Similarity metrics for classification: A Review

Shaker K. Ali¹, Zahoor M. Aydam² and Biadaa M. Rashed³

Computer sciences and mathematics college, University of Thi_Qar, Thi_Qar, Iraq.
¹ shaker@utq.edu.iq , ² zahoormosad@gmail.com and ³ bidaaalsafy@gmail.com

Abstract. In this paper, fourteen similarity metrics are reviews, which will be the most important part in Diagnosis, Classification, Clustering and Recognition. Most researchers may not sure to choose which metric will be powerful and give high accuracy in them researches. Therefore, this paper will be as a guide for them to select which metric useful for them research by try one of fourteen metrics that listed in this paper and can compare one of these metrics by advantage and disadvantage of each one. In addition, there are new metrics modified to give more accuracy by testing them in some clustering application.

Keywords: similarity metrics measurements, Diagnosis , Classification, Clustering, Recognition and distance-based measurements

1. Introduction

The similarity is a basic and common use concept and reflects the amount of relationship between the two elements of the database, while the difference deals with the measurement of the difference between the two data elements. The similarity scale should be between 0 and 1. Several similarity metrics methods, such as the content of information and information exchanged Dice coefficient, cosine coefficient, the distance-based measurements, and the feature contrast model [1]. A difficulty with the previous similarity measures is that each of them related to a specific application or assumes a particular domain model. Similarities play an increasingly important role in research and text-related applications in tasks such as information retrieval, classification, diagnosis, classification, clustering, recognition, document compilation, subject discovery, subject tracking, question generation, answer questions, results recording, short answer recording, machine translation, and others[2].
2. Data collection

Data collection methods are classify based on the method used. There are two methods of data collection [3]:

1. Quantitative Data

2. Qualitative Data

2.1. Quantitative Data

Data can be measure quantitatively and expressed, as numbers are quantitative data. Such as increasing the number of students in the class, the marks they received by the test, a number of news stories published on a topic, the number of times to repeat a specific word in the advice, etc. These ordinal and relative data can be represent. Be subject to statistical evaluation [3].

2.2. Qualitative Data

The qualitative data is the data that cannot expressed digitally. Data cannot expressed through nominal measures. This data can only expressed for example, nationalism, occupation, hobby…etc., this explain the events of behavioral interactions [4].

3. Feature selection Methods

The objective of this paper is to list the methods of similarity and difference of classes, by match in similarity or in difference between classes depending on features. The feature-selection ways must be finite and the number of features are big in training. We consider several popular feature-selection algorithms.

4. Similarity and differences measuring

This section briefly describes the similarity and differences of measurement that between classes.

4.1. Euclidean distance.

The most popular distance utilized to numerical data is the Euclidean distance. That is a private state of the Minkowski distance when n = 2. Euclidean distance functions good while used to datasets that contain consolidated or separated groups. Although Euclidean distance is quite popular in clustering, it has a disadvantage when the target data has no shared values; Distance may have smaller than the other consort vector data. A different issue for Euclidean distance as a class of the Minkowski metric is the larger feature will dominate the others. The Ecclesiastical equation between two points (p, q) is the length of the line connecting them \( \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \) as shown in the Equation 1 [5]:

\[
ED = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}
\]  

where ED is Euclidean distance, p, q is two points.

4.2. Cosine distance.

If two Vector measurement according to the corner amidst than, then the measure of their symmetry is considered meaningful. Also, in special, the appropriate similarity measure is the natural inner product as shown in Equation 2. Note that the scale of the cosine is fix to the rotation, is different from the linear transformations, and is independent of the length of the vector [6].

\[
C(p, q) = \frac{1}{\|p\| \cdot \|q\|} \sum_{i=1}^{m} p_i q_i
\]  

