

# Classifying token frequencies using angular Minkowski $p$ -distance

Oliver Urs Lenz<sup>[0000–0001–9925–9482]</sup> and Chris Cornelis<sup>[0000–0002–7854–6025]</sup>

Research Group for Computational Web Intelligence,  
Department of Applied Mathematics, Computer Science and Statistics,  
Ghent University {oliver.lenz, chris.cornelis}@ugent.be

**Abstract.** Angular Minkowski  $p$ -distance is a dissimilarity measure that is obtained by replacing Euclidean distance in the definition of cosine dissimilarity with other Minkowski  $p$ -distances. Cosine dissimilarity is frequently used with datasets containing token frequencies, and angular Minkowski  $p$ -distance may potentially be an even better choice for certain tasks. In a case study based on the *20-newsgroups* dataset, we evaluate classification performance for classical weighted nearest neighbours, as well as fuzzy rough nearest neighbours. In addition, we analyse the relationship between the hyperparameter  $p$ , the dimensionality  $m$  of the dataset, the number of neighbours  $k$ , the choice of weights and the choice of classifier. We conclude that it is possible to obtain substantially higher classification performance with angular Minkowski  $p$ -distance with suitable values for  $p$  than with classical cosine dissimilarity.

**Keywords:** Cosine dissimilarity · Fuzzy rough sets · Minkowski distance · Nearest Neighbours.

## 1 Introduction

Cosine (dis)similarity [12, 13] is a popular measure for data that can be characterised by a collection of token frequencies, such as texts, because it only takes into account the relative frequency of each token. Cosine dissimilarity is particularly relevant for distance-based algorithms like classical (weighted) nearest neighbours (NN) and fuzzy rough nearest neighbours (FRNN). In the latter case, cosine dissimilarity has been used to detect emotions, hate speech and irony in tweets [9].

A common way to calculate cosine dissimilarity is to normalise each record (consisting of a number of frequencies) by dividing it by its Euclidean norm, and then considering the squared Euclidean distance between normalised records. Euclidean distance can be seen as a special case of a larger family of Minkowski  $p$ -distances (namely the case  $p = 2$ ). It has previously been argued that in high-dimensional spaces, classification performance can be improved by using Minkowski  $p$ -distance with fractional values for  $p$  between 0 and 1 [1].

In light of this, we propose *angular Minkowski  $p$ -distance*: a natural generalisation of cosine dissimilarity obtained by substituting other Minkowski  $p$ -

distances into its definition. The present paper is a case study of angular Minkowski  $p$ -distance using the well-known *20-newsgroups* classification dataset. In particular, we investigate the relationship between the hyperparameter  $p$ , the dimensionality  $m$ , the number of neighbours  $k$ , and the choice of classification algorithm and weights.

To the best of our knowledge, this topic has only been touched upon once before in the literature. Unlike the present paper, the authors of [5] do not evaluate classification performance directly, but rather the more abstract notion of ‘neighbourhood homogeneity’, and they only consider a limited number of values for  $p$  and  $m$ .

The remainder of this paper is organised as follows. In Section 2, we motivate and define angular Minkowski  $p$ -distance. In Section 3, we recall the definitions of NN and FRNN classification. Then, in Section 4, we describe our experiment, and in Section 5 we present and analyse our results, before concluding in Section 6.

## 2 Angular Minkowski $p$ -distance

In this section, we will work in a general  $m$ -dimensional real vector space  $\mathbb{R}^m$ , for some  $m \in \mathbb{N}$ .

The cosine similarity between any two points  $x, y \in \mathbb{R}^m$  is defined as the cosine of the angle  $\theta$  between  $x$  and  $y$ . We obtain the cosine dissimilarity by subtracting the cosine similarity from 1. Defined thus, cosine similarity and dissimilarity take values in, respectively,  $[-1, 1]$  and  $[0, 2]$ . However, when all records are located in  $\mathbb{R}_{\geq 0}^m$ , such as token frequencies, both measures take values in  $[0, 1]$ .

