p-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2025. № 2
e-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2025. № 2

UDC 004.942, 004.89

A STUDY ON THE USE OF NORMALIZED L_2-METRIC IN CLASSIFICATION TASKS

Kondruk N. E. – PhD, Associate Professor, Associate Professor of Department of Cybernetics and Applied Mathematics, Uzhhorod National University, Uzhhorod, Ukraine.

ABSTRACT

Context. In machine learning, similarity measures, and distance metrics are pivotal in tasks like classification, clustering, and dimensionality reduction. The effectiveness of traditional metrics, such as Euclidean distance, can be limited when applied to complex datasets. The object of the study is the processes of data classification and dimensionality reduction in machine learning tasks, in particular, the use of metric methods to assess the similarity between objects.

Objective. The study aims to evaluate the feasibility and performance of a normalized L_2-metric (Normalized Euclidean Distance, NED) for improving the accuracy of machine learning algorithms, specifically in classification and dimensionality reduction.

Method. We prove mathematically that the normalized L_2-metric satisfies the properties of boundedness, scale invariance, and monotonicity. It is shown that NED can be interpreted as a measure of dissimilarity of feature vectors. Its integration into *k*-nearest neighbors and *t*-SNE algorithms is investigated using a high-dimensional Alzheimer's disease dataset. The study implemented four models combining different approaches to classification and dimensionality reduction. Model M1 utilized the *k*-nearest neighbors method with Euclidean distance without dimensionality reduction, serving as a baseline; Model M2 employed the normalized L_2-metric in *k*NN; Model M3 integrated *t*-SNE for dimensionality reduction followed by kNN based on Euclidean distance; and Model M4 combined *t*-SNE and the normalized L_2-metric for both reduction and classification stages. A hyperparameter optimization procedure was implemented for all models, including the number of neighbors, voting type, and the perplexity parameter for *t*-SNE. Cross-validation was conducted on five folds to evaluate classification quality objectively. Additionally, the impact of data normalization on model accuracy was examined.

Results. Models using NED consistently outperformed models based on Euclidean distance, with the highest classification accuracy of 91.4% achieved when it was used in *t*-SNE and the nearest neighbor method (Model M4). This emphasizes the adaptability of NED to complex data structures and its advantage in preserving key features in high and low-dimensional spaces.

Conclusions. The normalized L_2-metric shows potential as an effective measure of dissimilarity for machine learning tasks. It improves the performance of algorithms while maintaining scalability and robustness, which indicates its suitability for various applications in high-dimensional data contexts.

KEYWORDS: normalized Euclidean distance, machine learning, classification, *t*-SNE, similarity measures, *k*-Nearest Neighbors.

ABBREVIATIONS

*k*NN is the *k*-Nearest Neighbors algorithm;
NED is the Normalized Euclidean Distance;
t-SNE is a *t*-distributed Stochastic Neighbor Embedding.

NOMENCLATURE

$|\cdot|$ is a Euclidean norm;
R^n is a *n*-dimensional vector space;
$\boldsymbol{u}$ is a non-zero *n*-dimensional feature vector;
$\boldsymbol{v}$ is a non-zero *n*-dimensional feature vector;
α is a scalar.

INTRODUCTION

In machine learning and data analysis, one key task is finding and using effective ways to measure the similarity between objects. Distance metrics and similarity measures serve as mathematical tools for this purpose. They are the basis for many algorithms, including classification, clustering, dimensionality reduction, and recommender systems. Traditional metrics, such as Euclidean, cosine, and Manhattan distances, are widely used because of their simplicity and efficiency. However, their use can limit the effectiveness of algorithms in the context of specific data. In particular, standard metrics cannot always adequately assess the similarity between objects when the data have different measurement scales when there is a complex correlation between features, etc.

Accordingly, there is a need to find new, more flexible, and accurate tools that can take into account the specifics of different types of data and tasks and not only more accurately reflect the similarity between objects but also ensure correctness in the context of different metric spaces for some applied tasks. On the other hand, it is necessary to investigate whether these new tools can be easily integrated into existing machine learning algorithms and ensure their efficiency.

The object of the study is the process of data classification and dimensionality reduction in machine learning tasks, in particular, the use of metric methods to assess the similarity between objects.

Any function that satisfies the basic properties of non-negativity and reflexivity can be considered a similarity measure. If it satisfies the triangle inequality, it can be considered a distance metric, so the number of such functions is infinite. There are many well-known distance measures in the literature, which, although they have a common goal, differ significantly in focus and formulation. Choosing the optimal measure for a particular task should consider additional properties that may affect the effectiveness of its application. This leads to the need not only to develop measures but also to study their properties.

DOI 10.15588/1607-3274-2025-2-9

The subject of this study is the normalized L_2-metric as a means of assessing the similarity between objects and its impact on the efficiency of algorithms.

This paper aims to investigate the feasibility of using normalized L_2-metric in classification and reduction algorithms to improve their efficiency.

1 PROBLEM STATEMENT

Let $\boldsymbol{u}$ and $\boldsymbol{v}$ from R^n be vectors, in particular, describing the numerical values of some features of objects.

The normalized L_2-metric between two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ can be defined using the formula (1):

$$\mathrm{NED}(\boldsymbol{u},\boldsymbol{v}) = \frac{|\boldsymbol{u}-\boldsymbol{v}|}{|\boldsymbol{u}|+|\boldsymbol{v}|}. \qquad (1)$$

Here, $|\cdot|$ denotes the Euclidean norm of the vector. This distance finds the ratio of the norm of the difference between the vectors to the sum of their norms, which ensures the normalization of the result. Note that (1) is defined only for non-zero vectors $\boldsymbol{u}$, $\boldsymbol{v}$. In the case of zero vectors, NED can be redefined to be zero.

As is known [1], of all the normalized L_p-distances, only when p=2 is a metric, i.e., it satisfies the triangle inequality. We will study and generalize other mathematical properties of the metric NED and evaluate the effectiveness of its integration into distance-based algorithms: k-nearest neighbors and t-distributed stochastic proximity embedding for high-dimensional data.

2 REVIEW OF THE LITERATURE

Since many real-world problems are based on finding similarities between groups of objects or populations, the list of fields of knowledge that use similarity measures is quite broad: biology, physics, chemistry, geography, ecology, social sciences, anthropology, algebra, statistical mathematics, engineering, and computer science [1].

Distance and similarity measures play a critical role in many machine learning tasks, including case-based reasoning, clustering, and classification [2–6], often impacting model performance more than the choice of algorithm [1]. However, this aspect receives insufficient attention in the literature, as it requires deep knowledge of the subject area and is difficult to generalize.

Modern studies on assessing the predictive capabilities of various similarity metrics for different datasets and conditions were conducted, particularly in [7–11]. Researchers have studied the relationship between commonly used distance measures, their performance in various machine learning tasks, and their robustness to noise [9]. The article [8] investigates the predictive capabilities of various similarity metrics (Euclidean, Pearson's correlation, Spearman's rank correlation coefficient, Kendall's coefficient) based on their application to data sets of different dimensions and properties, as well as the evaluation of the results obtained. Some metrics have been shown to be better suited for large datasets, while others are more reliable when applied to smaller outlier datasets. Data quality, correlation, and data types also play a role. Thus, research in this area emphasizes carefully selecting similarity measures depending on the application and data.

Selecting an effective tool for measuring the similarity between objects is critical in distance-based machine learning algorithms such as the k-nearest neighbors method [10, 11], clustering [4–6], and reduction. For example, in clustering tasks, using different similarity measures allows to obtain more clearly interpreted groups and apply a systematic approach to identifying different types of relationships between data [5, 6].

The effectiveness of various normalized L_1-metrics has been studied in [9–11], but the use of normalized Euclidean distance (1) in machine learning tasks has hardly been investigated. Again, It should be emphasized that among all normalized L_p-metrics, only NED is a distance metric.

This article is devoted to analyzing and summarizing the properties of the normalized L_2-metric (1) and evaluating its effectiveness in distance-based algorithms and high-dimensional data.

3 MATERIALS AND METHODS

Let's explore the properties of the NED metric.

1. Boundedness: $0 \le \mathrm{NED}(\boldsymbol{u}, \boldsymbol{v}) \le 1$.

Proof. The lower bound follows from the fact that NED is a distance metric [1]. Let us prove the upper bound using the triangle property.

For any vectors $\boldsymbol{u}, \boldsymbol{v} \in R^n$, it holds: $|\boldsymbol{u}-\boldsymbol{v}| \le |\boldsymbol{u}| + |\boldsymbol{v}|$.

This inequality reflects that the length of a side of a triangle (the difference of two vectors) is always less than or equal to the sum of the lengths of the other two sides.

$$\mathrm{NED}(\boldsymbol{u},\boldsymbol{v}) = \frac{|\boldsymbol{u}-\boldsymbol{v}|}{|\boldsymbol{u}|+|\boldsymbol{v}|} \le \frac{|\boldsymbol{u}|+|\boldsymbol{v}|}{|\boldsymbol{u}|+|\boldsymbol{v}|} = 1.$$

Proven.