(2)
Where \( C(p, q) \) is cosine distance, where \( \|p\|_2 \) is the Euclidean norm of vector \( p = (p_1, p_2, \ldots, p_m) \) defined as 
\[
\|p\|_2 = \sqrt{p_1^2 + p_2^2 + \cdots + p_m^2}
\]
and \( \|q\|_2 \) is the Euclidean norm of vector \( q = (q_1, q_2, \ldots, q_m) \) defined as 
\[
\|q\|_2 = \sqrt{q_1^2 + q_2^2 + \cdots + q_m^2}
\]

4.3. City block distance.

The City block distance between vectors is the sum of absolute differences in their elements. Mathematically, the distance between the two vectors \( p \) and \( q \) is offered by Equation 3.[7]

\[
C(p, q) = \sum_{i=1}^{N} |p_i - q_i| 
\]  

(3)

4.4. Minkowski distance

Specific two q-dimensional cases,

\[
p_i = (p_{i1}, p_{i2}, p_{i3}, \ldots, p_{iq})
\]

\[
p_j = (p_{j1}, p_{j2}, p_{j3}, \ldots, p_{jq})
\]

The distance between an object and an object is calculated using Minkowski distance. As it is show in the Equation 4 [8].

\[
d(i, j) = ((p_{i1} - p_{j1})^2 + (p_{i2} - p_{j2})^2 + \cdots + (p_{iq} - p_{jq})^2)^{\frac{1}{q}}
\]

(4)

Where \( d(i, j) \) is minkowski distance, \( p_{iq}, p_{jq} \) is two q-dimensional.

4.5. Chebychev distance

Is the maximum of the differences of two points in different vectors as explain in Equation 5 [9].

\[
d(p, q) = \max i = 1,2,\ldots,m |p_i - q_i|
\]

(5)

Where \( d(p, q) \) is Chebychev distance, \( p_i, q_i \) is two points in different vectors.

4.6. Soft Cosine Similarity

It is novel modification of the cosine similarity by (Goutam Majumder, Partha Pakray, Alexander Gelbukh, David Pinto) when they take into account similarity of features in Vector space model. It is shown in Equation 6 [10].

\[
Soft\_cos(\Theta) = \frac{\sum_{i,j=1}^{m} p_{ij} X_i Y_j}{\sqrt{\sum_{i,j=1}^{m} p_{ij} X_i X_j} \sqrt{\sum_{i,j=1}^{m} p_{ij} Y_i Y_j}}
\]

(6)

where \( p_{ij} \) is the value of the matrix of symmetry between properties \( i \) and \( j \).

4.7. Modify of the standardized Euclidean distance

It is modification of standardized Euclidean distance, its ability to evaluate the value between two points in one dimension or two dimension, the Equation 7 is the modification of the standards Euclidean distance[11].

\[
Dst = \sqrt{\sum_{i=1}^{m} \frac{(p_i - q_i)^2}{\frac{1}{m-1} \sum_{j=1}^{m} (p_j - \bar{p})^2 + (q_j - \bar{q})^2}}
\]

(7)

Where :

\( p_i \): is the \( i \)th value of first vector value.

\( q_i \): is the \( i \)th value of second vector value.

\( m \): is the number of elements in vector.
is the mean value of first and second vector.

4.8. Standardized Euclidean distance

It is one way of evaluate the similarities between two points in the matrix or vector, the Equation 8 is the formula from the standards Euclidean.\cite{12}

\[ d_{pq} = \sqrt{\sum_{i=1}^{m} \frac{(p_i - q_i)^2}{s_i^2}} \]  

(8)

where \( s_i \) is the sample standard deviation of the \( i \)-th variable

4.9. Weighted Euclidean distance

It is also modification of Euclidean distance, the measured uses the weighted Euclidean in case of provides relative importance according to each feature define distance as explained in Equation 9\cite{13}.

\[ d_w = \left( \sum_{i=1}^{m} w_i (p_i - q_i)^2 \right)^{\frac{1}{2}} \]

(9)

where \( w_i \) is the weight specific to the \( i \)-th component

4.10. Mahalanobis distance

Mahalanobis distance is one of useful multivariate distance as other metrics that measures the distance between vector, which considered as point and a distribution\cite{14}.