It is a well-known fact that cosine dissimilarity is proportional to the squared Euclidean distance between  $x$  and  $y$  once these points have been normalised by their Euclidean norm (note that  $\cdot$  denotes the vector in-product):

$$\begin{aligned}
 1 - \cos \theta &= 1 - \frac{x \cdot y}{|x| |y|} \\
 &= 1 - \frac{x}{|x|} \cdot \frac{y}{|y|} \\
 &= \frac{1}{2} \left| \frac{x}{|x|} \right| + \frac{1}{2} \left| \frac{y}{|y|} \right| - \frac{x}{|x|} \cdot \frac{y}{|y|} \\
 &= \frac{1}{2} \left( \frac{x}{|x|} \cdot \frac{x}{|x|} + \frac{y}{|y|} \cdot \frac{y}{|y|} - 2 \frac{x}{|x|} \cdot \frac{y}{|y|} \right) \\
 &= \frac{1}{2} \left( \frac{x}{|x|} - \frac{y}{|y|} \right)^2
 \end{aligned} \tag{1}$$

The Euclidean norm is the special case  $p = 2$  of the more general Minkowski  $p$ -size, defined for any  $x \in \mathbb{R}^m$  as:

$$|x|_p = \left( \sum |x_i^p| \right)^{\frac{1}{p}}, \tag{2}$$

where  $p$  is allowed to be any positive real number. Note that this is only a norm for  $p \geq 1$ . The Minkowski  $p$ -distance between any two  $x, y \in \mathbb{R}^m$  is defined as the  $p$ -size of their difference  $|y - x|_p$ . This is a metric if  $p \geq 1$ .

Similarly, we can also view the squared Euclidean norm (distance) as the special case  $p = 2$  of the *rootless* Minkowski  $p$ -size (distance), defined for any  $x \in \mathbb{R}^m$  as:

$$|x|_p^p = \sum |x_i^p|, \quad (3)$$

The rootless  $p$ -size is not a norm for any  $p$  (other than  $p = 1$ , for which it coincides with the ordinary 1-norm); rootless  $p$ -distance is a metric for  $p \leq 1$ .

With these definitions in place, we can define the angular Minkowski  $p$ -distance between any two vectors  $x, y \in \mathbb{R}^m$  as:

$$\left| \frac{y}{|y|_p} - \frac{x}{|x|_p} \right|_p. \quad (4)$$

as well as their rootless angular Minkowski  $p$ -distance:

$$\left| \frac{y}{|y|_p} - \frac{x}{|x|_p} \right|_p^p. \quad (5)$$

Thus, cosine dissimilarity corresponds to rootless angular Minkowski 2-distance, and we can consider angular Minkowski  $p$ -distance with different values for  $p$  as alternatives to cosine dissimilarity.

### 3 Classical and fuzzy rough nearest neighbour classification

We will now briefly review the definition of classical weighted nearest neighbour (NN) classification [4, 2, 3] and fuzzy rough nearest neighbour classification (FRNN) [7, 10]. Both approaches require a choice of a dissimilarity measure, weights, and a positive integer  $k$  determining the number of nearest neighbours to be considered. In what follows, we will specify the class prediction that each method makes for a test instance  $y$ , given a training set  $X$  and a decision class  $C \subseteq X$ .

#### 3.1 Nearest neighbour classification

For NN, let  $x_i$  be the  $i$ th nearest neighbour of  $y$  in  $X$ . Then the class score for  $C$  is given by:

$$\sum_{i \leq k | x_i \in C} w_i / \sum_{i \leq k} w_i \quad (6)$$

where  $w_i$  is the weight attributed to the  $i$ th nearest neighbour of  $y$ . Two popular choices [2, 3] for the weights are linear distance weights:

$$w_i = \begin{cases} \frac{d_k - d_i}{d_k - d_1} & k > 1; \\ 1 & k = 1, \end{cases} \quad (7)$$

and reciprocally linear distance weights:

$$w_i = \frac{1}{d_i}, \quad (8)$$

where  $d_i$  is the distance between  $y$  and  $x_i$ .

### 3.2 Fuzzy rough nearest neighbour classification

Properly speaking, FRNN consists of two different classifiers, the upper and the lower approximation, which can be combined to form the mean approximation. For the upper approximation, let  $d_i$  be the distance between  $y$  and its  $i$ th nearest neighbour in  $C$ . Then the class score for  $C$  is given by:

$$\overline{C}(y) = \sum_{i \leq k} w_i \cdot \min(0, 1 - d_i/2). \quad (9)$$

For the lower approximation, let  $d_i$  be the distance between  $y$  and its  $i$ th nearest neighbour in  $X \setminus C$ . Then the class score for  $C$  is given by:

$$\underline{C}(y) = \sum_{i \leq k} w_i \cdot \max(d_i/2, 1). \quad (10)$$

For the mean approximation, the class score for  $C$  is given by:

$$(\overline{C}(y) + \underline{C}(y)) / 2. \quad (11)$$

In the definition of both the upper and the lower approximation,  $\langle w_i \rangle_{i \leq k}$  is a weight vector of values in  $[0, 1]$  that sum to 1. As with NN, two popular weight choices are linear weights:

$$w_i = \frac{2(k+1-i)}{k(k+1)}, \quad (12)$$

and reciprocally linear weights:

$$w_i = \frac{1}{i \cdot \sum_{i \leq k} \frac{1}{i}}. \quad (13)$$

## 4 Experimental setup

To evaluate angular Minkowski  $p$ -distance, we conduct a case study on the well known text dataset *20-newsgroups* [8]. Originally, this contained 20 000 usenet posts from 20 different newsgroups (1000 each) from the period February-May 1993, and was collected by Ken Lang. We use the version of this dataset provided by the Python machine learning library *scikit-learn* [11], which comprises a training set (11 314 records) and a test set (7532 records, consisting of later posts than those in the training set), preprocessed to remove headers, footers and quotes.

We first convert each text into a set of words, defined as any sequence of at least two alphanumeric characters separated by non-alphanumeric characters, regardless of case. Next, we count the word frequencies per text and transform this into an  $m$ -dimensional dataset by selecting the top- $m$  overall most frequent words, and discarding the rest.

In order to evaluate the behaviour of NN and FRNN with angular Minkowski  $p$ -distance, we systematically vary different values for  $p$ ,  $m$  as well as the number of nearest neighbours  $k$ . In the case of FRNN, we consider the upper, lower and mean approximations separately. For both NN and FRNN, we will consider linear and reciprocally linear weights, as described in Section 3.

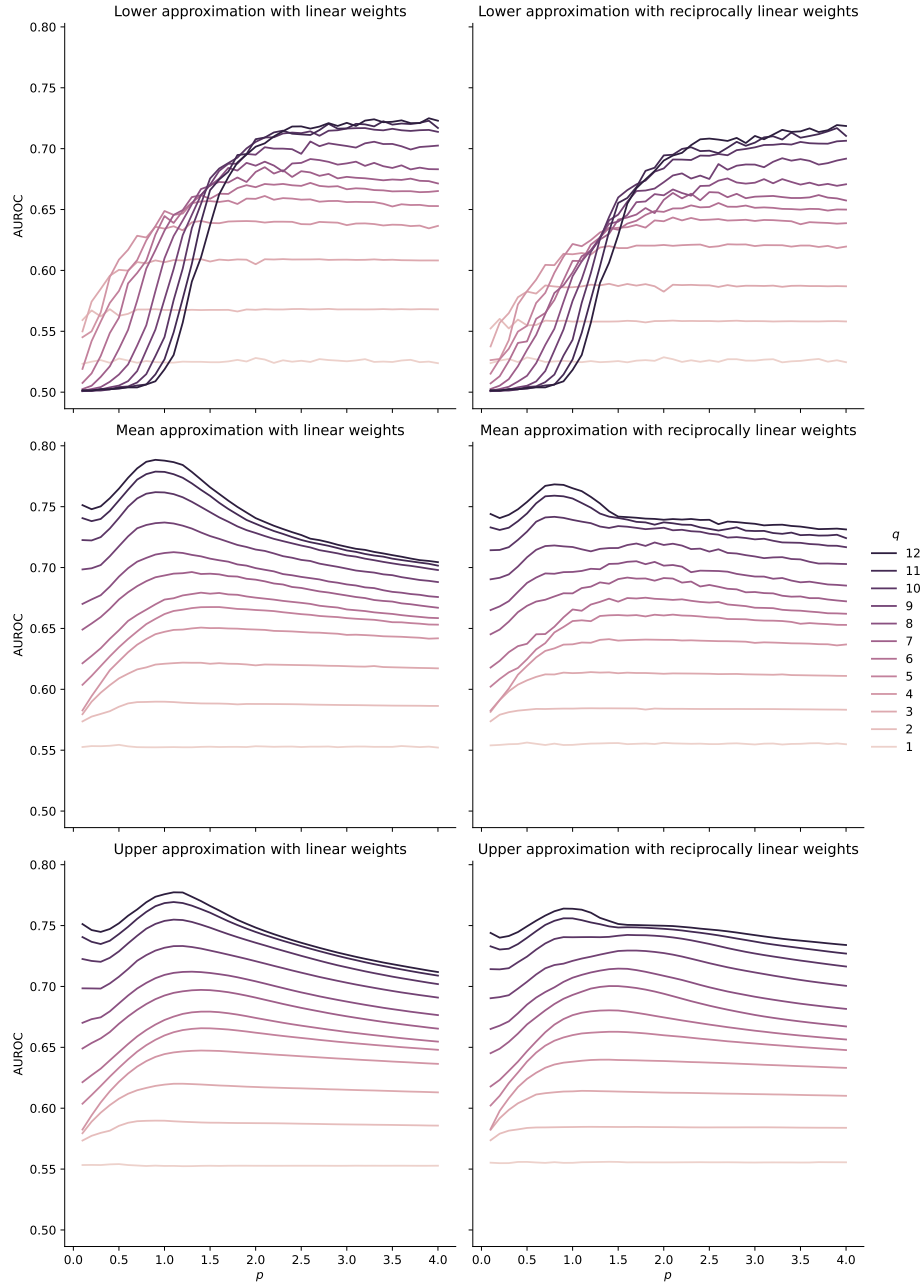
For  $p$ , we consider all multiples of 0.1 in the range of  $[0.1, 4]$ , centred on the canonical values of 1 and 2. Since  $k$  and  $m$  encode magnitudes, we investigate them on a logarithmic scale, with values corresponding to powers of 2 in the range of, respectively,  $[1, 256]$  and  $[2, 4096]$ .

We measure classification performance using the area under the receiver operator characteristic (AUROC) [6].

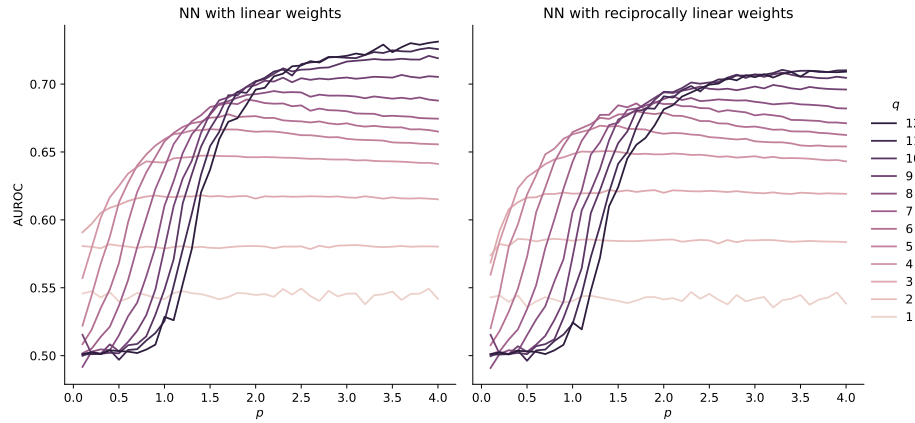
## 5 Results

Figures 1 and 2 display AUROC as a function of dimensionality (the number of most frequent tokens taken into consideration) and as a function of  $p$ , for  $k = 256$ . There are a few things to be noted from these response curves:

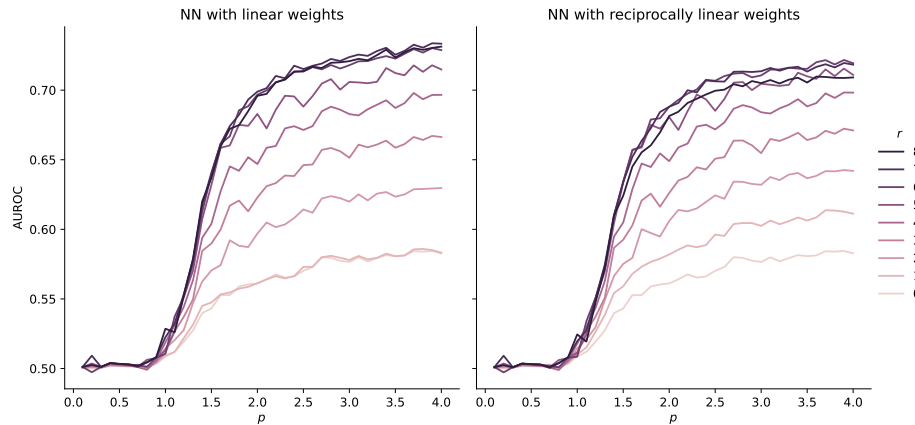
- The choice of weights doesn't appear to play a role in the overall behaviour of these response curves.
- The response curves are substantially smoother for the upper approximation than for the lower approximation and for NN. The mean approximation appears to inherit some of this smoothness from the upper approximation. This qualitative difference is somewhat surprising, but it can perhaps be explained by the fact that for the upper approximation, neighbours are drawn from a uniform concept (each decision class), whereas for the lower approximation and NN, neighbours are drawn from across decision classes.
- The upper approximation is a better classifier (in terms of AUROC) than the lower approximation and NN for the *20-newsgroups* dataset. Given the relatively poor performance of the lower approximation, it is surprising that



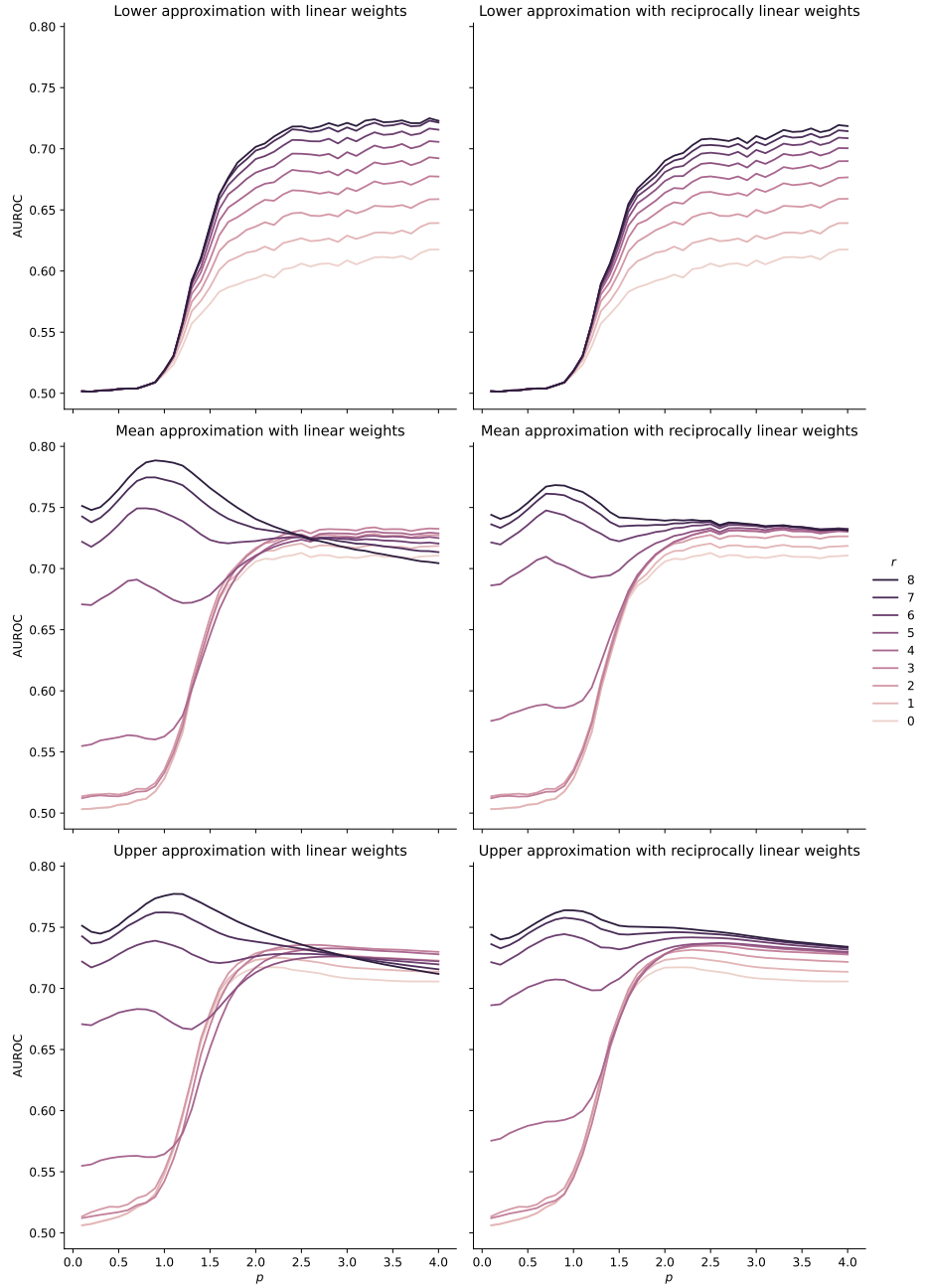
**Fig. 1.** AUROC obtained on the *20-newsgroups* dataset with FRNN, number of neighbours  $k = 256$ , dimensionality  $m = 2^q$  and angular Minkowski  $p$ -distance.



**Fig. 2.** AUROC obtained on the *20-newsgroups* dataset with NN, number of neighbours  $k = 256$ , dimensionality  $m = 2^q$  and angular Minkowski  $p$ -distance.

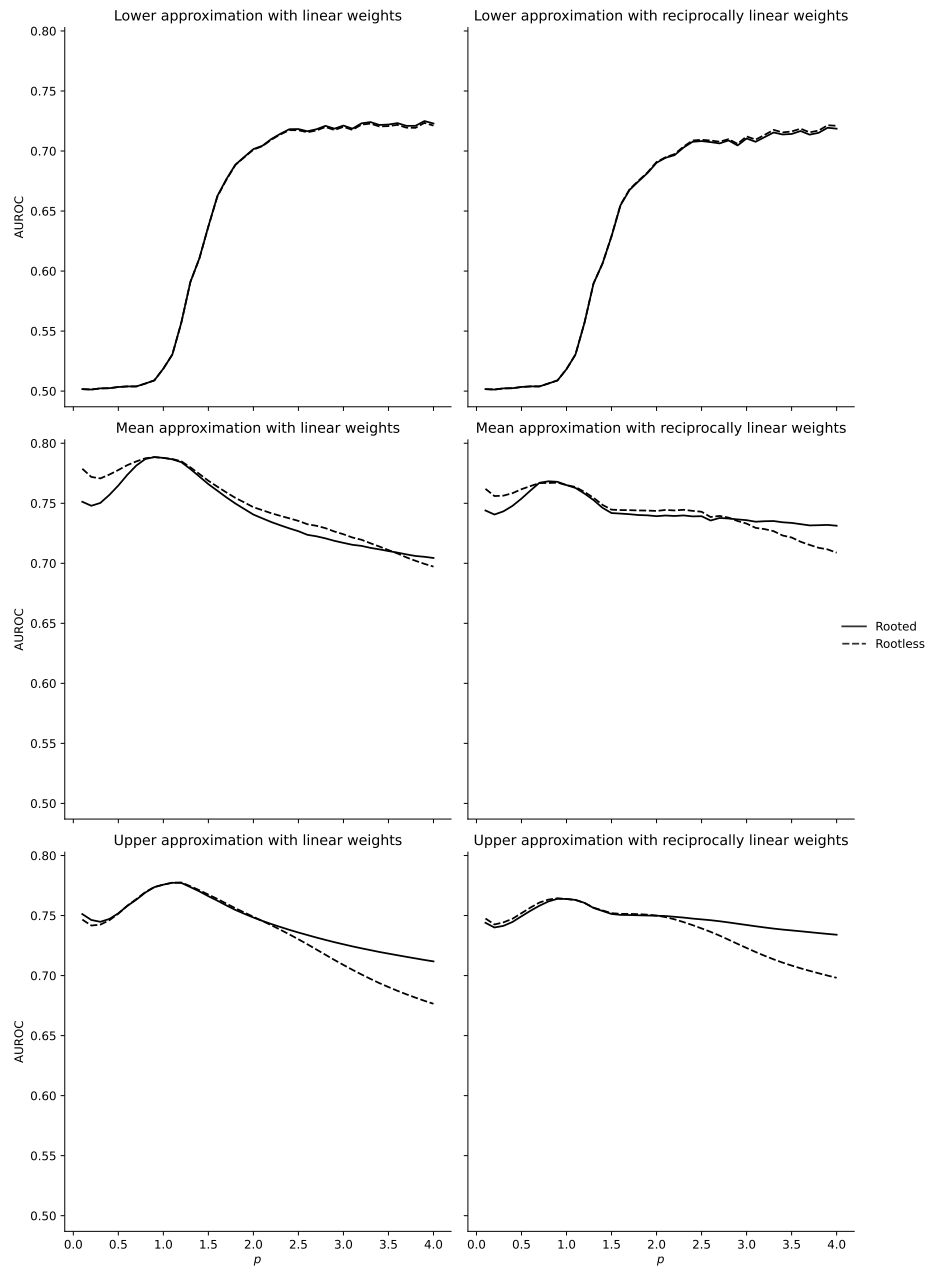


**Fig. 3.** AUROC obtained on the *20-newsgroups* dataset with NN, dimensionality  $m = 4096$ , number of neighbours  $k = 2^r$  and angular Minkowski  $p$ -distance.

