Do Fractional Norms and Quasinorms Help to Overcome the Curse of Dimensionality?

1rd Evgeny M. Mirkes
Department of mathematics
University of Leicester
Leicester, UK
Lobachevsky State University
Nizhni Novgorod, Russia
em322@leicester.ac.uk

2nd Jeza Allohibi
Department of mathematics
University of Leicester
Leicester, UK
jeza-2011@hotmail.com

3st Alexander N. Gorban
Department of mathematics
University of Leicester
Leicester, UK
Lobachevsky State University
Nizhni Novgorod, Russia
a.n.gorban@leicester.ac.uk

***Abstract*—The curse of dimensionality causes well-known and widely discussed problems for machine learning methods. There is a hypothesis that usage of Manhattan distance and even fractional quasinorms lp (for p less than 1) can help to overcome the curse of dimensionality in classification problems. In this study, we systematically test this hypothesis for 37 binary classification problems on 25 databases. We confirm that fractional quasinorms have greater relative contrast or coefficient of variation than Euclidean norm l2, but we demonstrate also that the distance concentration shows qualitatively the same behaviour for all tested norms and quasinorms and the difference between them decays while dimension tends to infinity. Estimation of classification quality for kNN based on different norms and quasinorms shows that the greater relative contrast does not mean the better classifier performance and the worst performance for different databases was shown by the different norms (quasinorms). A systematic comparison shows that the difference in performance of kNN based on lp for p=2, 1, and 0.5 is statistically insignificant.**

***Index Terms*—curse of dimensionality, blessing of dimensionality, kNN, metrics, high dimension**

I. Introduction

The term "curse of dimensionality" was introduced by Bellman [1]. Nowadays, this is a general term for problems related to high dimensional data, for example, for Bayesian modelling [2], neural networks [3], nearest neighbour prediction [4] and search [5], and many others. Many authors [6], [7], [8], [9], [10] studied the "meaningfulness" of distance based classification in high dimensions. These studies are related to the distance concentration, which means that in high dimensional space the distances between almost all pairs of points have almost the same value.

The term "blessing of dimensionality" was introduced by Kainen in 1997 [11]. The "blessing of dimensionality" considers the same distance concentration effect from the different point of view [12], [13], [14], [15], [16]. The distance concentration was discovered in the foundation of statistical physics and analysed further in the context of probability theory, functional analysis, and geometry (reviewed by [17], [14], [18], [19]). The blessing of dimensionality allows us to use some specific high dimensional properties to solve problems [21], [20]. The important property is linear separability of points from random sets in high dimensions [14], [16].

The project is supported by the Ministry of Education and Science of the Russian Federation (Project No 14.Y26.31.0022).

The l_p functional $\|x\|_p$ in d dimensional vector space is defined as

$$\|x\|_p = \left(\sum_{i=1}^{d} x_i^p\right)^{1/p}. \quad (1)$$

It is a norm for $p \geq 1$ and a quasinorm for $0 < p < 1$ because of violation of the triangle inequality [30]. It is well known that for $p < q$ we have $\|x\|_p \geq \|x\|_q, \forall x$.

Measurement of dissimilarity and errors by subquadratic functionals reduces the effect of outliers and can help to construct more robust data analysis methods [31], [8], [19]. Utilisation of these functionals for struggling with the curse of dimensionality was proposed in several works [8], [19], [32], [33] [34], [35].

In 2001, C.C. Aggarwal with co-authors [8] described briefly the effect of using fractional quasinorms for high-dimensional problems. They demonstrated that using of l_p ($p \leq 1$) can compensate the distance concentration. This idea was used further in many works [22], [23], [10]. One of the main problems of l_p quasinorms usage for $p < 1$ is time of calculation of minimal distances and solution of optimization problems with l_p functional (which is even non-convex for $p < 1$). Several methods have been developed to accelerate the calculations [22], [24]. The main outcome of [8] was the use of Manhattan distance instead of Euclidean one [25], [26], [27]. The main reason for this is the fact that for $p < 1$ functional l_p is not a norm but is a non-convex quasinorm. All methods and algorithms which assume triangle inequality [26], [28], [29] cannot use such a quasinorm.

Comparison of different l_p functionals for data mining problems is yet fragmental, see, for example, [8], [36], [37]. In our study, we performed systematic testing. In general case distance concentration for l_p functionals was less for lower p but for all p the shape of distance concentration as a function of dimension is qualitatively the same. Moreover, the difference in distance concentration for different p decreases with dimension increasing. We selected 25 databases and

978-1-7281-2009-6/$31.00

systematically tested the hypothesis that the measurement of dissimilarity by subquadratic norms $l_p (1 \leq p < 2)$ or even quasinorms $(0 < p < 1)$ can help to overcome the curse of dimensionality in classification problems. We demonstrated that these norms and quasinorms *do not* improve k Nearest Neighbour (kNN) classifiers in high dimensions systematically and significantly.

There are two main results in this study: (i) usage of l_p functionals with small p does not prevent the distance concentration and (ii) the smaller distance concentration does not mean the better accuracy of kNN classification.

The further part of our paper is organised as follows. Section 'Measure concentration' presents results of an empirical test of distance concentration for Relative Contrast (RC) and Coefficient of Variation (CV) also known as relative variance. Section 'Comparison of l_p functionals' describes the approaches used for l_p functionals comparison, the used databases and the classification quality measures. Results of the described tests are presented. In Section 'Discussion', discussion and outlook are presented.

All software and databases used for this study can be found in [38].

II. Measure concentration

Let us consider a database X with n data points $X = x_1, \dots, x_n$ and d real-valued attributes, $x_i = (x_{i1}, \dots, x_{id})$. We consider databases of two types: randomly generated database with i.i.d. components from the uniform distribution on the interval $[0, 1]$ (this section) and real life databases (Section III). The l_p functional for vector x is defined by (1). For comparability of results in this study, we consider set of norms and quasinorms used in [8] with one more quasinorm ($l_{0.01}$): $l_{0.01}, l_{0.1}, l_{0.5}, l_1, l_2, l_4, l_{10}, l_\infty$.

Fig. 1 shows forms of unit level sets for all considered norms and quasinorms excluding $l_{0.01}$ and $l_{0.1}$. For these two quasinorms, graphs are visually indistinguishable from the central cross.

Several different indicators were used to study distance concentration:

- Relative Contrast (RC) [6], [8], [19]

$$\mathrm{RC}_p(X, y) = \frac{\left| \max_i \|x_i - y\|_p - \min_i \|x_i - y\|_p \right|}{\min_i \|x_i - y\|_p}; \tag{2}$$

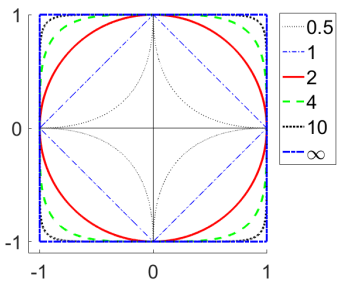

Fig. 1. Unit level sets for l_p functionals ("Unit spheres").

TABLE I
Comparison of RC for l_1 and l_2 for different dimension (Dim) and different number of points

Dim	$P(\mathbf{RC}_2 < \mathbf{RC}_1)$ for # of points			
	10 [8]	**10**	**20**	**100**
1	0	0	0	0
2	0.850	0.850	0.960	1.00
3	0.887	0.930	0.996	1.00
4	0.913	0.973	0.996	1.00
10	0.956	0.994	1.00	1.00
15	0.961	1.000	1.00	1.00
20	0.971	0.999	1.00	1.00
100	0.982	1.000	1.00	1.00

- Coefficient of Variations (CV) or relative variance [28], [29], [19]

$$\mathrm{CV}_p(X, y) = \frac{\sqrt{\mathrm{var}(\|x_i - y\|_p)}}{\mathrm{mean}(\|x_i - y\|_p)}, \tag{3}$$

  where $\mathrm{var}(x)$ is variance and $\mathrm{mean}(x)$ is mean value of random variable x;
- Hubness (popular nearest neighbours) [10].

In our study, we use RC and CV.

Table 2 in [8] shows that fraction of cases where $\mathrm{RC}_1 > \mathrm{RC}_2$ increases with dimension. It can be easily shown that for specific choice of X and y all three relations between RC_1 and RC_2 are possible $\mathrm{RC}_1(X, y) > \mathrm{RC}_2(X, y)$, $\mathrm{RC}_1(X, y) = \mathrm{RC}_2(X, y)$, or $\mathrm{RC}_1(X, y) < \mathrm{RC}_2(X, y)$. To evaluate the probabilities of these three outcomes, we performed the following experiment. We generated dataset X with k points and 100 coordinates. Each coordinate of each point was uniformly randomly generated for the interval $[0, 1]$. For each dimension $d = 1, 2, 3, 4, 10, 15, 20, 100$, we create d dimensional database X_d by selection of the first d coordinates of points in X. We calculated $\mathrm{RC}_p(X)$ as the mean value of RC for each point of X_d:

$$\mathrm{RC}_p(X) = \frac{1}{k} \sum_{i=1}^{k} \mathrm{RC}_p(X_d \backslash \{x_i\}, x_i),$$

where $X \backslash \{y\}$ is the database X without point y. We repeated this procedure 1000 times and calculated the fraction of cases when $\mathrm{RC}_1(X) > \mathrm{RC}_2(X)$. Results of this experiment are presented in Table I. Table I shows that for $k = 10$ points our results are very similar to the results presented in Table 2 in [8]. Increasing of number of points shows that already for relatively small number of points ($k \approx 20$) for almost all databases $\mathrm{RC}_1(X) > \mathrm{RC}_2(X)$.

This means that appearance of non-negligible fraction of cases where $\mathrm{RC}_2 > \mathrm{RC}_1$ is caused by very small size of a sample. For not so small samples we almost always have $\mathrm{RC}_2 < \mathrm{RC}_1$. The main reason for this is different pairs of closest and furthest points for different metrics. Several examples of such sets are presented in Fig. 2. Fig. 2 shows that $\mathrm{RC}_2 < \mathrm{RC}_\infty$ in rows 3, 5, 6, and 8 and $\mathrm{RC}_1 < \mathrm{RC}_2$ in row 6. These results allow us to formulate a hypothesis that in general case almost always $\mathrm{RC}_p < \mathrm{RC}_q, \forall p > q$. RC

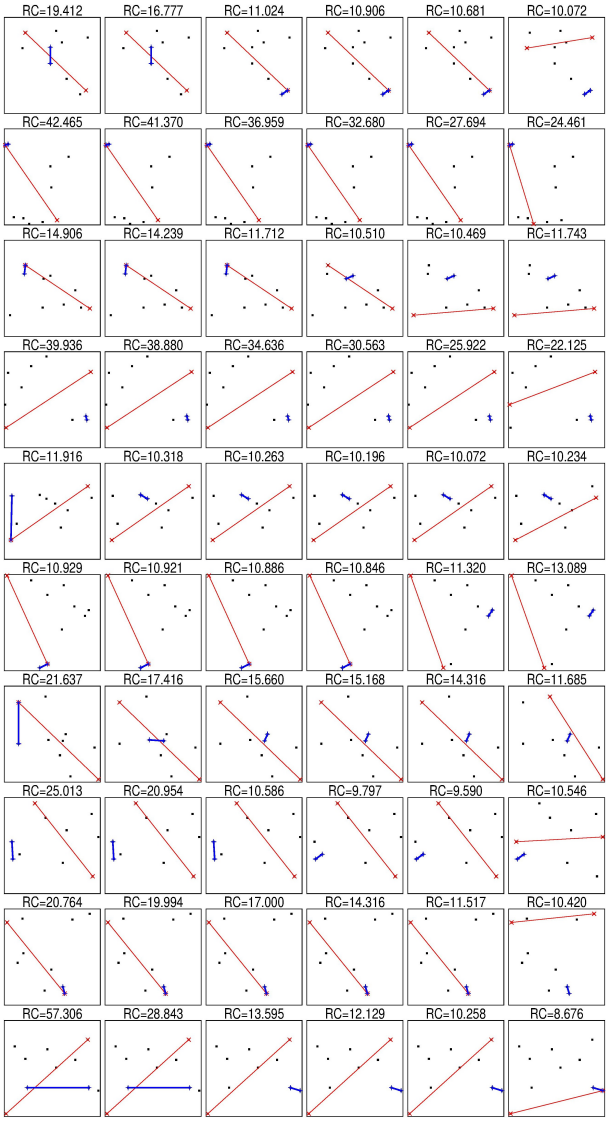

Fig. 2. 10 randomly generated sets of 10 points, thin red line connects the furthest points and bold blue line connects closest points, columns (from left to right) corresponds to $p = 0.01, 0.1, 0.5, 1, 2, \infty$

widely used to study properties of finite set of points but for distributions of points the CV is more appropriate. We assume that for CV hypothesis $\mathrm{CV}_p < \mathrm{CV}_q, \forall p > q$ is also true.

To check this hypothesis we performed the following experiment. We generated database X with 10,000 points in 200 dimensional space. Each coordinate of each point was uniformly randomly generated for the interval $[0, 1]$. We selected the set of dimensions $d = 1, 2, 3, 4, 5, 10, 15, \ldots, 195, 200$ and the set of l_p functionals $l_{0.01}, l_{0.1}, l_{0.5}, l_1, l_2, l_4, l_{10}, l_\infty$. For each dimension d, we prepared the database X_d as the set of the first d coordinates of points in database X. For each database X_d and each l_p functional, we calculate the set of all pairwise distances D_{dp}. Then we estimated the following values:

$$\mathrm{RC}_p = \frac{\max D_{dp} - \min D_{dp}}{\min D_{dp}}, \mathrm{CV}_p = \frac{\sqrt{\mathrm{var}(D_{dp})}}{\mathrm{mean}(D_{dp})}.$$

Graphs of RC_p and CV_p are presented in Fig. 3. Fig. 3 shows that our hypotheses are true. We can see that RC and CV as functions of dimension have qualitatively the same shape but in different scales: RC in the logarithmic scale. The paper [8] states that qualitatively different behaviour of $\max_i \|x_i\|_p - \max_i \|x_i\|_p$ for different p. We can state that for relative values we observe qualitatively the same behaviour with small quantitative difference $\mathrm{RC}_p - \mathrm{RC}_q$ which decreases with dimensionality increasing. This means that there could be some preference in usage of lower values of p but the fractional metrics do not provide a panacea from the curse of dimensionality. To analyse this hypothesis, we study the real live benchmarks in the next section.

III. Comparison of l_p functionals

In the previous section, we demonstrated that RC_p is higher for smaller p. Paper [6] shows that greater RC means 'more meaningful' task for kNN. We decide to compare different l_p functions for kNN classification. Classification has one additional benefit in comparison with regression and clustering problems: classification quality measure is classifier independent and similarity measure independent [39]. For this study, we selected three classification quality measures: Total Number of Neighbours of the Same Class (TNNSC), accuracy (fraction of correctly recognised cases), sum of sensitivity (fraction of correctly solved cases of positive class) and specificity (fraction of correctly solved cases of negative class). TNNSC is not an obvious measure of classification quality and we use it for comparability of our results with [8]. The 11 nearest neighbours as the method of classification was selected also for comparability with [8].

A. Databases for comparison

We selected 25 databases from UCI data repository [40]. We applied the following criteria for the database selection:

1) Data are not time-series.
2) Database is formed for the binary classification problem.
3) Database does not contain any missed values.
4) Number of attributes is less than number of observations and is greater than 3.
5) All predictors are binary or numeric.

Totally, we selected 25 databases and 37 binary classification problems. For simplicity, further we call each problem

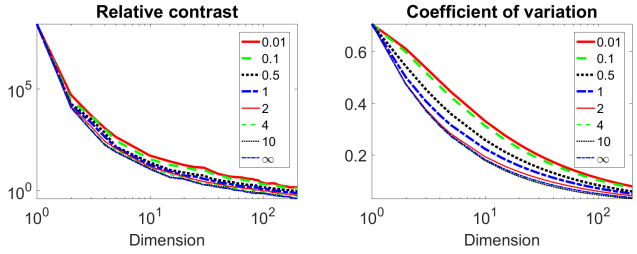

Fig. 3. Changes of RC (left) and CV (right) with dimension increasing for several metrics

TABLE II
DATABASES SELECTED FOR ANALYSIS

Name	Source	Dim.	Cases
Blood	[41]	4	748
Banknote authentication	[42]	4	1,372
Cryotherapy	[43], [44], [45]	6	90
Vertebral Column	[46]	6	310
Immunotherapy	[43], [44], [47]	7	90
HTRU2	[48], [49], [50]	8	17,898
ILPD (Indian Liver Patient Dataset)	[51]	10	579
Planning Relax	[52]	10	182
MAGIC Gamma Telescope	[53]	10	19,020
EEG Eye State	[54]	14	14,980
Climate Model Simulation Crashes	[55]	18	540
Diabetic Retinopathy Debrecen	[56], [57]	19	1,151
SPECT Heart	[58]	22	267
Breast Cancer	[59]	30	569
Ionosphere	[60]	34	351
QSAR biodegradation	[61], [62]	41	1,055
SPECTF Heart	[58]	44	267
MiniBooNE particle identification	[63]	50	130,064
First-order theorem proving (6 tasks)	[64], [65]	51	6,118
Connectionist Bench (Sonar)	[66]	60	208
Quality Assessment of Digital Colposcopies (7 tasks)	[67], [68]	62	287
Musk 1	[69]	166	476
Musk 2	[69]	166	6,598
Madelon	[70], [71]	500	2,600
Gisette	[72], [70]	5,000	7,000

a 'database'. The list of selected databases is presented in Table II.

We do not try to identify the best database preprocessing for each database. We simply use three preprocessing for each database:

- without preprocessing means usage data 'as is';
- standardisation means to shift and scale data to have zero mean and unit variance;
- standard dispersion means to shift and scale data to belong interval $[0, 1]$.

B. Approaches to comparison

Our purpose is to compare metrics but not to create the best classifier to solve each problem. Following [8] we use 11NN classifier. One of the reasons to select kNN is strong dependence of kNN on selected metrics and, on the other hand, the absence of any assumption about data, excluding the principle: tell me your neighbours, and will I tell you what you are. In our study, we consider 11NN with $l_{0.01}, l_{0.1}, l_{0.5}, l_1, l_2, l_4, l_1 0, l_\infty$ as different classifiers.

We applied several approaches to compare several l_p functionals (algorithms, classifiers):

- number of databases for which algorithm is the best [73];
- number of databases for which algorithm is the worst [73];
- number of databases for which algorithm has performance which statistically insignificantly different from the best;
- number of databases for which algorithm has performance which statistically insignificantly different from the worst;
- Friedman test [74], [75] and post hoc Nomenyi test [76] which were specially developed for multiple algorithms comparison;
- Wilcoxon signed rank test was used for comparison of three pairs of metrics.

The first four approaches we call frequency comparison. To avoid discrepancies, a description of all statistical tests used is presented below.

1) Proportion estimation: Since accuracy and $\mathrm{TNNSC}/(11 \times n)$, where n is number of cases in database, are proportions we can apply z-test of proportion estimations [77]. We compare two proportions with the same sample size, hence, we can use simplified formula for test statistics:

$$z = \frac{|p_1 - p_2|}{\sqrt{\frac{p_1+p_2}{n}\left(1 - \frac{p_1+p_2}{2}\right)}},$$

where p_1 and p_2 are two proportions to compare. *P*-value of this test is probability of observing by chance the same or greater z if both samples are taken from the same population. *P*-value is $p_z = \Phi(-z)$, where $\Phi(z)$ is standard cumulative normal distribution. We also meet the problem of reasonable selection of significance level. Selected databases contain from 90 to 130,064 cases. Usage of the same threshold for all databases is meaningless [78], [79]. The necessary sample size n can be estimated through the specified significance level of $1 - \alpha$, the statistical power $1 - \beta$, the expected effect size e, and the population variance s^2. For the normal distribution (since we use z-test):

$$n = \frac{2(z_{1-\alpha} + z_{1-\beta})^2 s^2}{d^2}.$$

In this study, we assume that the significance level is equal to the statistical power $\alpha = \beta$, the expected effect size is 1% (1% difference in accuracy is big enough), and the population variance can be estimated by

$$s^2 = n\frac{n_+}{n}\left(1 - \frac{n_+}{n}\right) = \frac{n_+(n - n_+)}{n},$$

where n_+ is number of cases in the positive class. Under this assumptions, we can estimate reasonable significance level as

$$\alpha = \Phi\left(\frac{d}{s}\sqrt{\frac{n}{8}}\right).$$

Usage of 8 l_p functionals means multiple testing. To avoid overdetection problem, we apply the Bonferroni correction [80]. From the other side, usage of too big significance level is also meaningless [78]. As a result we select the significance level as

$$\alpha = \max\left\{\frac{1}{28}\Phi\left(\frac{d}{s}\sqrt{\frac{n}{8}}\right), 0.00001\right\}.$$

Differences between two proportions (TNNSC or accuracies) is statistically significant if $p_z < \alpha$. It is necessary to stress that for TNNSC the number of cases is $11n$ because we consider 11 neighbours for each point.

2) Friedman test and post hoc Nomenyi test: One of the widely used statistical tests for algorithms comparison on many databases is Friedman test [74], [75]. To apply this test, we need firstly to apply tied ranking for the classification quality measure for one database: if several classifiers provide exactly the same quality measure then rank of all such classifiers will be equal to average value of the ranks for which they were tied [75]. Let us denote the number of used databases as N, the number of used classifiers as m and the rank of classifier i for database j as r_{ji}. Mean rank of classifier i is

$$R_i = \frac{1}{N}\sum_{j=1}^{N} r_{ji}.$$

Test statistics is

$$\chi_F^2 = \frac{4N^2(m-1)\Big(\sum_{i=1}^{m} R_i^2 - \frac{m(m+1)^2}{4}\Big)}{4\sum_{i=1}^{m}\sum_{j=1}^{N} r_{ji}^2 - Nm(m+1)^2}.$$

Test statistics under null hypothesis that all classifiers have the same performance follows χ^2 distribution with $m-1$ degrees of freedom. *P*-value of this test is probability of observing by chance the same or greater χ_F^2 if all classifiers have the same performance. *P*-value is $p_\chi = 1 - F(\chi_F^2; m-1)$, where $F(\chi; df)$ is cumulative χ^2 distribution with *df* degrees of freedom. Since we have 37 databases only we decide to use 95% significance level.

If Friedman test shows enough evidence to reject null hypothesis then we can conclude that not all classifiers have the same performance. To identify the pairs of classifiers with significantly different performance we applied post hoc Nomenyi test [76]. Test statistics for comparison of i and j classifiers is $|R_i - R_j|$. Critical distance

$$\text{CD} = q_{\alpha m}\sqrt{\frac{m(m+1)}{6N}}.$$

is used to identify pairs with statistically significant differences. $q_{\alpha m}$ is critical value for Nomenyi test with significance level of $1-\alpha$ and m degrees of freedom. The difference of classifiers performances is statistically significant with significance level of $1-\alpha$ if $|R_i - R_j| > CD$.

3) Wilcoxon signed rank test: To compare the performance of two classifiers on several databases we applied Wilcoxon signed rank test [81]. For this test we used standard Matlab function **signrank** [82].

C. Results of comparison

Results of frequency comparison are presented in Table III. Table III shows that indicator 'The best' is not robust and cannot be considered as a good tool for performance comparison [73]. For example, for TNNSC without preprocessing $l_{0.1}$ is the best for 11 databases and it is maximal value but $l_{0.5}, l_1$ and l_2 are essentially better if we consider indicator 'Insignificantly different from the best': 26 databases for $l_{0.1}$ and 31 databases for $l_{0.5}, l_1$ and l_2. Unfortunately, we cannot estimate this indicator for 'sensitivity plus specificity' quality measure by used way (it can be done by t-test). Analysis of Table III shows that in average $l_{0.5}, l_1, l_2$ and l_4 are the best and $l_{0.01}$ and l_∞ are the worst. Results of Friedman and post hoc Nomenyi tests are presented in Table IV. It can be seen that l_1 is the best for 6 of 9 tests and $l_{0.5}$ is the best for the remaining 3 tests. From the other side, performances of $l_{0.5}, l_1$ and l_2 are insignificantly different for all 9 tests.

We compared 8 different l_p functionals on 37 databases. Authors of [8] formulated the hypotheses that: (i) l_1 based kNN is better than l_2 based one and (ii) that the "fractional" metrics can further improve performance. We can test the differences between $l_{0.5}, l_1$ and l_2 based kNN by direct usage of Wilcoxon test. This comparison does not take into account the multiple testing. Results of comparisons are presented in Table V. The left table shows that for most cases $l_{0.5}$ and l_1 based kNN have insignificantly different performances and for the most cases l_2 based kNN is slightly worse than the previous two. Right table shows, that $l_{0.5}$ and l_2 based kNN are insensitive to preprocessing (performances of both methods are not significantly different for different preprocessing). In contrast with these two methods, l_1 based kNN shows significant difference for standard dispersion preprocessing in comparison with two other preprocessing.

IV. Discussion

Authors of [8] found that l_1 "is consistently more preferable than the Euclidean distance metric for high dimensional data mining applications". Our study partially confirmed this finding: kNN with l_1 distance frequently demonstrates the better performance in comparison with $l_{0.01}, l_{0.1}, l_{0.5}, l_2, l_4, l_{10}, l_\infty$ but this difference is not statistically significant. Really performance of kNN on the base of $l_{0.5}, l_1$ and l_2 functionals is indistinguishable. Detailed pairwise comparison of $l_{0.5}, l_1$ and l_2 functionals shows that the performance of l_1 based kNN is more sensitive to used data preprocessing than l_2. There is no unique and unconditional leader in l_p functionals. We can conclude, that l_p based kNN with very small $p < 0.1$ and very big $p > 4$ are almost always worse than with $0.1 < p < 4$. Our wide test shows that for all used preprocessing and all considered classifier quality measures, the performance of l_p based kNN classifiers for $l_{0.5}, l_1$ and l_2 are not statistically significantly different.

There are many questions for further study: does the kNN performance depend on the "real" data dimension? How can we measure this dimension? Can the number of l_2 based major principal components be considered as a reasonable estimate of the "real" data dimension or it is necessary to use l_1 based PCA? Recently developed PQSQ PCA [24] gives the possibility to create PCA with various subquadratic functionals, including l_p for $0 < p \leq 2$.

Authors of [8] state that "fractional distance metrics can significantly improve the effectiveness of standard clustering

TABLE III
FREQUENCY COMPARISON FOR TNNSC, ACCURACY AND SENSITIVITY PLUS SPECIFICITY

Indicator\ p for l_p functional	**0.01**	**0.1**	**0.5**	**1**	**2**	**4**	**10**	∞
TNNSC								
Without preprocessing								
The best	2	11	5	10	7	1	1	1
The worst	19	0	1	0	1	3	4	8
Insignificantly different from the best	17	26	31	31	31	30	23	22
Insignificantly different from the worst	34	23	17	19	21	21	25	29
Standardisation								
The best	0	5	10	11	6	2	1	1
The worst	18	2	0	0	1	2	4	10
Insignificantly different from the best	19	26	33	32	31	30	25	24
Insignificantly different from the worst	35	24	20	19	20	21	25	28
Standard dispersion								
The best	1	5	10	13	4	6	1	3
The worst	23	4	2	2	3	3	4	7
Insignificantly different from the best	19	26	32	31	30	29	26	26
Insignificantly different from the worst	36	24	22	21	22	22	26	26
Accuracy								
Without preprocessing								
The best	3	9	9	15	6	5	1	2
The worst	13	3	1	2	4	4	9	14
Insignificantly different from the best	29	31	34	35	35	35	33	30
Insignificantly different from the worst	35	32	28	28	29	29	30	31
Standardisation								
The best	2	5	12	18	7	3	1	1
The worst	13	4	0	0	2	6	7	13
Insignificantly different from the best	30	31	34	34	33	31	32	30
Insignificantly different from the worst	35	32	29	29	30	31	33	33
Standard dispersion								
The best	2	7	15	8	8	3	3	6
The worst	18	6	3	4	5	9	8	8
Insignificantly different from the best	30	31	34	33	33	32	31	32
Insignificantly different from the worst	36	33	31	31	31	32	33	32
Sensitivity plus specificity								
Without preprocessing								
The best	4	8	7	12	7	5	1	1
The worst	14	2	1	1	3	5	8	12
Standardisation								
The best	4	7	8	15	7	2	1	0
The worst	13	3	0	0	2	5	4	15
Standard dispersion								
The best	5	8	13	6	9	3	4	5
The worst	15	4	2	3	3	7	8	13

TABLE IV
RESULTS OF THE FRIEDMAN TEST AND POST HOC NOMENYI TEST

Preprocessing	**Quality measure**	**Friedman's *p*-value**	**The best l_p**		**Set of insignificantly different**							
			p	R_i	**0.01**	**0.1**	**0.5**	**1**	**2**	**4**	**10**	∞
Without preprocessing	TNNSC	< 0.0001	1	6.2639		X	X	X	X	X		
	Accuracy	< 0.0001	1	6.2639		X	X	X	X			
	Se+Sp	< 0.0001	0.5	6.0556		X	X	X	X			
Standardisation	TNNSC	< 0.0001	1	6.6944			X	X	X			
	Accuracy	< 0.0001	1	6.8056			X	X	X			
	Se+Sp	< 0.0001	1	6.4722		X	X	X	X			
Standard dispersion	TNNSC	< 0.0001	1	6.4722			X	X	X	X		
	Accuracy	< 0.0001	0.5	6.0000		X	X	X	X			
	Se+Sp	< 0.0001	0.5	6.0000			X	X	X	X		

TABLE V

P-VALUES OF WILCOXON TEST FOR DIFFERENT l_p FUNCTIONS (LEFT) AND DIFFERENT TYPE OF PREPROCESSING (RIGHT): W FOR DATA WITHOUT PREPROCESSING, S FOR STANDARDISED AND D FOR STANDARD DISPERSION PREPROCESSING, SE+SP STANDS FOR SENSITIVITY PLUS SPECIFICITY

Preprocessing	**Quality measure**	***p*-value for l_p and l_q**		
		0.5 & 1	**0.5 & 2**	**1 & 2**
Without preprocessing	TNNSC	0.6348	0.3418	0.0469
	Accuracy	0.9181	0.0657	0.0064
	Se+Sp	0.8517	0.0306	0.0022
Standardised	TNNSC	0.3098	0.1275	0.0014
	Accuracy	0.6680	0.0202	0.0017
	Se+Sp	0.8793	0.0064	0.0011
Standard dispersion	TNNSC	0.7364	0.0350	0.0056
	Accuracy	0.1525	0.0218	0.2002
	Se+Sp	0.1169	0.0129	0.3042

Quality measure	p **of** l_p **function**	***p*-value for pair of preprocessing**		
		W & S	**W & D**	**S & D**
TNNSC	0.5	0.5732	0.8382	0.6151
	1	0.9199	0.5283	0.1792
	2	0.9039	0.3832	0.1418
Accuracy	0.5	0.8446	0.5128	0.3217
	1	0.8788	0.0126	0.0091
	2	0.5327	0.3127	0.3436
Se+Sp	0.5	0.6165	0.2628	0.0644
	1	0.5862	0.0054	0.0067
	2	0.6292	0.3341	0.4780

algorithms". Our experiments show that for both considered distance concentration measures inequalities $RC_p < RC_q, \forall p > q$ and $CV_p < CV_q, \forall p > q$ hold. From the other side, our results show that there is no direct relation between distance concentration (e.g. RC or CV) and quality of classifiers: $l_{0.01}$ based kNN has one of the worst performance but the greatest RC and CV. The question about performance of clustering algorithms with different l_p functionals remains still open. This problem seems to be less clearly posed than for classification because there is no unconditional criteria of "proper clustering".

REFERENCES

[1] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton university press, 1961.

[2] C. M. Bishop, "The curse of dimensionality," *Pattern Recognition and Machine Learning*, pp. 33–38, 2006.

[3] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[4] H. Trevor, T. Robert, and F. JH, "The elements of statistical learning: data mining, inference, and prediction," 2009.

[5] F. Korn, B.-U. Pagel, and C. Faloutsos, "On the "dimensionality curse" and the "self-similarity blessing"," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 1, pp. 96–111, 2001.

[6] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?" in *International conference on database theory*. Springer, 1999, pp. 217–235.

[7] A. Hinneburg, C. C. Aggarwal, and D. A. Keim, "What is the nearest neighbor in high dimensional spaces?" in *26th Internat. Conference on Very Large Databases*, 2000, pp. 506–515.

[8] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*. Springer, 2001, pp. 420–434.

[9] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *ACM Sigmod Record*, vol. 30, no. 2. ACM, 2001, pp. 37–46.

[10] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in space: Popular nearest neighbors in high-dimensional data," *Journal of Machine Learning Research*, vol. 11, no. Sep, pp. 2487–2531, 2010.

[11] P. C. Kainen, "Utilizing geometric anomalies of high dimension: When complexity makes computation easier," in *Computer Intensive Methods in Control and Signal Processing*. Springer, 1997, pp. 283–294.

[12] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3025–3032.

[13] A. N. Gorban, I. Y. Tyukin, and I. Romanenko, "The blessing of dimensionality: separation theorems in the thermodynamic limit," *IFAC-PapersOnLine*, vol. 49, no. 24, pp. 64–69, 2016.

[14] A. N. Gorban and I. Y. Tyukin, "Blessing of dimensionality: mathematical foundations of the statistical physics of data," *Phil. Trans. R. Soc. A*, vol. 376, no. 2118, p. 20170237, 2018.

[15] G. Liu, Q. Liu, and P. Li, "Blessing of dimensionality: Recovering mixture data via dictionary pursuit," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 47–60, 2017.

[16] A. Gorban, A. Golubkov, B. Grechuk, E. Mirkes, and I. Tyukin, "Correction of ai systems by linear discriminants: Probabilistic foundations," *Information Sciences*, vol. 466, pp. 303–322, 2018.

[17] A. Giannopoulos and V. Milman, "Concentration property on probability spaces," *Advances in Mathematics*, vol. 156, no. 1, pp. 77 – 106, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001870800919492

[18] M. Ledoux, *The concentration of measure phenomenon*. American Mathematical Soc., 2001, no. 89.

[19] D. Francois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 873–886, 2007.

[20] J. Anderson, M. Belkin, N. Goyal, L. Rademacher, and J. Voss, "The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures," in *Conference on Learning Theory*, 2014, pp. 1135–1164.

[21] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS math challenges lecture*, vol. 1, no. 2000, p. 32, 2000.

[22] G. Cormode, P. Indyk, N. Koudas, and S. Muthukrishnan, "Fast mining of massive tabular data via approximate distance computations," in *Data Engineering, 2002. Proceedings. 18th International Conference on*. IEEE, 2002, pp. 605–614.

[23] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the twentieth annual symposium on Computational geometry*. ACM, 2004, pp. 253–262.

[24] A. N. Gorban, E. M. Mirkes, and A. Zinovyev, "Data analysis with arbitrary error measures approximated by piece-wise quadratic pqsq functions," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[25] E. A. Allen, E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, and V. D. Calhoun, "Tracking whole-brain connectivity dynamics in the resting state," *Cerebral Cortex*, vol. 24, no. 3, pp. 663–676, 2014. [Online]. Available: http://dx.doi.org/10.1093/cercor/bhs352

[26] C. Elkan, "Using the triangle inequality to accelerate k-means," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003, pp. 147–153.

[27] E. Chang, K. Goh, G. Sychay, and G. Wu, "Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 1, pp. 26–38, 2003.

[28] P. Demartines, "Analyse de données par réseaux de neurones auto-organisés," Ph.D. dissertation, Grenoble INPG, 1994 (in French).

[29] P. N. Yianilos, "Excluded middle vantage point forests for nearest neighbor search," in *In DIMACS Implementation Challenge, ALENEX'99*. Citeseer, 1999.

[30] G. Köthe, *Topological Vector Spaces. Translated by DJH Garling*. Springer-Verlag, 1969.

[31] D. François, V. Wertz, M. Verleysen *et al.*, "Non-euclidean metrics for similarity search in noisy datasets." in *ESANN*, 2005, pp. 339–344.

[32] A. Dik, K. Jebari, A. Bouroumi, and A. Ettouhami, "Fractional metrics for fuzzy c-means," *International Journal of Computer and Information Technology*, vol. 3, no. 6, pp. 1490–1495, 2014.

[33] B. Jayaram and F. Klawonn, "Can unbounded distance measures mitigate the curse of dimensionality?" *International Journal of Data Mining, Modelling and Management*, vol. 4, no. 4, 2012.

[34] S. L. France, J. D. Carroll, and H. Xiong, "Distance metrics for high dimensional nearest neighborhood recovery: Compression and normalization," *Information Sciences*, vol. 184, no. 1, pp. 92–110, 2012.

[35] K. Doherty, R. Adams, and N. Davey, "Non-euclidean norms and data normalisation," in *Proceedings of the*, 2004.

[36] A. Singh, A. Yadav, and A. Rana, "K-means with three different distance metrics," *International Journal of Computer Applications*, vol. 67, no. 10, 2013.

[37] L.-Y. Hu, M.-W. Huang, S.-W. Ke, and C.-F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *SpringerPlus*, vol. 5, no. 1, p. 1304, 2016.

[38] A. N. Gorban, J. Allohibi, and E. M. Mirkes, "Databases and code for l_p functional comparison." [Online]. Available: https://github.com/Mirkes/Databases-and-code-for-l_p-functional-comparison

[39] C. Sammut and G. I. Webb, *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.

[40] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[41] "Blood transfusion service center." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

[42] "Banknote authentication." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/banknote+authentication

[43] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, and S. Nahavandi, "An expert system for selecting wart treatment method," *Computers in biology and medicine*, vol. 81, pp. 167–175, 2017.

[44] F. Khozeimeh, F. Jabbari Azad, Y. Mahboubi Oskouei, M. Jafari, S. Tehranian, R. Alizadehsani, and P. Layegh, "Intralesional immunotherapy compared to cryotherapy in the treatment of warts," *International journal of dermatology*, vol. 56, no. 4, pp. 474–478, 2017. [Online]. Available: https://doi.org/10.1111/ijd.13535

[45] "Cryotherapy." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Cryotherapy+Dataset+

[46] "Vertebral column." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Vertebral+Column

[47] "Immunotherapy." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Immunotherapy+Dataset

[48] "HTRU2." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/HTRU2

[49] R. Lyon, "HTRU2," 2016. [Online]. