Robust Distance Measures for *k*NN Classification of Cancer Data

Rezvan Ehsani[1,2] and Finn Drabløs[3] iD

[1]Department of Mathematics, University of Zabol, Zabol, Iran. [2]Department of Bioinformatics, University of Zabol, Zabol, Iran. [3]Department of Clinical and Molecular Medicine, NTNU – Norwegian University of Science and Technology, Trondheim, Norway.

ABSTRACT: The *k*-Nearest Neighbor (*k*NN) classifier represents a simple and very general approach to classification. Still, the performance of *k*NN classifiers can often compete with more complex machine-learning algorithms. The core of *k*NN depends on a "guilt by association" principle where classification is performed by measuring the similarity between a query and a set of training patterns, often computed as distances. The relative performance of *k*NN classifiers is closely linked to the choice of distance or similarity measure, and it is therefore relevant to investigate the effect of using different distance measures when comparing biomedical data. In this study on classification of cancer data sets, we have used both common and novel distance measures, including the novel distance measures Sobolev and Fisher, and we have evaluated the performance of *k*NN with these distances on 4 cancer data sets of different type. We find that the performance when using the novel distance measures is comparable to the performance with more well-established measures, in particular for the Sobolev distance. We define a robust ranking of all the distance measures according to overall performance. Several distance measures show robust performance in *k*NN over several data sets, in particular the Hassanat, Sobolev, and Manhattan measures. Some of the other measures show good performance on selected data sets but seem to be more sensitive to the nature of the classification data. It is therefore important to benchmark distance measures on similar data prior to classification to identify the most suitable measure in each case.

KEYWORDS: *k*-nearest neighbors, *k*NN, distance measures, Fisher, Sobolev

Background

Classification and pattern recognition are important challenges in data analysis. The *k*-nearest neighbor (*k*NN) approach was proposed by Fix and Hodges in 1951[1] and later modified by Cover and Hart in 1967.[2] It is a simple, robust and versatile algorithm for classification and regression and has been used for different types of problems such as pattern recognition,[3] ranking of models,[4] text categorization,[5] and object recognition,[6] and in many different areas, including bioinformatics and medicine.[7-9] It is a non-parametric[10] and lazy learning classifier. Being non-parametric makes *k*NN free of assumptions on underlying data properties, so there is no need to have prior knowledge about the data. In lazy learning, any generalization of the training data is postponed until the test data are presented to the system.[11]

The *k*NN algorithm is conceptually simple but can still be used on complex biological data, for example, from cancer. A search in the PubMed database for "*k*-NN OR *k*NN" retrieves more than 2000 hits from 1980 to 2020 (August 2020), and a joint search with "cancer" shows that more than 330 of these hits (16%) are related to using the *k*NN approach in cancer research. The popularity of *k*NN actually seems to be increasing; the largest number of hits for both *k*NN itself and the combination of *k*NN and cancer is found in 2019, and for the combination of *k*NN and cancer more than 60% of the hits are found in 2014 or later.

The *k*NN algorithm depends upon a neighborhood of close (or similar) patterns relative to a query pattern, and an important challenge is to find the best distance or similarity measure.[12] Different distance measures will lead to different shapes that define the neighborhood which directly impacts the performance of the *k*NN classifier, as illustrated in Figure 1. However, most applications of *k*NN seem to rely on a limited set of distance measures like Euclidean or Spearman by default, without considering whether alternative distance measures may lead to improved performance.

Several general benchmarking studies have investigated how the performance of the *k*NN algorithm is affected by the choice of distance measure. Chomboon et al[13] tested the performance of *k*NN with 11 different distance measures including Euclidean, Minkowski, Mahalanobis, Cosine, Manhattan, Chebyshev, Correlation, Hamming, Jaccard, Standardized Euclidean, and Spearman, and they used these distance measures on 8 different binary synthetic data sets. They used cross-validation (70% training and 30% testing) to measure performance and showed that similar accuracy could be obtained using either Manhattan, Minkowski, Chebyshev, Euclidean, Mahalanobis, or Standardized Euclidean, and that these distance measures could outperform several other measures.