\[ D_{mah} = \sqrt{(p - q)c^{-1}(p - q)^T} \]

(10)

where \( c \) is the covariance matrix of the dataset

4.11. Euclidean distance

The squared distance between two vectors \( p \) and \( q \):

\[ p = [p_1, p_2, \ldots, p_i, p_m] \]
\[ q = [q_1, q_2, \ldots, q_i, q_m] \]

is the total of squared Differences in the coordinates, to indicate the distance amidst vectors \( p \) and \( q \). The entry \( d_p, q \) can be written as follows\cite{15}:

\[ d_p, q = (p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_i - q_i)^2 + (p_m - q_m)^2 \]

(11)

4.12. Hamming distance

The Hamming distance is equal to the numeral of features essentially various in \( p \) and \( q \). Allows the computer program in which the exclusive logic is applied (XOR) is completed allows fast calculation the hamming distance\cite{16}

\[ D = \sum_{i=1}^{n} xor(p_i, q_i) \]

(12)
4.13. Spearman rank coefficient

The Spearman rank coefficient utilized to measure the symmetry between the two vectors (p vector) and (q vector). Spearman rank coefficient is describe as in the Equation 13[17]:

\[
D(p, q) = 1 - \frac{6 \sum_{i=1}^{d}(R(p_i) - R(q_i))^2}{d(d^2 - 1)}
\] (13)

4.14. Average distance

With regard to the defects in the Euclidean distance, the average distance is the amendment at the Euclidean distance to get better the results. For two data points p, q at m-dimensional space, the average distance is defined as in the Equation 14[18]:

\[
d_{av} = \left(\frac{1}{m} \sum_{i=1}^{m}(p_i - q_i)^2\right)^{\frac{1}{2}}
\] (14)

5. Comparison and analysis

In this section, we will discuss the comparative results of similarity measures and this section show the target and the conditions imposed on the similarity coefficients as shown in Table.1.

6. Discussion

This paper collect fourteen similarity measurements metrics, each metrics of similarity is necessary for anyway. The similarity measure efficiency conclusion includes TP values and TN values. The Equation 15 shows the accuracy of each fourteen metrics.

\[
Accuracy = \frac{TN + TP}{TN + TP + FP + FN}
\] (15)

7. Conclusions

This section explains the objective and limit of ways are gain from this review. Table 1 portrays the problem of measure similarity which is gained of this review.

| No | similarity Measure | Time complex | Advantage | Dis advantage |
|----|--------------------|--------------|-----------|--------------|
| 1  | Euclidean distance | O(n)         | very popular, simple to calculation and action well together datasets with consolidate or separated clusters. | Sensitive to extreme values |
| 2  | Cosine distance    | O(3n)        | separate of vector length and Fixed to the rotation | It is non fixed to a linear conversion |
| 3  | Cityblock distance | O(n)         | Is commonly and like another Minkowski-driven distance that runs good together with datasets by consolidated or separated groups. | Delicate to values extremistes. |
| 4  | Minkowski distance | O(n)         | The Minkowski distance works good when the dataset groups are If the data set is not separated and compressed, at the great-scale | |

Table 1. Comparison of multiclass measure similarity
| No. | Distance Method                          | Time Complexity | Description                                                                 |
|-----|-----------------------------------------|-----------------|-----------------------------------------------------------------------------|
| 5   | Chebychev distance                      | $O(n)$          | Effectively smaller distance parallel to column                             |
| 6   | Soft Cosine                             | $O(n^2)$        | soft cosine measure gained the best test results as compare to the standard cosine measure in the bulk of status. |
| 7   | Modify of the standardized Euclidean distance | $O(n)$          | Hyper way                                                                   |
| 8   | Standardized Euclidean distance         | $O(n)$          | simple to calculation and action well together datasets with consolidate or separated clusters. |
| 9   | Weighted euclidean                      | $O(n)$          | The weight matrix supports increasing the impact of more significant data points than Least important one |
| 10  | Mahalanobis                             | $O(3n)$         | Mahalanobis is a data-driven measure that Works the ability to reduce distortion distance through a linear set of features |
| 11  | Euclidean                               | $O(n)$          | self-same as Euclidean Distance                                             |
| 12  | Average distance                        | $O(n)$          | best than Euclidean distance deal with values outliers                      |
| 13  | Spearman rank coefficient               | $O(2n)$         | display the importance of that data, refute correlation, Provides to more analysis, Effectsn't suppose natural distribution. |
| 14  | Average distance                        | $O(n)$          | The variables independently contribute to measuring the distance. Where duplicate values can be controlled by the similarity between data points |
References