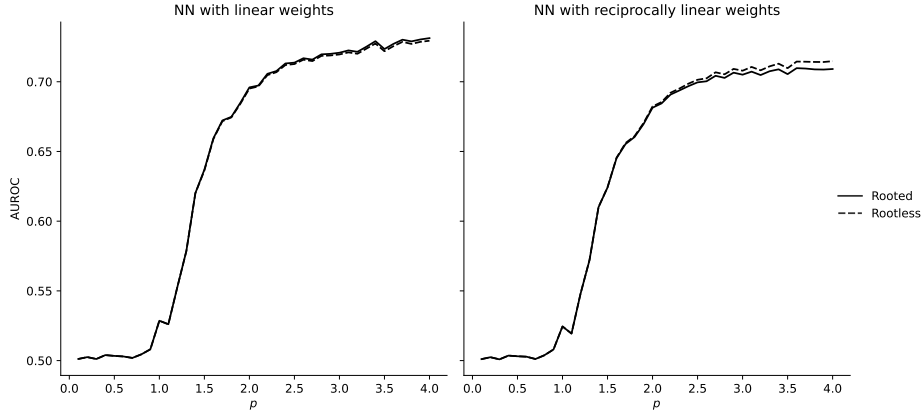


**Fig. 4.** AUROC obtained on the *20-newsgroups* dataset with FRNN, dimensionality  $m = 4096$ , number of neighbours  $k = 2^r$  and angular Minkowski  $p$ -distance.





**Fig. 5.** AUROC obtained on the *20-newsgroups* dataset with FRNN, number of neighbours  $k = 256$ , dimensionality  $m = 4096$  and rooted and rootless angular Minkowski  $p$ -distance.



**Fig. 6.** AUROC obtained on the *20-newsgroups* dataset with NN, number of neighbours  $k = 256$ , dimensionality  $m = 4096$  and rooted and rootless angular Minkowski  $p$ -distance.

the mean approximation produces even better results than the upper approximation.

- AUROC increases with dimensionality, but the difference between 2048 and 4096 dimensions is quite small. It appears that up until that point, the additional information encoded in each additional dimension outweighs the noise. Note, however, that even before that point, we get diminishing returns. For each subsequent curve we need to double the dimensionality, and we obtain a performance increase that is smaller than the previous one.
- For NN and the lower approximation, the choice for  $p$  becomes more important as dimensionality increases. Not only is a good choice for  $p$  necessary to make use of the potential performance increase from adding more dimensions, choosing  $p$  poorly can actually cause performance to decrease with dimensionality.
- There is a marked difference with respect to the optimal values for  $p$  between the different classifiers. For NN and the lower approximation, higher values appear to be better within the range  $[0.1, 4]$  that we have investigated, albeit with diminishing returns. For the upper and mean approximations, the optimum is located near  $p = 1$  for high dimensionalities.

As mentioned above, Figures 1 and 2 reflect a choice of the number of neighbours  $k = 256$ . The effect of  $k$  on performance is illustrated in Figures 3 and 4, for  $m = 4096$ .

- For NN and the lower approximation, the overall behaviour of the response curve does not change with  $k$ . Higher values for  $k$  lead to higher AUROC, and within the range of investigated values, the relationship appears to be similar to the relationship between AUROC and  $m$ : each doubling of  $k$  leads

to an increase in AUROC that is slightly smaller than the previous increase. From  $k = 128$  to  $k = 256$ , the increase is already quite small.

- In contrast, for the upper and mean approximations, AUROC starts out quite high for high values of  $p$ , and increases only little thereafter. However, from  $k = 8$  upwards, AUROC starts to strongly increase for lower values of  $p$ , eventually surpassing the AUROC obtained with higher values of  $p$  from  $k = 64$  upwards. This means that the good performance of the mean and upper approximations around  $p = 1$  is only realised for high values of  $k$ .

Finally, we may also ask whether it makes a difference whether we use rooted (‘ordinary’) or rootless angular Minkowski  $p$ -distance. The results discussed above were obtained using rooted angular Minkowski  $p$ -distance. It turns out that using rootless angular Minkowski  $p$ -distance, which generalises cosine dissimilarity more closely, does not make much difference (Figures 5 and 6). In particular, there is (by definition) no difference for  $p = 1$ , which maximises classification performance for the upper and mean approximations.

**Table 1.** Highest AUROC and corresponding value for  $p$  obtained on the *20-newsgroups* dataset, with linear weights, number of neighbours  $k = 256$ , dimensionality  $m = 4096$  and rooted angular Minkowski  $p$ -distance.

Classifier	$p$	AUROC
NN	4.0	0.731
FRNN (lower approximation)	3.9	0.725
FRNN (mean approximation)	0.9	0.788
FRNN (upper approximation)	1.1	0.777

In summary (Table 1), we obtain the best classification performance on the *20-newsgroups* dataset with the upper and mean approximation and angular Minkowski  $p$ -distance with values of  $p$  around 1, but only when  $k$  is high enough ( $\geq 64$ ).

## 6 Conclusion

We have presented angular Minkowski  $p$ -distance, a generalisation of the popular cosine (dis)similarity measure. In an exploratory case study of the large *20-newsgroups* text dataset, we showed that the choice of  $p$  can have a large effect on classification performance, and in particular that the right choice of  $p$  can increase classification performance over cosine dissimilarity (which corresponds to  $p = 2$ ).

We have also examined the interaction between  $p$  and the dimensionality  $m$  of a dataset, the choice of classification algorithm (NN or FRNN), the choice of weights (linear or reciprocally linear), and the choice of the number of neighbours  $k$ . We found that while the choice of weights was not important, the best

value for  $p$  can depend on  $m$ ,  $k$  and the classification algorithm. Under optimal circumstances (high  $k$  and high  $m$ ), the best-performing values for  $p$  are in the neighbourhood of 1 (FRNN with upper or mean approximation) and around 4 (NN and FRNN with lower approximation).

A major advantage of angular Minkowski  $p$ -distance is that it is defined in terms of ordinary Minkowski  $p$ -distance, which is widely available. Thus, angular Minkowski  $p$ -distance does not require any dedicated implementation and can easily be used in experiments by other researchers.

The most important open question to be investigated in future experiments is to which extent these results generalise to other text datasets, as well as to other datasets containing token frequencies. Depending on the outcome of these experiments, it may be possible to formulate more general conclusions about the best choice for  $p$ , or we may be forced to conclude that this is a hyperparameter that must be optimised for each individual dataset.

**Acknowledgements** The research reported in this paper was conducted with the financial support of the Odysseus programme of the Research Foundation – Flanders (FWO).

## References

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8. pp. 420–434. Springer (2001)
2. Dudani, S.A.: An experimental study of moment methods for automatic identification of three-dimensional objects from television images. Ph.D. thesis, The Ohio State University (1973)
3. Dudani, S.A.: The distance-weighted  $k$ -nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics **6**(4), 325–327 (1976)
4. Fix, E., Hodges, Jr, J.: Discriminatory analysis — nonparametric discrimination: Consistency properties. Tech. Rep. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas (1951), <https://apps.dtic.mil/sti/citations/ADA800276>
5. France, S.L., Carroll, J.D., Xiong, H.: Distance metrics for high dimensional nearest neighborhood recovery: Compression and normalization. Information Sciences **184**(1), 92–110 (2012)
6. Hand, D.J., Till, R.J.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine learning **45**(2), 171–186 (2001)
7. Jensen, R., Cornelis, C.: A new approach to fuzzy-rough nearest neighbour classification. In: RSCTC 2008: Proceedings of the 6th International Conference on Rough Sets and Current Trends in Computing. pp. 310–319. No. 5306 in Lecture Notes in Artificial Intelligence, Springer (2008)
8. Joachims, T.: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. Tech. Rep. CMS-CS-96-118, Carnegie Mellon University, School of Computer Science, Pittsburgh (1996)
9. Kaminska, O., Cornelis, C., Hoste, V.: Fuzzy rough nearest neighbour methods for detecting emotions, hate speech and irony. Information Sciences (2023)

10. Lenz, O.U.: Fuzzy rough nearest neighbour classification on real-life datasets. Doctoral thesis, Universiteit Gent
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**(85), 2825–2830 (2011)
12. Rosner, B.S.: A new scaling technique for absolute judgments. *Psychometrika* **21**(4), 377–381 (1956)
13. Salton, G.: Some experiments in the generation of word and document associations. In: *Proceedings of the 1962 Fall Joint Computer Conference. AFIPS Conference Proceedings*, vol. 22, pp. 234–250. Spartan Books (1962)