Hu et al[14] evaluated distance measures for *k*NN classification using medical data sets. They focused on 3 different types
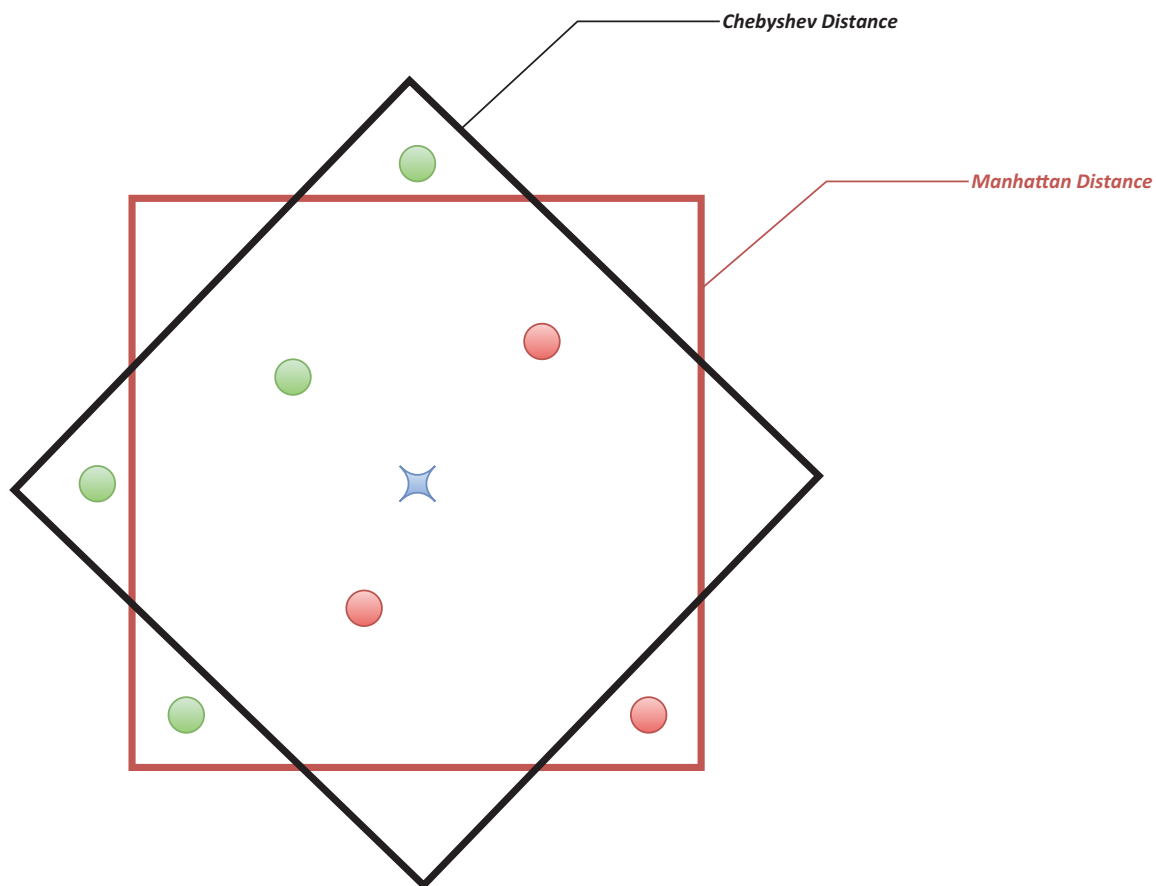
Figure 1. Impact of alternative distance measures on *k*NN performance.
Diamond and square shaped neighborhoods are generated by the Chebyshev and Manhattan distances, respectively. In this case, a new query pattern (blue star) would be classified as either green or red by Chebyshev and Manhattan distances, respectively.

of data, consisting of categorical, numerical, and mixed data types. The sets were from the UCI repository of data sets for machine learning, and they used 4 different distance measures; Euclidean, cosine, chi-square, and Minkowski. They used cross-validation (70% training and 30% testing) to measure the performances, with *k*-values between 1 and 15. The experiments showed the chi-square distance measure to be best for the 3 different data types, whereas the cosine, Euclidean, and Minkowski distances lead to the lowest accuracy on the mixed-type data set.

Punam and Nitin[15] used the KDD data set[16] and the *k*NN classifier with Chebyshev, Euclidean, and Manhattan distance measures. The KDD data set contains numeric data for 41 features in 2 classes. They estimated accuracy, sensitivity, and specificity to evaluate the performance of *k*NN for each distance. The Manhattan distance outperformed the other distances, with 97.8% accuracy, 96.76% sensitivity and 98.35% specificity.

Todeschini et al[17,18] investigated the *k*NN classifier on 8 benchmark data sets with 18 different distance measures, including Manhattan, Euclidean, Soergel, Lance-Williams, contracted Jaccard-Tanimoto, Bhattacharyya, Lagrange, Mahalanobis, Canberra, Wave-Edge, Clark, Cosine, Correlation, and 4 locally centered Mahalanobis distances. The rate of

non-errors and average rank for each distance was determined to evaluate the efficiency of the measure. The results indicated that the highest accuracy was achieved for the Manhattan, Euclidean, Soergel, contracted Jaccard-Tanimoto, and Lance-Williams distance measures.

In a comprehensive review study Prasath and colleagues[19] investigated the impact of 54 different distance measures on 28 various data sets that were obtained from the UCI machine-learning repository. On most data sets, their work showed the best performance by using the Hassanat distance, compared to the other distances.

In summary, these benchmarking studies (and others) have shown that no distance metric is optimal for all data types. Each data type may require a different distance metric for optimal performance in *k*NN, which is consistent with the principle of "no free lunch." This makes it relevant to ask how we can guide users with respect to the choice of distance metrics for *k*NN classification of complex data sets to achieve optimal performance. Here, we have tried to answer that question by identifying metrics with relatively consistent performance across a range of complex data sets, using a selection of both common and more novel metrics.

Specifically, we have investigated the performance of *k*NN classification with 12 different distance metrics, including 8

Table 1. Summary information for the cancer data sets that were used in this study.