[1] F. Hassan, D. Cai-lin and Z. M. Hussain 2014 An Information-Theoretic Image Quality Measure: Comparison with Statistical Similarity( Journal of Computer Science).

[2] M.K.Vijaymeena and K.Kavitha 2016 A Survey On Similarity Measures In Text Mining( Machine Learning and Applications: An International Journal (MLAIJ) Vol.3, No.1).

[3] Male, T 2016 Collecting Quantitative Data( In I. Palaiologou, D. Needham and T. Male (Eds). Doing Research in Education: Theory and Practice. London: SAGE)PP. 192-208.

[4] Zubin Austin and Jane Sutton 2015 Qualitative Research: Data Collection, Analysis, and Management ( Can J Hosp Pharm  Vol. 68,No.3)PP.226–231.

[5] Sadina Gagula-Palalic 2008 Fuzzy Clustering Models and Algorithms for Pattern Recognition( Ph.D. International University of Sarajevo Mechatronics and System Engineering Graduate Program, Sarajevo ).

[6] Xu R, Wunsch D 2005 Survey of Clustering Algorithms [Internet]( IEEE Transactions on Neural Networks) PP. 645–678.

[7] DK H PHM Yassin, S Hoque, and F. Deravi 2013 Age Sensitivity of Face Recognition Algorithms( International Conference on Emerging Security Technologies, Cambridge, United Kingdom) pp. 12-15.

[8] Arindam Banerjee 2000 Clustering with Bregman Divergences( journal of Machine Learning Research Vol.6 ) PP.1705–1749.

[9] Khalakshan Kouser and Sunita 2013 A Comparative Study of K Means Algorithm by Different Distance Measures( International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 9).

[10] Sidorov, G., Gelbukh, A. F., G´omez-Adorno, H.and Pinto, D 2014 Soft Similarity and Soft Cosine Measure: Similarity of Features In Vector Space Model(Computaci´on y Sistemas, Vol. 18, No. 3) pp. 491–504.

[11] Shaker K .Ali and Zahoor M Aydam 2019 Convert Gestures of Arabic Numbers into Voice( Journal of Computational and Theoretical Nanoscience Vol.16,No.3)PP. 874-879(6).

[12] Pavol OR.ANSKÝ 2009 Fundamentals of Mathematical Statistics( Statistics Faculty of Management, University of Presov, Slovakia).

[13] Ji M, Xie F and Ping Y 2013 A Dynamic Fuzzy Cluster Algorithm for Time Series( Abstr Appl Anal, Vol. 2013 )PP. 1–7.

[14] Sujan Dahal 2015 Effect of Different Distance Measures in Result of Cluster Analysis( Thesis .Aalto University School of Engineering, Department of Real Estate, Planning and Geoinformatics).

[15] Shaker K. Ali and Zahoor M. Aydam 2018 Gestures Conversion to Arabic Letters( Journal of Thi-Qar University  Vol.13, No.4).

[16] Mingjin Yan 2005 Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion(Ph.D. Virginia Polytechnic Institute and State University).

[17] Diksha Kurchaniya and Punit Kumar Johari 2017 Analysis of Different Similarity Measures in Image Retrieval Based on Texture and Shape( International Research Journal of Engineering and Technology (IRJET) Vol. 04 ,Issue).

[18] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi and Teh Ying Wah 2015 A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data( PLOS ONE, DOI:10.1371/journal.pone.0144059).