

Dynamic Similarity Distance with Mean Average Precision Tool

Nur Atikah Arbain*, Mohd Sanusi Azmi, Sharifah Sakinah Syed Ahmad, Azah Kamilah Muda, Intan Ermahami A. Jalil and King Ming Tiang

Faculty of Information Technology and Communication, Universiti Teknikal Malaysia Melaka, Malaysia

ABSTRACT

In recent years, many classification models have been developed and applied to increase their accuracy. The concept of distance between two samples or two variables is a fundamental concept in multivariate analysis. This paper proposed a tool that used different similarity distance approaches with ranking method based on Mean Average Precision (MAP). In this study, several similarity distance methods were used, such as Euclidean, Manhattan, Chebyshev, Sorenson and Cosine. The most suitable distance measure was based on the smallest value of distance between the samples. However, the real solution showed that the results were not accurate as and thus, MAP was considered the best approach to overcome current limitations.

Keywords: Accuracy, Mean average precision, Ranking, Similarity distance

INTRODUCTION

Distance measure approach is essential for ranking method based on Mean Average Precision (MAP) performance (Kalofolias, 2015). For every ranking solution, a suitable

distance measurement should be decided earlier because it can determine whether the objects or items are naturally relevant to be in the same groups or clusters. This ranking is efficiently contracted by the best performance of MAP that is easily derived from the probability of a data point being relevant to the given query (Yue, Finley, Radlinski, & Joachims, 2007). Therefore, the determination of relevant items is based on the perfect choice of the distance measure as the latter plays an important role as the study target in order to obtain an interpretable result (Ahmad, 2014).

Euclidean distance is a popular distance measure used for metric variables. However, the Euclidean distance is sensitive to outliers

ARTICLE INFO

Article history:

Received: 15 August 2016

Accepted: 18 May 2017

E-mail addresses:

nuratikah9.arbain@gmail.com (Nur Atikah Arbain),
sanusi@utem.edu.my (Mohd Sanusi Azmi),
sakinah@utem.edu.my (Sharifah Sakinah Syed Ahmad),
azah@utem.edu.my (Azah Kamilah Muda),
b031210272@student.utem.edu.my (King Ming Tiang),
Jalil_ermahani@utem.edu.my (Intan Ermahami A.)

*Corresponding Author

and expensive in terms of resources and computing time (Weihs & Szepannek, 2009). Moreover, Euclidean distance assumes that data points are distributed around the mean value in a spherical manner, but in most cases, distribution of the data point is non-spherical. To overcome the limitation of Euclidean distances, we implement other distance measures such as Manhattan, Canberra, Sorensen, Chebyshev, and Angular Separation or Cosine.

The distance formula for calculating the similarity between items or objects (Cha, 2007) is presented. Improving the ranking result arises from the distances measure that is used to calculate the distances between data points. Therefore, different formulas will lead to different ranking solution. Table 1 shows the list of distance measures used in this study.

Table 1
List of similarity distance's formulation

Euclidean L_2	Manhattan	Canberra
$d_{\text{Eucl}} = \sqrt{\sum_{i=1}^d P_i - Q_i ^2} \quad (1)$	$d_{\text{Man}} = \sum_{i=1}^d P_i - Q_i \quad (1)$	$d_{\text{Can}} = \sqrt{\sum_{i=1}^d \frac{ P_i - Q_i }{P_i + Q_i}} \quad (1)$
Sorensen L_1	Chebyshev L_∞	Angular Separation or Cosine
$d_{\text{Sor}} = \sqrt{\frac{\sum_{i=1}^d P_i - Q_i }{\sum_{i=1}^d (P_i + Q_i)}} \quad (1)$	$d_{\text{Cheb}} = \max_i P_i - Q_i \quad (1)$	$d_{\text{Cos}} = \frac{\sum_{i=1}^d P_i Q_i}{\sqrt{\sum_{i=1}^d P_i^2} \sqrt{\sum_{i=1}^d Q_i^2}} \quad (1)$

The ranking R system that takes into account the value of recall and precision can be used to classify the image data correctly. The following formula evaluates precision value for different number m for the requested relevant items (outcomes):

$$AP = \frac{1}{|M|} \sum_{m \in M} P@m \quad (7)$$

Whereas the value of M is the precision value that we used to encounter relevant items in the ranking system in the proposed method. The symbol of $P@m$ is called Precision at (rank) m . The formula for calculating M is as follows;

$$M = \{m | 1_{REL}(R[m]) = 1, 1 \leq m \leq N\} \quad (8)$$

Where $R[m]$ denotes the index of the items while ranking R return at position m . Here, $1_{REL}(R[m])$ is 1 for the m in the ranking that contains relevant item (Kalofolias, 2015).