Consequence. If $\boldsymbol{u}$, $\boldsymbol{v}$ are antiparallel, then $\mathrm{NED}(\boldsymbol{u}, \boldsymbol{v}) = 1$.

Proof. Two vectors are called antiparallel if they are collinear and oppositely directed.

So, $\boldsymbol{v} = -\alpha\cdot\boldsymbol{u}$, $\alpha > 0$, then

$$\mathrm{NED}(\boldsymbol{u},\boldsymbol{v}) = \frac{|\boldsymbol{u}-(-\alpha\boldsymbol{u})|}{|\boldsymbol{u}|+|-\alpha\boldsymbol{u}|}.$$

In the numerator:

$$|\boldsymbol{u}-(-\alpha\cdot\boldsymbol{u})| = |\boldsymbol{u}+\alpha\cdot\boldsymbol{u}| = (1+\alpha)\cdot|\boldsymbol{u}|.$$

In the denominator:

$$|\boldsymbol{u}|+|-\alpha\cdot\boldsymbol{u}| = |\boldsymbol{u}|+|\alpha|\cdot|\boldsymbol{u}| = |\boldsymbol{u}|+\alpha\cdot|\boldsymbol{u}| = (1+\alpha)\cdot|\boldsymbol{u}|.$$

So,

DOI 10.15588/1607-3274-2025-2-9

$$\mathrm{NED}(\boldsymbol{u},\boldsymbol{v}) = \frac{1+\alpha}{1+\alpha} = 1.$$

Proven.

Remarks. Based on the boundedness property and the consequence, we can conclude that this distance metric (1) is also a similarity measure. That is, the more similar the feature vectors of objects are, the closer the NED value will be to zero. On the other hand, if the feature vectors are as dissimilar as possible, the closer the NED value will approach one.

Therefore, it is proposed to interpret NED as a measure of dissimilarity.

2. Invariance to scale: NED($\alpha\boldsymbol{u}$, $\alpha\boldsymbol{v}$) = NED($\boldsymbol{u}$, $\boldsymbol{v}$), where α is some number.

Proof.

$$\mathrm{NED}(\alpha\boldsymbol{u},\alpha\boldsymbol{v}) = \frac{|\alpha\boldsymbol{u}-\alpha\boldsymbol{v}|}{|\alpha\boldsymbol{u}|+|\alpha\boldsymbol{v}|} = \frac{|\alpha(\boldsymbol{u}-\boldsymbol{v})|}{|\alpha||\boldsymbol{u}|+|\alpha||\boldsymbol{v}|} =$$
$$= \frac{|\alpha|\|(\boldsymbol{u}-\boldsymbol{v})|}{|\alpha|(|\boldsymbol{u}|+|\boldsymbol{v}|)} = \mathrm{NED}(\boldsymbol{u},\boldsymbol{v}).$$

Proven.

3. The monotonicity property.

If $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{w}$ are non-zero vectors, $|\boldsymbol{w}| \le |\boldsymbol{v}|$ and $|\boldsymbol{u}-\boldsymbol{v}| \le |\boldsymbol{u}-\boldsymbol{w}|$, then NED ($\boldsymbol{u}$, $\boldsymbol{v}$) ≤ NED ($\boldsymbol{u}$, $\boldsymbol{w}$).

Proof. Consider the difference:

$$\mathrm{NED}(\boldsymbol{u},\boldsymbol{v}) - \mathrm{NED}(\boldsymbol{u},\boldsymbol{w}) = \frac{|\boldsymbol{u}-\boldsymbol{v}|}{|\boldsymbol{u}|+|\boldsymbol{v}|} - \frac{|\boldsymbol{u}-\boldsymbol{w}|}{|\boldsymbol{u}|+|\boldsymbol{w}|} =$$
$$= \frac{|\boldsymbol{u}-\boldsymbol{v}|\cdot(|\boldsymbol{u}|+|\boldsymbol{w}|) - |\boldsymbol{u}-\boldsymbol{w}|\cdot(|\boldsymbol{u}|+|\boldsymbol{v}|)}{(|\boldsymbol{u}|+|\boldsymbol{v}|)\cdot(|\boldsymbol{u}|+|\boldsymbol{w}|)} \le$$
$$\le \frac{|\boldsymbol{u}-\boldsymbol{w}|\cdot(|\boldsymbol{u}|+|\boldsymbol{w}|) - |\boldsymbol{u}-\boldsymbol{w}|\cdot(|\boldsymbol{u}|+|\boldsymbol{v}|)}{(|\boldsymbol{u}|+|\boldsymbol{v}|)\cdot(|\boldsymbol{u}|+|\boldsymbol{w}|)} =$$
$$= \frac{|\boldsymbol{u}-\boldsymbol{w}|\cdot(|\boldsymbol{u}|+|\boldsymbol{w}|-|\boldsymbol{u}|-|\boldsymbol{v}|)}{(|\boldsymbol{u}|+|\boldsymbol{v}|)\cdot(|\boldsymbol{u}|+|\boldsymbol{w}|)} =$$
$$= \frac{|\boldsymbol{u}-\boldsymbol{w}|\cdot(|\boldsymbol{w}|-|\boldsymbol{v}|)}{(|\boldsymbol{u}|+|\boldsymbol{v}|)\cdot(|\boldsymbol{u}|+|\boldsymbol{w}|)} \le 0.$$

Given that all Euclidean norms are positive values and $|\boldsymbol{w}| \le |\boldsymbol{v}|$, the difference in the numerator will take on non-positive values. Consequently, the expression will also be non-positive.

Proven.

4 EXPERIMENTS

The DARWIN dataset [12], which contains data on Alzheimer's disease based on handwriting analysis (452 features about 174 respondents), was selected for the experiments. As a result of data pre-processing, 450 predictor features and one target feature, which is binary, were retained. This made it possible to train models in high-dimensional data. The *k*-nearest neighbors algorithm and the Euclidean distance metric were used as a benchmark to evaluate the effectiveness of the dissimilarity measure (1). The index for assessing the quality of classification is a metric that determines the proportion of correct predictions of the model, i.e., the ratio of the number of correct predictions to the total number of observations.

Next, a procedure was implemented to find the optimal hyperparameters of the kNN algorithm adapted to use the Euclidean metric and the dissimilarity measure (1). The classifiers were trained on unstandardized and standardized data using standard normalization and different approaches to neighbor voting.

The approach of reducing the high-dimensional data space to two dimensions based on the *t*-SNE method was also used. The perplexity parameter was varied in the range from 5 to 40 in increments of 5 to study the effect of this parameter on classification accuracy.

We optimized the kNN hyperparameters using a grid search, which included the number of neighbors in the range from 2 to 15 and the type of voting – simple (uniform) and weighted (distance). During each iteration, the value of the *t*-SNE perplexity parameter was updated, and the classifier was optimized for the reduced data. The dimensionality of the feature space was reduced using Euclidean distance and NED.

Cross-validation was used to evaluate the classification quality on five blocks, corresponding to 20% of all data, to form the test sample. The best results of classification accuracy were recorded, and for each combination of parameters, it was determined whether they outperformed the previous results. As a result, the optimal hyperparameters of the models were determined, including the *t*-SNE perplexity parameters, the number of neighbors, the type of voting, and the accuracy achieved for these settings.

5 RESULTS

Table 1 show the results of the experiments for non-standardized and standardized data: the optimal hyperparameters of the four models, including the value of the *t*-SNE perplexity parameter, the number of kNN neighbors, the type of voting, and the achieved accuracy of the classifiers.

Fig. 1 graphically illustrates the comparison of the dependence of the accuracy of classifiers built based on M1–M4 models under optimally tuned hyperparameters (Table 1) and the number of kNN neighbors with and without using data standardization approaches.

In Fig. 1, the solid line represents the performance of models based on the NED dissimilarity measure (M2, M4), while the dashed line represents models based on Euclidean distance (M1, M3). Bold points indicate where the highest accuracy is achieved. Fig. 1a and 1c illustrate the comparison of model performance for non-standardized data, while Fig. 1b and 1d show the performance for standardized data. Fig. 1a and 1b show the accuracies of the M1 and M2 models, while Fig. 1c and 1d illustrate the accuracies of the M3 and M4 models.