DATA SET	NO. OF INSTANCES	NO. OF CLASSES	NO. OF ATTRIBUTES	TYPE OF DATA	REFERENCE
Brain cancer	3064	3	64*64 matrix	MRI image	[20]
Breast cancer	699	2	9	Float	[21]
Lung cancer	32	2	55	Integer	[21]
Prostate cancer	97	2	9	Float	[22]

common and well-known metrics (Euclidean, Manhattan, Canberra, Chebyshev, Bray-Curtis, Clark, Hamming, Bhattacharyya), 2 more novel metrics (Hassanat and Soergel), and 2 new metrics presented by us (Sobolev and Fisher). We have tested these metrics on 4 different data sets on cancer; for breast cancer (cytology), brain cancer (imaging), lung cancer (multivariate), and prostate cancer (clinical). We have evaluated the overall performance of each metric by ranking the metrics according to classification performance across these data sets.

Method

Data sets

The experiments were done on 4 cancer data sets, for brain, lung, breast, and prostate cancer (see Table 1).

For brain cancer, we used a data set consisting of 2-dimensional (2D) slices of CE-MRI images for 3 types of tumors; glioma, meningioma, and pituitary tumor. Data for 233 patients with a total of 3,064 images (axial, coronal, and sagittal views) were available. The original size of each image was 512×512 pixels, which has been decreased to 64×64 to make the calculation faster. The breast and lung cancer data sets were benchmark data sets obtained from the UCI Machine-Learning Repository. The Wisconsin Breast Cancer Data set (WBCD) has 699 instances with 9 attributes for cytology data on 2 types of tumors (i.e. malignant and benign). The lung cancer data set is a multivariate data set with 55 attributes for 32 instances. The prostate cancer data set is a data frame with 97 rows and 9 features with data from a study examining the correlation between the level of prostate-specific antigen and several clinical parameters, using data from participants about to receive a radical prostatectomy.

Distance measures

Here, we give mathematical formulas for distance measures estimating the closeness between 2 vectors x and y, with $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ having numerical attributes. The $d_m(x, y)$ is the distance between x and y as measured by m. Formulations and terminologies are mainly taken from Abu Alfeilat et al,[19] with additional definitions as specified.

Minkowski, Euclidean, Manhattan, and Chebyshev distance. This family of distances is defined as:

$$d_{Minkowski} = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p} \tag{1}$$

where p is a positive value. It is the Manhattan distance when $p = 1$, and the Euclidean distance when $p = 2$, whereas the Chebyshev distance is a variant of Minkowski distance where $p = \infty$. This is also known as maximum value distance,[23] Lagrange,[17] and chessboard distance,[24] and can be formulated as:

$$d_{Chebyshev} = \max_i |x_i - y_i| \tag{2}$$

Canberra distance. This weighted version of the Manhattan distance was introduced and later modified by Lance and Williams.[25,26]

$$d_{Canberra} = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{3}$$

Hamming distance. This distance is based on the number of differences between 2 vectors.[27] It is mainly used to analyze nominal data but can also be used for numerical data.

$$d_{Hamming} = \sum_{i=1}^{n} 1_{x_i \neq y_i} \tag{4}$$

Bhattacharyya distance. This distance represents the similarity of 2 probability distributions.[28]

$$d_{Bhattacharyya} = -ln \sum_{i=1}^{n} \sqrt{x_i y_i} \tag{5}$$

Sorensen distance. This distance is often used to describe relationships in areas like ecology and environmental sciences,[29] and it is also known as Bray-Curtis. It is a modified Manhattan distance, where the total sum of the values is used to standardize the difference over the vectors x and y.[30] It will be between 0 and 1 when all values of the vectors are positive.

$$d_{Bray-Curtis} = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} (x_i + y_i)} \tag{6}$$

Clark distance. This distance[31] is also known as the coefficient of divergence and is the square root of half the divergence distance.

$$d_{\text{Clark}} = \sqrt{\sum_{i=1}^{n}\left(\frac{x_i - y_i}{|x_i| + |y_i|}\right)^2} \qquad (7)$$

Soergel distance. This distance (also known as the Ruzicka distance) is widely used for calculating evolutionary distances.[32] It is identical to the complement of the Jaccard or Tanimoto similarity coefficient for binary variables,[32] and it is in accordance with all 4 metric properties provided that all the attributes have non-negative values.[33]

$$d_{\text{Soergel}} = \frac{\sum_{i=1}^{n}|x_i - y_i|}{\sum_{i=1}^{n}\max(x_i, y_i)} \qquad (8)$$

Hassanat distance. This non-convex distance was introduced by Hassanat.[34]

$$d_{Hassanat} = \sum_{i=1}^{n} D(x_i, y_i)$$

where

$$D(x_i, y_i) = \begin{cases} 1 - \dfrac{1 + \min(x_i, y_i)}{1 + \max(x_i, y_i)}, & \min(x_i, y_i) \geq 0 \\[2em] 1 - \dfrac{1 + \min(x_i, y_i) + |\min(x_i, y_i)|}{1 + \max(x_i, y_i) + |\min(x_i, y_i)|}, & \min(x_i, y_i) < 0 \end{cases} \qquad (9)$$

Sobolev distance. Definitions and notations for this distance are as given by Villmann.[35] Starting with the standard *p*-inner product

$$\langle x, y \rangle_p = \left(\sum_{i=1}^{n}|x_i \cdot y_i|^p\right)^{\frac{1}{p}} \qquad (10)$$

the Sobolev inner product, norm, and metric of degree *k* can be defined as follows:

$$\langle x, y \rangle_{p,\alpha}^S = \langle x, y \rangle_p + \alpha \langle D^k x, D^k y \rangle_p \qquad (11)$$

$$\| x \|_{p,k,\alpha}^S = \sqrt{\langle x, x \rangle_{p,\alpha}^S} \qquad (12)$$

$$d_{p,k,\alpha}^S(x, y) = \| x - y \|_{p,k,\alpha}^S \qquad (13)$$

where D^k is the *k*th differential operator. There is a connection to the Fourier transform for the special case $p = 2$ and $\alpha = 1$. Let $\hat{x}$ be the Fourier transform *y*

$$\hat{x}(\omega_k) = \sum_{j=1}^{N-1} y_j \exp\left(-i\frac{2\pi kj}{N}\right) \qquad (14)$$

where $\omega_k = 2\pi k / N$ and $i = \sqrt{-1}$. The norm can be defined as

$$\|x\|_{2,k,1}^S = \sqrt{\sum_{j=1}^{N-1}(1 + \omega_j)^k |\hat{x}(\omega_j)|^2} \qquad (15)$$

Here, we have used metric (13) with norm (15) and *k = 1*.

Fisher distance. Definitions and notations are as given by Lebanon.[36] We first define the n-simplex P_n.

$$P_n = \left\{x \in R^{n+1} : \forall i, x_n \geq 0, \sum_{i=1}^{n+1} x_i = 1\right\} \qquad (16)$$

The sequence $\{x_i\}$ is the probability of different outputs in each experiment. The Fisher information metric on P_n can be defined by

$$J_{ij} = \sum_{k=1}^{n+1} \frac{1}{x_k} \frac{\partial x_k}{\partial x_i} \frac{\partial x_k}{\partial x_j} \qquad (17)$$

The Fisher information is defined as a pull-back metric from the positive *n*-sphere S_n^+;

$$S_n^+ = \left\{x \in R^n; \forall i, x_n \geq 0, \sum_{i=1}^{n+1} x^2 = 1\right\} \qquad (18)$$

The transformation $T : P_n \to S_n^+$ defined by

$$T(x) = \left(\sqrt{x_1}, \ldots, \sqrt{x_{n+1}}\right) \qquad (19)$$

pulls back the Euclidean metric on the surface of the sphere to the Fisher information on P_n. Now Fisher metric for $x, y \in P_n$ can be defined as the length of the great circle (geodesic) between $T(x)$ and $T(y)$ on S_n^+.

$$d(x, y) = \text{acos}\left(\sum_{i=1}^{n+1} \sqrt{x_i y_i}\right) \qquad (20)$$

Performance measures

We used 4 complementary measures for evaluating the performance of each classifier; accuracy, precision, recall, and F_1. These measures can be computed from the following classifications results, where a subset of patterns (the positive set)

belongs to a specific class, whereas the remaining patterns (the negative set) do not belong to this class:

- True positive (TP): The number of patterns of the positive set that is correctly classified as belonging to the positive set.
- True negative (TN): The number of patterns outside of the positive set that are correctly classified as not belonging to the positive set.
- False positive (FP): The number of patterns of the negative set that is incorrectly classified as belonging to the positive set.
- False negative (FN): The number of patterns of the positive set that is incorrectly classified as not belonging to the positive set.

The relevant performance measures can then be defined as:

$$precision = \frac{TP}{TP + FP} \qquad (21)$$

$$recall = \frac{TP}{TP + FN} \qquad (22)$$

$$F_1 = 2\frac{precision \times recall}{precision + recall} \qquad (23)$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (24)$$

Ranking of distance measures

For each distance and performance score, we considered the best (maximum) score among scores across all different $k \in K$ values as the final score. If S_{dpe}^k is the score of distance d for performance p and experiment e, the final score can be defined as:

$$S_{dpe} = \max_k S_{dpe}^k \qquad (25)$$

We then ranked distances according the final score for each individual experiment, using 2 different approaches. The first approach was simply to compute the average of the ranks across all experiments. That is, for a given experiment e and a given performance measure p, the score S was computed for each distance metric d, and the distance metrics were ranked according to the score. This was repeated for each combination of e and p, giving $e \times p$ rankings in total. The final ranking was then estimated as the average ranking of distance metric d over all these $e \times p$ rankings. For the second approach, we used RankAggreg tool,[37] an *R* package for weighted rank aggregation, and we used it on the complete set of ranked lists as described above, using the Cross Entropy Monte Carlo (CE) method, Kendall distances, and a value of rho as 0.1 (please see the RankAggreg documentation).

Table 2. Best scores among all tested *k*-values for the brain cancer data set.

DISTANCE	PRECISION	RECALL	F_1	ACCURACY
Fisher	0.442	0.387	0.378	0.487
Sobolev	0.462	0.460	0.441	0.526
Clark	0.435	0.414	0.404	0.493
Bhattacharyya	0.434	0.398	0.394	0.492
Soergel	0.450	0.444	0.428	0.518
Hassanat	0.462	0.460	0.443	**0.529**
Euclidean	0.461	0.455	0.432	0.522
Manhattan	0.458	**0.461**	0.440	0.524
Chebyshev	0.452	0.455	0.438	0.521
Hamming	0.445	**0.461**	0.432	0.520
Canberra	**0.466**	0.457	**0.448**	0.527
Bray-Curtis	0.450	0.445	0.428	0.518

Maximum scores for each performance measure are shown in bold.

In addition to the ranking, we used the *k*-means algorithm to cluster the distance measures based on the scores over all experiments, and plotted this using the factoextra[38] tool in *R*. This highlights in a visual way the similarities and differences between the tested distance measures.

Software implementation

The Python programming language (version 3.7.1) was used for scripts, which were implemented under Anaconda3. We used libraries from the scikit-learn package (version 0.20.1) to apply the *k*NN algorithm for Euclidean, Manhattan, Chebyshev, Hamming, Canberra, and Bray-Curtis distances.

Results

We applied all 12 distance measures on the 4 cancer data sets. For the brain, breast, and prostate cancer data sets, we used ranges from 1 up to 20 for *k*. For the lung cancer data, the range of *k* was limited to values from 1 up to 11, due to more limited data.

The best scores for the brain cancer data are shown in Table 2. The best precision score is for Canberra followed by Sobolev and Hassanat. For recall the maximum is shared between Manhattan and Hamming. Second and third places are for Sobolev and Hassanat. The best performances based on F_1 and accuracy were for Canberra and Hassanat, respectively.

The scores for the breast cancer data are shown in Table 3. The Clark distance achieved the best score for 3 performance measures: recall, F_1, and accuracy. The best precision was for the Bray-Curtis distance.

For the lung cancer data, the Sobolev distance outperformed the other distances, as it had the best performance according to

precision, F_1, and accuracy. The second rank was for Fisher distance, which achieved the best score for recall and shared F_1 with Sobolev.

Finally, for the prostate cancer data, the Canberra distance clearly outperformed the other distances according to all performance measures.

To have a total and robust ranking scale we used 2 approaches as described under Methods: a basic average of ranks for each distance measure estimated over 16 different rankings (i.e. all possible combinations of data set and performance measure), and the weighted rank aggregation of these rankings by the RankAggreg tool.

To compare the ranking of these 2 approaches, we plotted the 2 rankings, as shown in Figure 2. This shows a good correlation between these rankings, indicating that the overall ranking of the distance measures is robust.

The result of the k-means clustering on the performance scores over all experiments for $k = 3$ are shown in Figure 3. The set of (Hassanat, Canberra, Sobolev, Manhattan, Euclidean, Soergel, Bray-Curtis) forms a relatively tight cluster, whereas the 2 additional clusters are (Hamming, Chebyshev, Clark) and (Bhattacharyya, Fisher). This is quite consistent with the ranking in Figure 2, where the main cluster is seen to consist of the measures with the best overall performance. A clustering with $k = 4$ splits the main cluster into 2 subclusters consisting of (Sobolev, Manhattan, Euclidean) and (Hassanat, Canberra, Soergel, Bray-Curtis), but the general clustering is stable. In summary, the k-means clustering confirms the ranking of the performance data shown in Figure 2.

Discussion

The results presented here show clear differences between distance measures with respect to classification performance on the cancer data sets. Some distance measures have a quite robust performance across most data sets, whereas other measures show a clearly lower performance on some data sets. This seems to be largely independent of which performance measures that are used (precision, recall, F_1 or accuracy), which seems to be confirmed by the loading plot of a principal component analysis (PCA) of the performance data from Tables 2 to 5 (Supplemental Figure S2 in Additional file 1). The plot shows very similar loadings for all performance measures for each data set, in particular for the data on breast cancer and lung cancer.

The individual classification results in Tables 2 to 6 show important differences (and similarities) between the distance measures, depending on data type. If we focus on the F_1 performance measure, we see that both Fisher and Bhattacharyya seem to have relatively low performance on brain cancer (Table 2), breast cancer (Table 3), and prostate cancer (Table 5), in addition to Hamming for prostate cancer. This is different for lung cancer (Table 4), where it is Clark and Chebyshev that is associated with

Table 3. Best scores among all tested k-values for the breast cancer data set.

DISTANCE	PRECISION	RECALL	F_1	ACCURACY
Fisher	0.896	0.896	0.892	0.903
Sobolev	0.964	0.958	0.960	0.964
Clark	0.967	**0.972**	**0.969**	**0.971**
Bhattacharyya	0.905	0.895	0.896	0.907
Soergel	0.964	0.966	0.964	0.967
Hassanat	0.963	0.966	0.964	0.967
Euclidean	0.966	0.962	0.963	0.967
Manhattan	0.963	0.954	0.957	0.962
Chebyshev	0.963	0.961	0.960	0.964
Hamming	0.950	0.925	0.934	0.943
Canberra	0.966	0.969	0.967	0.970
Bray-Curtis	**0.969**	0.969	0.968	**0.971**

Maximum scores for each performance measure are shown in bold.

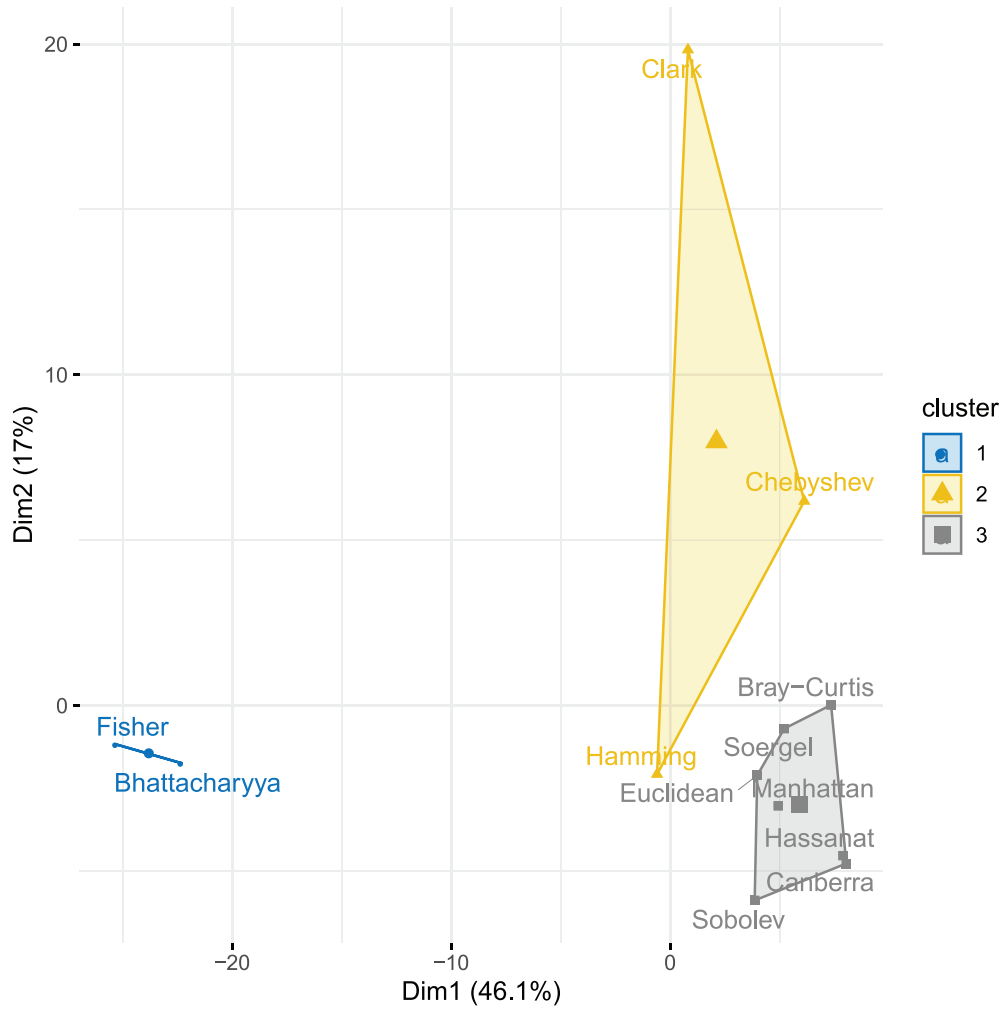**Figure 2.** Comparison of average ranking and RankAggreg ranking.

Figure 3. The *k*-means clustering of all scores over all experiments. Axes represent variance in a PCA plot of the data.

low performance. These differences seem to be confirmed by the *k*-means clustering (Figure 3), where both (Fisher, Bhattacharyya) and (Clark, Chebyshev, Hamming) form separate clusters, and by the PCA analysis, where the loadings for breast cancer data are clearly separated from the other cancer types (Supplemental Figure S2 in Additional file 1). It is also consistent with the ranking data shown in Figure 2, where these same distance measures are ranked together as having low performance.

The ranking of the well-performing measures shows some variation, but this is mainly due to the generally good performance of these measures, with only small (and partly random) differences between cases. However, it is important to realize that the performance of a given distance measure depends on the input data. For example, in the data on lung cancer (Table 4) the Fisher measure shows one of the best performances, whereas it shows low performance on the other data sets. Similarly, the Clark measure is the best-performing measure on breast cancer data (Table 3) but has very low performance on lung cancer data. Apart from intrinsic effects of the type and distribution of data, these differences could arise from the distance functions, which is something that is relevant for further studies.

Table 4. Best scores among all tested *k*-values for the lung cancer data set.

DISTANCE	PRECISION	RECALL	F_1	ACCURACY
Fisher	0.611	**0.656**	**0.602**	0.613
Sobolev	**0.650**	0.633	**0.602**	**0.618**
Clark	0.144	0.356	0.198	0.307
Bhattacharyya	0.553	0.600	0.511	0.545
Soergel	0.550	0.578	0.522	0.545
Hassanat	0.489	0.544	0.464	0.516
Euclidean	0.617	0.633	0.582	0.590
Manhattan	0.618	0.611	0.582	0.585
Chebyshev	0.268	0.389	0.262	0.351
Hamming	0.449	0.500	0.418	0.459
Canberra	0.584	0.544	0.503	0.507
Bray-Curtis	0.550	0.578	0.522	0.545

Maximum scores for each performance measure are shown in bold.

Table 5. Best scores among all tested *k*-values for the prostate cancer data set.

DISTANCE	PRECISION	RECALL	F_1	ACCURACY
Fisher	0.868	0.792	0.769	0.813
Sobolev	0.856	0.832	0.823	0.832
Clark	0.892	0.836	0.840	0.861
Bhattacharyya	0.864	0.783	0.755	0.803
Soergel	0.836	0.820	0.812	0.822
Hassanat	0.864	0.835	0.829	0.841
Euclidean	0.856	0.845	0.834	0.841
Manhattan	0.846	0.828	0.821	0.832
Chebyshev	0.864	0.845	0.834	0.841
Hamming	0.625	0.625	0.594	0.598
Canberra	**0.926**	**0.882**	**0.877**	**0.892**
Bray-Curtis	0.836	0.820	0.812	0.822

Maximum scores for each performance measure are shown in bold.

Table 6. Total ranking of distance measures over all experiments according average rank and rank aggregation by RankAggreg.

DISTANCE	AVERAGE RANK	RANK_AVE	RANKAGGREG
Hassanat	3.50	1	1
Manhattan	4.18	2	4
Sobolev	4.56	3	2
Canberra	4.64	5	6
Bray-Curtis	4.62	4	5
Euclidean	4.81	6	3
Soergel	5.43	7	7
Clark	5.56	8	9
Chebyshev	5.68	9	8
Fisher	5.87	10	10
Bhattacharyya	6.06	11	11
Hamming	6.68	12	12

The analysis presented here may be influenced by the quality of the input data, for example, whether cases in the training set are correctly annotated with respect to class (e.g. cancer versus normal). In principle, we can estimate the quality of training data by looking for consistent misclassifications, experiments where a case consistently is classified to a different class compared to its annotation. Such cases may represent potential annotation errors in the data set and may be considered for removal. However, we should probably expect to have some examples of such cases in most data sets consisting of experimental data. In particular for complex properties like cancer, where it may be difficult to decide unambiguously in each case whether a given sample should represent "cancer" or "normal." In the data presented here, the somewhat lower classification performance on brain cancer data and lung cancer data can possibly be linked partly to misannotated cases. However, such cases will be a natural part of most experimental data and removing them may introduce user bias into the analysis. Also, *k*NN is supposed to be somewhat robust with respect to errors in training data, in particular for higher values of *k*, as the classification will represent an average over multiple cases. Therefore, we have not considered removing such cases from the analysis.

This analysis will also be influenced by the choice of features, for example, if we select only specific features for analysis, compared to the full range of features of a data set. This may for example be relevant if the features represent very different properties. Again, selecting subsets of features may introduce user bias into the analysis. Here, we wanted to test the robustness of the various distance metrics, and therefore, we decided to use all features as given in the original data sets, without any feature selection.

Conclusions

The performance analysis of *k*NN classification of cancer data with different distance measures identifies important differences between both distance measures and data sets. It is possible to identify a subset of distance measures that show robust performance across several data sets, and this includes the Hassanat, Sobolev, and Manhattan measures. However, the study also confirms that no single distance measure will be optimal for all data sets, and the recommendation must be that several measures should be tested on suitable reference data that are as similar to the actual data as possible when selecting distance measure for a particular study.

Author Contributions

RE initiated and designed the project, implemented and performed the *k*NN analysis, and drafted the first version of the manuscript. FD analyzed the data on classification performance and participated in discussion of the results and writing of the final publication.

Availability of Data and Materials

All data used for the analysis are available in public repositories:

- "Brain tumor data set" (https://figshare.com/articles/brain_tumor_dataset/1512427/5).
- "Breast Cancer Wisconsin (Original) Data Set" (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29).
- "Lung Cancer Data Set" (https://archive.ics.uci.edu/ml/datasets/Lung+Cancer).
- Prostate cancer data from the *R* package ElemStatLearn (https://cran.r-project.org/web/packages/ElemStatLearn/index.html)

ORCID iD

Finn Drabløs https://orcid.org/0000-0001-5794-828X

Supplemental Material

Supplementary material for this article is available online.

REFERENCES

1. Silverman BW, Jones MC, Fix E, Hodges JL. An important contribution to nonparametric discriminant analysis and density estimation: commentary on Fix and Hodges (1951). *Int Stat Rev*. 1989;57:233-238.
2. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inform Theory*. 1967;13:21-27.
3. Xu S, Wu Y. An algorithm for remote sensing image classification based on artificial immune B-cell network. In: Jun C, Jie J, Cho K, eds. Xxist ISPRS *Congress, Youth Forum*, Vol. 37. Beijing, China: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; 2008:107-112.
4. Geng X, Liu T-Y, Qin T, Arnold A, Li H, Shum H-Y. Query dependent ranking using K-nearest neighbor. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Singapore)*. New York, NY: Association for Computing Machinery; 2008:115-122.
5. Manne S, Kotha SK, Sameen Fatima S. Text categorization with k-nearest neighbor approach. In: *Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012)*, Visakhapatnam, India; Berlin, Germany; Heidelberg, Germany: Springer; 2012:413-420.
6. Bajramovic F, Mattern F, Butko N, Denzler J. *A Comparison of Nearest Neighbor Search Algorithms for Generic Object Recognition*. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 4179. Berlin, Germany: Springer.
7. Khamis HS, Cheruiyot KW, Kimani S. Application of k-nearest neighbour classification in medical data mining. *Int J Inform Commun Technol Res*. 2014;4:121-128.
8. Kusmirek W, Szmurlo A, Wiewiorka M, Nowak R, Gambin T. Comparison of kNN and k-means optimization methods of reference set selection for improved CNV callers performance. *BMC Bioinform*. 2019;20:266.
9. Roder J, Oliveira C, Net L, Tsypin M, Linstid B, Roder H. A dropout-regularized classifier development approach optimized for precision medicine test discovery from omics data. *BMC Bioinform*. 2019;20:325.
10. Kataria A, Singh MD. A review of data classification using k-nearest neighbour algorithm. *Int J Emerg Technol Adv Eng*. 2013;3:354-360.
11. Wettschereck D, Aha DW, Mohri T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artif Intell Rev*. 1997;11:273-314.
12. Mehta S, Shen X, Gou J, Niu D. A new nearest centroid neighbor classifier based on k local means using harmonic mean distance. *Information*. 2018;9:234.
13. Chomboon K, Chujai P, Teerarassamee P, Kerdprasop K, Kerdprasop N. An empirical study of distance metrics for k-Nearest Neighbor algorithm. In: *The 3rd International Conference on Industrial Application Engineering 2015 (ICIAE2015)*, Kitakyushu, Japan. Japan: The Institute of Industrial Applications Engineers; 2015:280-285.
14. Hu LY, Huang MW, Ke SW, Tsai CF. The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*. 2016;5:1304.
15. Mulak P, Talhar N. Analysis of distance measures using k-nearest neighbor algorithm on KDD dataset. *Int J Sci Res*. 2015;4:2101-2104.
16. Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: *Proceedings of the Second IEEE International Conference on Computational Intelligence for Security and Defense Applications*. Ottawa, ON, Canada: IEEE Press; 2009:53-58.
17. Todeschini R, Ballabio D, Consonni V. Distances and other dissimilarity measures in chemometrics. *Encyclop Anal Chem*. 2015:1-34.
18. Todeschini R, Ballabio D, Consonni V, Grisoni F. A new concept of higher-order similarity and the role of distance/similarity measures in local classification methods. *Chem Intell Lab Syst*. 2016;157:50-57.
19. Abu Alfeilat HA, Hassanat ABA, Lasassmeh O, et al. Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big Data*. 2019;7:221-248.
20. Cheng J. Brain tumor dataset. https://figshare.com/articles/brain_tumor_dataset/1512427/5. Accessed April 27, 2020.
21. Dua D, Graff C. *UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/. Accessed April 27, 2020.
22. Stamey TA, Kabalin JN, McNeal JE, et al. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *J Urol*. 1989;141:1076-1083.
23. Grabusts P. The choice of metrics for clustering algorithms. In: *Proceedings of Environment, Technology, Resources. 8th International Scientific and Practical Conference*, Vol. 2. Rezekne, Latvia: Rezeknes Augstskola; 2011:70-76.
24. Premaratne P. *Human Computer Interaction Using Hand Gestures*. Singapore: Springer; 2014.
25. Lance GN, Williams WT. Computer programs for hierarchical polythetic classification ("similarity analyses"). *Comput J*. 1966;9:60-64.
26. Lance GN, Williams WT. Mixed—data classificatory programs I—agglomerative systems. *Austr Comput J*. 1967;1:15-20.
27. Hamming RW. Error detecting and error correcting codes. *Bell Syst Tech J*. 1950;29:147-160.
28. Bhattacharyya A. On a measure of divergence between two multinomial populations. *Sankhyā*. 1946;7:401-406.
29. Sørensen T. A method of establishing groups of equal amplitudes in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab, Biologiske Skrifter*. 1948;5:1-34.
30. Szmidt E. *Distances and Similarities in Intuitionistic Fuzzy Sets*. Berlin, Germany: Springer; 2013.
31. Clark PJ. An extension of the coefficient of divergence for use with multiple characters. *Copeia*. 1952;1952:61-64.
32. Zhou T, Chan KCC, Wang Z. TopEVM: using co-occurrence and topology patterns of enzymes in metabolic networks to construct phylogenetic trees. In: Chetty M, Ngom A, Ahmad S, eds. Pattern Recognition in Bioinformatics. Berlin, Germany; Heidelberg, Germany: Springer; 2008:225-236.
33. Willett P, Barnard JM, Downs GM. Chemical similarity searching. *J Chem Inform Comput Sci*. 1998;38:983-996.
34. Hassanat AB. Dimensionality invariant similarity measure. *J Am Sci*. 2014;10:221-226.
35. Villmann T. Sobolev metrics for learning of functional data—mathematical and theoretical aspects. In: *Machine Learning Reports*. Leipzig, Germany: Research Group on Computational Intelligence, University of Leipzig; 2007:1-13.
36. Lebanon G. Learning Riemannian metrics. In: *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*. Acapulco, Mexico: Morgan Kaufmann Publishers; 2002:362-369.
37. Pihur V, Datta S, Datta S. RankAggreg. https://cran.r-project.org/package=RankAggreg. Accessed April 27, 2020.
38. factoextra. https://CRAN.R-project.org/package=factoextra. Accessed May 4, 2020.