In this paper, we focus on the problem of handwriting recognition, where the goal is to classify automatically an image of 10 classes of different symbols of number. The success of the classification depends on the use of distance measure. The objective of this paper is to compare the usage of various distance measures and choose the most suitable one for each dataset used in the experiment.

This paper is organised as follows. First, in Section 2, related works and formula of distance measures used in this paper are presented. In Section 3, the proposed method is analysed while Section 4 highlights the dataset used in the experiments. Section 5 presents the results of comparison of various distance measures by cross-validation. Section 6 concludes the paper.

MATERIALS AND METHODS

The Dynamic Similarity Distance with Mean Average Precision (SD Tool) was proposed. It applied six types of distance methods, namely Euclidean, Manhattan, Canberra, Sorensen, Chebyshev and Angular Separation. This tool is dynamic because it is flexible to be applied to any different length of features and is also compatible with large database. Both train and test data can be any length of features (column) but the size of train and test data must be in the same length. However, this proposed tool only accepts the format of Comma Separated Value (CSV) file and applicable in digit or numeric form for both test and train data. The ranking measurement uses Mean Average Precision (MAP) in order to increase the results of similarity distance's accuracy. Figure 1 shows the first interface of SD tool while Figure 2 shows the interactive interface that contains file section, filter section, result section, ranking section and several buttons.

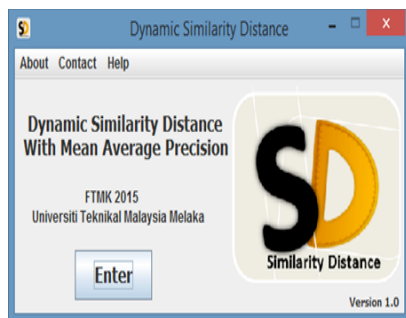


Figure 1. First interface of SD tool

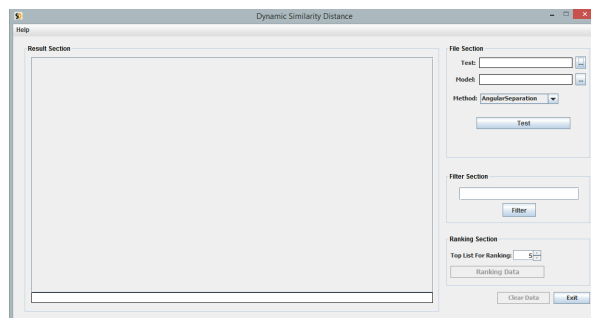


Figure 2. Interactive interface in SD tool

There were two stages in our proposed tool: a) first stage was the process of obtaining the nearest class based on the distance value of selected similarity distance method; b) the second stage involved a process that applied the ranking method which used the collection of results obtained from selected similarity distance method using MAP approach. Figure 3 is a flow chart describing the first stage while Figure 4 describes the second stage in SD tool.

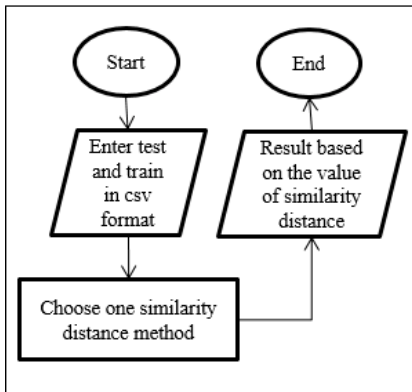


Figure 3. Flow chart of first stage

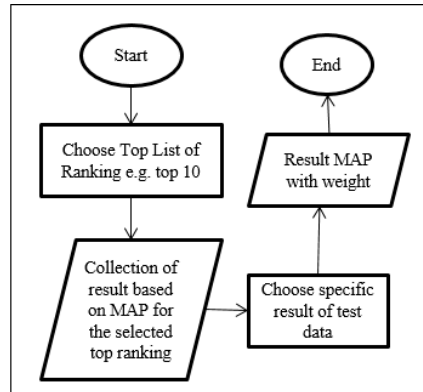


Figure 4. Flow chart of second stage

Equation (7) and (8) (Kalofolias, 2015) were applied into our ranking system. After applying the ranking method, the full results for each test data along with the value of weight are shown in Figure 5 while Figure 6 shows the results of the top ranking, i.e., top 10 ranking for selected test data. The results showed the nearest class to the selected test data was based on the highest value of weight. The ranking approach has also been used by Azmi, Nasrudin, Omar, & Ghazali (2012); Azmi, Omar, Nasrudin, Idrus, & Wan Mohd Ghazali (2013), and Azmi (2013).

