 278] is again applicable for both ordinal and quantitative variables. Another name for this distance is the maximum value distance. It calculates the maximum of the absolute differences between the features of a pair of objects. The formula for this distance is:

$$d_{ij} = \max_{k} |x_{ik} - x_{jk}|.$$

Consider the two points $A = (2, 3, 4)$ and $B = (5, 9, 10)$. Then, the Chebyshev distance between A and B is $d_{AB} = \max\{|2-5|, |3-9|, |4-10|\} = \max\{3, 6, 6\} = 6$. The Chebyshev distance is a special case of the Minkowski distance with $\lambda = \infty$ (taking the limit).

3.6.5 Canberra Distance

The Canberra distance [82, 173, 278] measures the sum of the absolute fractional differences between the features of a pair of objects. Each term of the fractional difference ranges between 0 and 1. Note that, if both features are zeros, then it is assumed that $\frac{0}{0} = 0$. This distance is very sensitive to a small change when both coordinates are near to zero. It is calculated as

$$d_{ij} = \sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}.$$

For example, for the two vectors $A = (2, 3, 4)$ and $B = (5, 9, 10)$, the Canberra distance between these two points is $d_{AB} = \frac{3}{7} + \frac{6}{12} + \frac{6}{14} = 0.4286 + 0.5 + 0.4286 = 1.3571$.

3.6.6 Bray-Curtis Distance

The Bray-Curtis distance [82, 173, 278] is sometimes also called the Sorensen distance. This is a distance measure commonly used in botany, ecology, and environmental sciences. When all the coordinates are positive, this distance function takes a value between zero and one. A distance of zero corresponds to exactly similar coordinates. The Bray-curtis distance is undefined if both objects have zero coordinates. This distance function is calculated using the absolute differences divided by the summation.

$$d_{ij} = \frac{\sum_{k=1}^{n} |x_{ik} - x_{jk}|}{\sum_{k=1}^{n} (x_{ik} + x_{jk})}.$$

For two vectors $A = (2, 3, 4)$ and $B = (5, 9, 10)$, the Bray-Curtis distance between these two points is $d_{AB} = \frac{|2-5|+|3-9|+|4-10|}{(2+5)+(3+9)+(4+10)} = \frac{3+6+6}{7+12+14} = \frac{15}{33} = 0.45$.

3.6.7 Angular Separation

This is sometimes termed the coefficient of correlation [82, 173, 278]. It primarily measures the cosine angle between two vectors. This is a similarity measurement taking values in the range $[-1, +1]$ rather than a distance measure. When two vectors are similar, the angular separation will take higher values. Again in order to define angular separation, it is assumed that $\frac{0}{0} = 0$. The angular separation between two vectors is defined as:

$$s_{ij} = \frac{\sum_{k=1}^{n} (x_{ik} \times x_{jk})}{\sqrt{(\sum_{k=1}^{n} x_{ik}^2 \times \sum_{r=1}^{n} x_{jr}^2)}}.$$

For two vectors $A = (2, 3, 4)$ and $B = (5, 9, 10)$, the angular separation similarity between these two points is $s_{AB} = \frac{2*5+3*9+4*10}{\sqrt{(2^2+3^2+4^2)(5^2+9^2+10^2)}} = \frac{77}{77.29} = 0.99624$.

The origin of the name of this distance is as follows: The cosine angle between two vectors is represented as the dot product divided by the lengths of the two vectors:

$$\cos\theta = \frac{AB}{|A|.|B|}.$$

The length of a vector A, which is also termed the modulus, is the root of the square of its coordinates, $|A| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$. Thus,

$$\cos\theta = \frac{a_1b_1 + a_2b_2 + \cdots + a_nb_n}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}\sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}}.$$

Changing the notation of vectors into coordinates, the angular separation formula become:

$$s_{ij} = \frac{\sum_{k=1}^{n} x_{ik}x_{jk}}{\sqrt{(\sum_{k=1}^{n} x_{ik}^2 \sum_{r=1}^{n} x_{jr}^2)}}.$$

3.6.8 Correlation Coefficient

The correlation coefficient [82, 173, 278], which is another similarity measurement rather than a distance measurement, is also termed the linear correlation coefficient or/Pearson correlation coefficient. It is a special case of angular separation taking values in the range $[-1, +1]$. It is angular separation standardized by centering the coordinates on its mean value, defined as

$$s_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)}{\sqrt{(\sum_{k=1}^{n}(x_{ik} - \overline{x}_i)^2 \sum_{r=1}^{n}(x_{jr} - \overline{x}_j)^2)}},$$

where $\overline{x}_i = \frac{1}{n}\sum_{k=1}^{n} \overline{x}_{ik}$ and $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} \overline{x}_{jk}$. Here it has been assumed that $\frac{0}{0} = 0$.

For two vectors $A = (2, 3, 4)$ and $B = (5, 9, 10)$, the correlation coefficient between these two points is defined, using $\overline{A} = \frac{2+3+4}{3} = \frac{9}{3} = 3$ and $\overline{B} = \frac{5+9+10}{3} = \frac{24}{3} = 8$, as

$$\begin{aligned} s_{AB} &= \frac{(2-3)\times(5-8) + (3-3)\times(9-8) + (4-3)\times(10-8)}{\sqrt{((2-3)^2 + (3-3)^2 + (4-3)^2)\times((5-8)^2 + (9-8)^2 + (10-8)^2)}} \\ &= \frac{3+0+2}{\sqrt{(1+0+1)\times(9+1+4)}} \\ &= \frac{5}{\sqrt{28}} = \frac{5}{5.29} = 0.9452. \end{aligned} \tag{3.1}$$

3.6.9 Mahalanobis Distance

Another important distance measure is the Mahalanobis distance [82, 277, 278], also termed as the quadratic distance. It generally determines the difference/distance between two groups of objects. Suppose there are two groups with means $\overline{x}_i$ and $\overline{x}_j$, the Mahalanobis distance is calculated by the following formula:

$$d_{ij} = \left((\overline{x}_i - \overline{x}_j)^T \times S^{-1} \times (\overline{x}_i - \overline{x}_j)\right)^{\frac{1}{2}}.$$

The points within each group should have the same number of features, but the number of points in each group may vary (i.e., the number of columns in each group should be the same, but the number of rows can vary).

3.7 Normalization Methods

The process of transforming the value of a feature/attribute into a fixed range, usually 0 to 1, is generally referred to as normalization [278]. Here, a brief discussion on how to transform values of feature/attribute into the range from 0 to 1 or [0, 1] is provided.

Let us assume that there is a feature/attribute which takes values in the range $[f_{max}$ to $f_{min}]$. Suppose it is required to transform this value into the range [0, 1]. Let the original feature be denoted by f and the normalized feature be denoted by δ. There are several ways to normalize a particular feature value. First, all the negative values are converted to positive, and then each number is divided by some value which is larger than the numerator. Below some such techniques for transforming the original feature value into a normalized value are discussed in brief [278].

- If the range of the given feature vector is known a priori then the feature can be normalized by using the following equation:

$$\delta = \frac{f - f_{min}}{f_{max} - f_{min}}, \tag{3.2}$$

  where f_{min} denotes the minimum value of this feature vector and f_{max} denotes its maximum value. By using the above transformation one can normalize the value in to the range [0, 1]. If $f_{min} = f$, then $\delta = 0$; if $f = f_{max}$, then $\delta = 1$. If for a given data set $f_{min} = 0$, then Eq. 3.2 can be simplified to: $\delta = \frac{f}{f_{max}}$.
- Suppose the maximum value of a particular feature is not known but it can be assumed that the feature will always take the value 0 or some positive value. If there are a total of n possible values for that particular feature, then normalization of the ith feature can be done as follows:

$$\delta_i = \frac{f}{\sum_{i=1}^{n} f_i}. \tag{3.3}$$

Note that normalization using Eq. 3.3 will provide much lower values than normalization using Eq. 3.2, because $f_{max} \leq (\sum_{i=1}^{n} f_i)$.

- If the maximum value of a particular feature is not known and it also takes negative values, then normalization can be done using the following equation:

$$\delta_i = \frac{|f|}{\sum_{i=1}^{n} |f_i|}. \tag{3.4}$$

- Normalization of negative values: For data sets with positive or zero values, the above-discussed normalization techniques work fine. However, if the data set contains some negative values, then these numbers first have to be shifted by adding the absolute value of the minimum of these numbers such that the most negative one will become zero and all other numbers become positive. After that, any of the above-mentioned techniques can be applied to normalize the data.

  Suppose our data set is $\{-6, -9, 0, 6, 7\}$. The minimum of these numbers is -9. Now the minimum of all these numbers, $|-9| = 9$, has to be added to all five numbers. Then, the modified numbers are $\{-6+9, -9+9, 0+9, 6+9, 7+9\} = \{3, 0, 9, 15, 16\}$. Now, any of the above-mentioned techniques can be applied to this data set to normalize it.
- z-Score normalization: This is a statistical normalization technique. Here the assumption is that the data follows a normal distribution. By using the below-mentioned transformation, any data following a normal distribution can be converted into another normal distribution with mean zero and variance $= 1$. The standard deviation of the ith attribute is

$$Z = \frac{X - u}{s}. \tag{3.5}$$

  Here, the original data set is X, and the transformed data set is Z. s is the calculated standard deviation, and u is the mean of the data.

3.8 Summary

In this chapter we have discussed in detail several distance measures available in the literature. We have provided the definitions of similarity and distance measures. Different distance measures for binary variables, categorical variables, ordinal variables, and quantitative variables are discussed. Sufficient examples are provided to make them understood. Finally, the last part of the chapter contains a discussion on normalization. Different normalization techniques are elaborately explained. A good review of different similarity measures is available in [278], on which a large part of this chapter is based.

Clustering [143, 281] is an important problem in data mining and pattern recognition. It has applications in a large number of fields. In the next chapter, some existing approaches to clustering are discussed in detail.