DOI 10.15588/1607-3274-2025-2-9

Table 1 – Experimental results on optimal hyperparameters and model accuracy

Model		M1: model without *t*-SNE reduction; the distance metric in *k*NN is Euclidean.	M2: model without *t*-SNE reduction; the distance metric in kNN is NED.	M3: model with *t*-SNE using Euclidean distance; the distance metric in kNN is Euclidean.	M4: model with *t*-SNE using NED distance; the distance metric in *k*NN is NED.
Non-standardized data	Optimal Hyperparameters	Number of neighbors – 4, weighted voting.	Number of neighbors – 4, weighted voting.	Perplexity – 30, number of neighbors – 4, weighted voting.	Perplexity – 10, number of neighbors – 10, weighted voting.
	Accuracy	75%	81 %	80 %	87 %
Standardized data	Optimal Hyperparameters	Number of neighbors – 2, weighted voting.	Number of neighbors – 6, weighted voting.	Perplexity – 25, number of neighbors – 3, simple voting.	Perplexity – 25, number of neighbors – 9, weighted voting.
	Accuracy	73 %	89 %	87 %	91.4 %

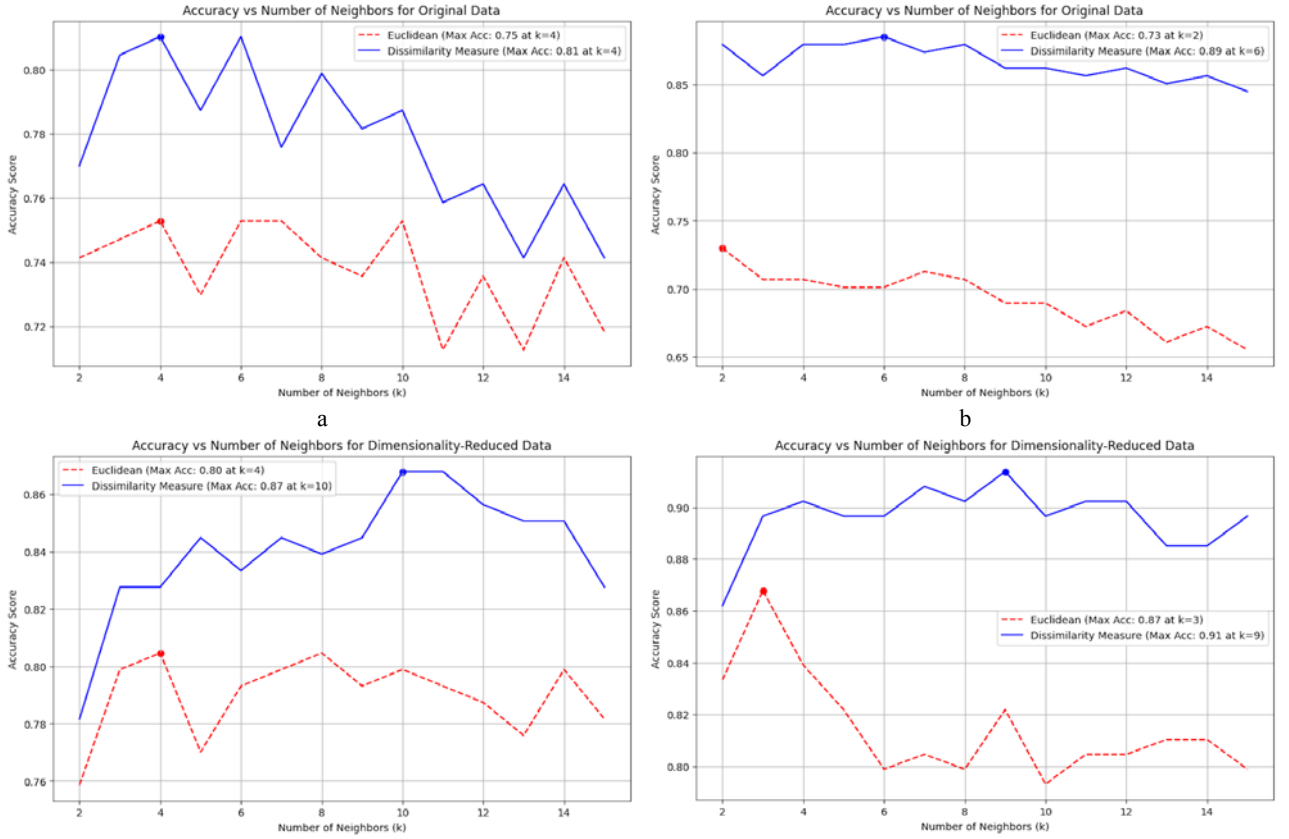

Figure 1 – Comparison of predictive accuracy of the models M1–M4

6 DISCUSSION

The results presented in Table 1 demonstrate the effectiveness of using the *t*-SNE dimensionality reduction technique to enhance model kNN performance. Specifically, models based on Euclidean distance (M1, M3) achieved a 5% improvement in accuracy for non-standardized data and a 14% improvement for standardized data. For models using the NED dissimilarity measure (M2, M4), performance gains were 6% for non-standardized data and 2% for standardized data.

When comparing models based on NED (M2, M4) and Euclidean distance (M1, M3), the NED-based models consistently showed higher accuracy across various values of the nearest neighbors parameter (Fig. 1). For non-standardized data, the highest accuracy of 87% was achieved with the NED measure, exceeding the corresponding result of 80% for the Euclidean metric by 7%. Similarly, for standardized data, the use of NED improved accuracy by 4%, with the best performance reaching 91.4% (model M4). These findings highlight the competitiveness of the proposed dissimilarity measure (1), which demonstrates greater adaptability to complex data structures compared to the Euclidean metric.

The experiments combining *t*-SNE for dimensionality reduction and *k*NN for classification emphasize the significant influence of the choice of distance measure and model parameters on classification outcomes. The NED measure (1) proved particularly effective when applied in combination with *t*-SNE, where it was also used to compute distances between points during data compression. This approach, implemented in model M4, outperformed all other configurations, achieving an accuracy of 91.4%.

DOI 10.15588/1607-3274-2025-2-9

This result indicates that the proposed measure (1) can better capture data structures in high-dimensional spaces and after dimensionality reduction.

As shown in Section 3 of this article, the normalized Euclidean distance (NED) metric simultaneously acts as a distance metric and a similarity measure and satisfies the important properties for algorithms: scale invariance, monotonicity, and boundedness. It has been experimentally shown that the use of NED consistently allows models to achieve higher accuracy than the traditional Euclidean metric. These characteristics emphasize the adaptability of NED to work with complex datasets and its ability to preserve key properties even in spaces with reduced dimensionality.

CONCLUSIONS

The problem of developing a mathematical framework to enhance the efficiency of existing machine learning algorithms is addressed. One of the central elements of distance-based methods is the approach to distance measurement.

The scientific novelty of the results lies in demonstrating that the normalized L_2 distance metric (1) can be interpreted as a measure of dissimilarity between feature vectors, making it valuable for comparing relative differences between vectors. It is proven that the NED satisfies the properties of boundedness, monotonicity, and scale invariance.

The practical significance of the results is reflected in the developed software that implements the dissimilarity measure (1) within the *k*-nearest neighbors method and *t*-SNE feature space reduction. A comparative analysis of the accuracy of four models was conducted using applied high-dimensional data. The highest accuracy achieved through cross-validation was 91.4%, obtained by Model M4, which integrates NED into the classifier and feature space reduction. It is worth noting that the experiments demonstrated the ability of measure (1) to significantly improve the efficiency of distance-based algorithms in solving specific classes of applied classification problems under standardized and non-standardized, high-dimensional, and reduced data conditions.

Future research prospects include studying the efficiency of applying the normalized Euclidean metric to other applied problems.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of Uzhhorod National University, “Computational Intelligence Methods for Data Processing and Analysis” (state registration number 0121U109279).

REFERENCES

1. Deza M. M., Deza E. Encyclopedia of Distances, *Encyclopedia of Distances*. Berlin, Heidelberg, Springer, 2009. DOI: 10.1007/978-3-642-00234-2_1.
2. Mathisen B. M., Aamodt A., Bach K., Langseth H. Learning similarity measures from data, *Progress in Artificial Intelligence*, 2019, Vol. 9, pp. 129–143. DOI: 10.1007/s13748-019-00201-2.
3. Vangipuram S. K., Appusamy R. A survey on similarity measures and machine learning algorithms for classification and prediction, *International Conference on Data Science, E-learning and Information Systems 2021 (DATA'21): Proceedings. – Association for Computing Machinery*, 2021, pp. 198–204. DOI: 10.1145/3460620.3460755.
4. Kondruk N. E. Methods for determining similarity of categorical ordered data, *Radio Electronics, Computer Science, Control,* 2023, Vol. 65, No. 2, pp. 31–36. DOI: 10.15588/1607-3274-2023-2-4.
5. Kondruk N. E. Use of length-based similarity measure in clustering problems, *Radio Electronics, Computer Science, Control,* 2018, Vol. 46, No. 3, pp. 98–105. DOI: 10.15588/1607-3274-2018-3-11.
6. Kondruk N. E., Malyar M. M. Analysis of Cluster Structures by Different Similarity Measures, *Cybernetics and Systems Analysis,* 2021, Vol. 57, pp. 436–441. DOI: 10.1007/s10559-021-00368-4.
7. Vital A., Amancio D. R. A comparative analysis of local similarity metrics and machine learning approaches: application to link prediction in author citation networks, *Scientometrics,* 2022, Vol. 127, pp. 6011–6028. DOI: 10.1007/s11192-022-04484-6.
8. Radisic I., Lazarevic S., Antović I., Stanojevic V. Evaluation of Predictive Capabilities of Similarity Metrics in Machine Learning, *2020 24th International Conference on Information Technology (IT),* 2020, pp. 1–4. DOI: 10.1109/IT48810.2020.9070437.
9. Blanco-Mallo E., Morán-Fernández L., Remeseiro B., Bolón-Canedo V. Do all roads lead to Rome? Studying distance measures in the context of machine learning, *Pattern Recognition*, 2023, Vol. 141. Article ID: 109646. DOI: 10.1016/j.patcog.2023.109646.
10. Pulungan A. F., Zarlis M., Suwilo S. Analysis of Braycurtis, Canberra and Euclidean Distance in KNN Algorithm, *Sinkron: Jurnal dan Penelitian Teknik Informatika,* 2019, Vol. 4, № 1, pp. 74–77. DOI: 10.33395/sinkron.v4i1.10207.
11. Sandhu G., Singh A., Lamba P. S., Virmani D., Chaudhary G. Modified Euclidean-Canberra blend distance metric for kNN classifier, *Intelligent Decision Technologies,* 2023. Vol. 17, № 2, pp. 527–541. DOI: 10.3233/IDT-220233.
12. Cilia N. D., De Stefano C., Fontanella F., Di Freca A. S. An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis, *Procedia Computer Science,* 2018, Vol. 141, pp. 466–471. DOI: 10.24432/C55D0K.

Received 05.02.2025.
Accepted 20.04.2025.

DOI 10.15588/1607-3274-2025-2-9

УДК 004.942, 004.89

ДОСЛІДЖЕННЯ ВИКОРИСТАННЯ НОРМАЛІЗОВАНОЇ L_2-МЕТРИКИ В ЗАДАЧАХ КЛАСИФІКАЦІЇ

Кондрук Н. Е. – канд. техн. наук, доцент, доцент кафедри кібернетики і прикладної математики Ужгородського національного університету, Ужгород, Україна.

АНОТАЦІЯ

Актуальність. У машинному навчанні міри подібності та метрики відстані відіграють ключову роль у задачах класифікація, кластеризація та зменшення розмірності. Ефективність традиційних метрик, зокрема евклідової відстані, може бути обмеженою при застосуванні до складних наборів даних. Об'єктом дослідження є процеси класифікації та зменшення розмірності у задачах машинного навчання, зокрема використання метричних методів для визначення подібності між об'єктами.

Мета роботи – оцінка доцільності та ефективності нормалізованої L_2-метрики (нормалізованої евклідової метрики, NED) для підвищення точності алгоритмів машинного навчання, зокрема в задачах класифікації та зменшення розмірності.

Метод. Математично доведено, що нормалізована L_2-метрика задовольняє властивості обмеженості, масштабної інваріантності та монотонності. Показано, що NED можна інтерпретувати як міру несхожості векторів ознак. Її інтеграція в алгоритми *k*-найближчих сусідів і *t*-SNE досліджується на основі даних про хворобу Альцгеймера високої розмірності. У дослідженні реалізовано чотири моделі, що поєднують різні підходи до класифікації та зменшення розмірності. Модель M1 використовувала метод *k*-найближчих сусідів з евклідовою відстанню без зменшення розмірності, як базова; модель M2 використовувала нормалізовану L_2-метрику в *k*NN; модель M3 інтегрувала *t*-SNE для зменшення розмірності, а потім *k*NN на основі евклідової відстані; модель M4 поєднувала *t*-SNE і нормалізовану L_2-метрику як для зменшення розмірності, так і класифікації. Для всіх моделей було застосовано процедуру оптимізації гіперпараметрів, включаючи кількість сусідів, тип голосування та параметр перплексії в *t*-SNE. Для об'єктивної оцінки якості класифікації було проведено перехресну перевірку на п'яти фолдах. Крім того, було досліджено вплив нормалізації даних на точність моделі.

Результати. Моделі, що використовували NED, стабільно перевершували моделі на основі евклідової відстані, причому найвища точність класифікації (91,4%) була досягнута при інтегруванні NED у *t*-SNE та методі найближчих сусідів (модель M4). Це підкреслює адаптивність NED до складних структур даних та її перевагу у збереженні ключових ознак як у високорозмірному, так і в низькорозмірному просторах.

Висновки. Нормалізована метрика L_2 демонструє потенціал як ефективна міра несхожості для задач машинного навчання. Вона покращує продуктивність алгоритмів, зберігаючи при цьому масштабованість і надійність, що вказує на її придатність для різних застосувань у контексті даних високої розмірності.

КЛЮЧОВІ СЛОВА: нормалізована евклідова відстань, машинне навчання, класифікація, *t*-SNE, міри подібності, *k*-найближчих сусідів.

ЛІТЕРАТУРА

1. Deza M. M. Encyclopedia of Distances / M. M. Deza, E. Deza // Encyclopedia of Distances. – Berlin, Heidelberg : Springer, 2009. DOI: 10.1007/978-3-642-00234-2_1.
2. Learning similarity measures from data / [B. M. Mathisen, A. Aamodt, K. Bach, H. Langseth] // Progress in Artificial Intelligence. – 2019. – Vol. 9. – P. 129–143. DOI: 10.1007/s13748-019-00201-2.
3. Vangipuram S. K. A survey on similarity measures and machine learning algorithms for classification and prediction / S. K. Vangipuram, R. Appusamy // International Conference on Data Science, E-learning and Information Systems 2021 (DATA'21): Proceedings. – Association for Computing Machinery, 2021. – P. 198–204. DOI: 10.1145/3460620.3460755.
4. Кондрук Н. Е. Способи визначення подібності категоріальних впорядкованих даних / Н. Е. Кондрук // Радіоелектроніка, інформатика, управління. – 2023. – № 2 (65). – С. 31–36. DOI: 10.15588/1607-3274-2023-2-4.
5. Кондрук Н. Е. Використання довжинної міри подібності в задачах кластеризації / Н. Е. Кондрук // Радіоелектроніка, інформатика, управління. – 2018. – № 3 (46). – С. 98–105. DOI: 10.15588/1607-3274-2018-3-11.
6. Kondruk N. E. Analysis of Cluster Structures by Different Similarity Measures / N. E. Kondruk, M. M. Malyar // Cybernetics and Systems Analysis. – 2021. – Vol. 57. – P. 436–441. DOI: 10.1007/s10559-021-00368-4.
7. Vital A. A comparative analysis of local similarity metrics and machine learning approaches: application to link prediction in author citation networks / A. Vital, D. R. Amancio // Scientometrics. – 2022. – Vol. 127. – P. 6011–6028. DOI: 10.1007/s11192-022-04484-6.
8. Evaluation of Predictive Capabilities of Similarity Metrics in Machine Learning / [I. Radisic, S. Lazarevic, I. Antović, V. Stanojevic] // 2020 24th International Conference on Information Technology (IT). – 2020. – P. 1–4. DOI: 10.1109/IT48810.2020.9070437.
9. Do all roads lead to Rome? Studying distance measures in the context of machine learning / [E. Blanco-Mallo, L. Morán-Fernández, B. Remeseiro, V. Bolón-Canedo] // Pattern Recognition. – 2023. – Vol. 141. – Article ID: 109646. DOI: 10.1016/j.patcog.2023.109646.
10. Pulungan A. F. Analysis of Braycurtis, Canberra and Euclidean Distance in KNN Algorithm / A. F. Pulungan, M. Zarlis, S. Suwilo // Sinkron: Jurnal dan Penelitian Teknik Informatika. – 2019. – Vol. 4, № 1. – P. 74–77. DOI: 10.33395/sinkron.v4i1.10207.
11. Modified Euclidean-Canberra blend distance metric for *k*NN classifier / [G. Sandhu, A. Singh, P. S. Lamba, D. Virmani, G. Chaudhary] // Intelligent Decision Technologies. – 2023. – Vol. 17, № 2. – P. 527–541. DOI: 10.3233/IDT-220233.
12. An experimental protocol to support cognitive impairment diagnosis by using handwriting analysis / [N. D. Cilia, C. De Stefano, F. Fontanella, A. S. Di Freca] // Procedia Computer Science. – 2018. – Vol. 141. – P. 466–471. DOI: 10.24432/C55D0K.

DOI 10.15588/1607-3274-2025-2-9