Filename (Test)	Class (Model)	Weight
te_i9460_0.bmp	mnist0	10.0000
te_i9646_1.bmp	mnist1	10.0000
te_i6798_2.bmp	mnist8	2.5714
te_i7402_3.bmp	mnist3	3.8917
te_i8401_4.bmp	mnist4	8.7889
te_i7404_5.bmp	mnist5	4.5556
te_i730_6.bmp	mnist6	7.5710
te_i98_7.bmp	mnist7	4.8810
te_i4851_8.bmp	mnist8	9.0000
te_i9702_9.bmp	mnist9	6.0833

Figure 5. Full result of ranking for each test data

Filename (Test)	Class (Model)	Weight
te_i9702_9.bmp	mnist9	6.0833
te_i9702_9.bmp	mnist0	0.3651
te_i9702_9.bmp	mnist4	0.2500

Figure 6. Top ranking result for selected test data

In this paper, two types of digit dataset, namely HODA and MNIST, were used. Both contained 10 classes for each test and train data. The HODA dataset was developed in 2005 while the HODA dataset was published in 2007 by Khosravi & Kabir (2007). This dataset can be found at <http://FarsiOCR.ir>. The total HODA samples was 102,352 whereas the number of test data was 20,000, train data was 60,000 and the remaining was classified as others. Table 2 shows the digit from HODA dataset.

Initially, MNIST dataset was known as National Institute of Standard Technology (NIST). After some improvement, this dataset was renamed MNIST (Borji, Hamidi, & Mahmoudi, 2008). It was developed in 1992 and it can be found at <http://yann.lecun.com/exdb/mnist/>. The total of MNIST was 102,532 whereas the number of test data was 10,000 and model data was 60,000. Table 3 shows the Roman digit from MNIST dataset.

Table 2
HODA dataset











Class	Filename	Image
hoda0	th_1852_0	
hoda1	th_3852_1	
hoda2	th_5839_2	
hoda3	th_6430_3	
hoda4	th_9048_4	

Table 3
MNIST dataset

Class	Filename	Image
mnist5	te_i7404_5	
mnist6	te_i730_6	
mnist7	te_i98_7	
mnist8	te_i4851_8	
mnist9	te_i9702_9	

RESULTS AND DISCUSSION

In this experiment, five types of similarity distance method were used, namely Euclidean, Manhattan, Sorensen, Chebyshev and Cosine. Top 10 ranking method was used and the features from Azmi's study (2013) were utilised which introduced 9 features for a triangle shape. Two different dimensional length of features on HODA and MNIST dataset were tested respectively. Therefore, we have used 9 features as the least number of features and 297 features as the largest.

The results of HODA and MNIST experiments were based on the nearest class of train data for each of test data. There were 5 HODA and 5 MNIST test data. While, the HODA samples were picked randomly from 20,000 test data, the MNIST samples were picked randomly from 10,000 test data. The results of HODA and MNIST test data in Table 4 can be seen in Appendix A.

The result of HODA with 297 features showed that the Euclidean and Manhattan distance methods produced good results with the nearest class for each test data compared with other similarity distance methods. However, HODA with 9 features showed otherwise. After applying the top 10 ranking (MAP), the HODA results were improved especially for Chebyshev and Sorensen distance methods respectively. Based on the HODA results, the Euclidean, Manhattan Chebyshev and Sorensen distance methods were found to be the best distance for HODA dataset with 297 features.

The MNIST results with 297 features showed good results compared with 9 features after applying the top 10 ranking. For example, the Manhattan distance with 297 features showed that most test data had accurate results after applying ranking approach compared to Manhattan with 9 features which had the least test data with accurate results.

Overall, the results show that most accurate results are provided when using the largest features compared with the least features. It shows the number of features have affected the results of accuracy. After applying the ranking approach, most of the result were improved.

CONCLUSION

This paper presents the proposed tool that applies similarity distance method with ranking method in order to increase the results of accuracy. The proposed tool is flexible to handle any different length of features i.e. three features as the minimum including the filename, digit features and type of class. Experiments show that not all similarity distance method are the best distance to measure some datasets. Thus, the ranking method can help to improve the results of accuracy of similarity distance method. Further research is needed to improve the performance of the proposed tool. The SD tool is shared with public users. The proposed tool can be accessed freely at this link: https://www.dropbox.com/sh/wm8umplys8xjmtq/AAAaznDfaZt7_1JkwPb0auyja?dl=0.

ACKNOWLEDGMENT

The authors thank the Ministry of Education for funding this study through the following grants: FRGS/2/2014/ICT/02/FTMK/02/F00246 and PJP/2016/FTMK/HI4/S01477. Gratitude is also due to Universiti Teknikal Malaysia Melaka and Faculty of Information Technology and Communication for providing excellent research facilities.

REFERENCES

- Ahmad, S. S. S. (2014, December). Feature and instances selection for nearest neighbor classification via cooperative PSO. *Fourth World Congress on Information and Communication Technologies (WICT)* (pp. 45-50). IEEE.
- Azmi, M. S., Nasrudin, M. F., Omar, K., & Ghazali, K. W. M. (2012). Farsi/Arabic digit classification using triangle based model features with ranking measures. In *Int. Conf. Image Inf. Process. (ICIIP 2012)* (Vol. 46, pp. 128-133).
- Azmi, M. S., Omar, K., Nasrudin, M. F., & Muda, A. K. (2013). *Fitur baharu dari kombinasi geometri segitiga dan pengezonan untuk paleografi jawi digital* (Doctoral dissertation, Universiti Kebangsaan Malaysia).
- Azmi, M. S., Omar, K., Nasrudin, M. F., Idrus, B., & Wan Mohd Ghazali, K. (2013, April). Digit recognition for Arabic/Jawi and Roman using features from triangle geometry. In *AIP Conference Proceedings* (Vol. 1522, No. 1, pp. 526-537). AIP.
- Borji, A., Hamidi, M., & Mahmoudi, F. (2008). Robust handwritten character recognition with features inspired by visual ventral stream. *Neural Processing Letters*, 28(2), 97-111.

- Cha, S. H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2), 1.
- Kalofolias, J. (2015). Towards mean average precision. In *Symposium on Natural Language Processing* (pp. 1–6).
- Khosravi, H., & Kabir, E. (2007). Introducing a very large dataset of handwritten Farsi digits and a study on their varieties. *Pattern Recognition Letters*, 28(10), 1133-1141.
- Weihs, C., & Szepannek, G. (2009). Distances in classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5633 LNAI(1), 1–12.
- Yue, Y., Finley, T., Radlinski, F., & Joachims, T. (2007, July). A support vector method for optimizing average precision. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 271-278). ACM.

APPENDIX A

Table 4
Result of HODA and MNIST dataset

Test Image File	Features	Similarity Distance (SD) and Mean Average Precision (MAP) for Top 10									
		Euclidean		Manhattan		Sorensen		Chebyshev		Cosine	
		SD	MAP	SD	MAP	SD	MAP	SD	MAP	SD	MAP
th_1852_0	9	hoda0	hoda0	hoda0	hoda0	hoda1	hoda1	hoda0	hoda0	hoda1	hoda1
	297	hoda0	hoda0	hoda0	hoda0	hoda0	hoda0	hoda0	hoda0	hoda1	hoda1
th_3852_1	9	hoda1	hoda1	hoda1	hoda1	hoda1	hoda1	hoda1	hoda1	hoda1	hoda1
	297	hoda1	hoda1	hoda1	hoda1	hoda1	hoda1	hoda1	hoda1	hoda1	hoda1
th_5839_2	9	hoda0	hoda2	hoda0	hoda0	hoda1	hoda1	hoda0	hoda0	hoda1	hoda1
	297	hoda2	hoda2	hoda2	hoda2	hoda2	hoda2	hoda4	hoda4	hoda1	hoda1
th_6430_3	9	hoda3	hoda3	hoda3	hoda3	hoda1	hoda1	hoda3	hoda3	hoda1	hoda1
	297	hoda3	hoda3	hoda3	hoda3	hoda3	hoda3	hoda1	hoda1	hoda1	hoda1
th_9048_4	9	hoda2	hoda2	hoda2	hoda2	hoda1	hoda1	hoda2	hoda2	hoda1	hoda1
	297	hoda4	hoda4	hoda4	hoda4	hoda4	hoda4	hoda3	hoda3	hoda1	hoda1
te_i7404_5	9	mnist0	mnist2	mnist0	mnist2	mnist1	mnist1	mnist0	mnist1	mnist1	mnist1
	297	mnist5	mnist5	mnist5	mnist5	mnist5	mnist8	mnist3	mnist3	mnist5	mnist5
te_i730_6	9	mnist0	mnist0	mnist0	mnist0	mnist1	mnist1	mnist0	mnist0	mnist1	mnist1
	297	mnist6	mnist6	mnist6	mnist6	mnist6	mnist6	mnist2	mnist2	mnist5	mnist5
te_i98_7	9	mnist0	mnist0	mnist0	mnist0	mnist1	mnist1	mnist0	mnist0	mnist1	mnist1
	297	mnist7	mnist7	mnist7	mnist7	mnist7	mnist7	mnist5	mnist5	mnist5	mnist5
te_i4851_8	9	mnist0	mnist0	mnist0	mnist0	mnist1	mnist1	mnist0	mnist0	mnist1	mnist1
	297	mnist8	mnist8	mnist8	mnist8	mnist8	mnist8	mnist0	mnist2	mnist5	mnist5
te_i9702_9	9	mnist0	mnist0	mnist0	mnist0	mnist1	mnist1	mnist0	mnist0	mnist1	mnist1
	297	mnist9	mnist9	mnist4	mnist9	mnist9	mnist9	mnist9	mnist9	mnist5	mnist5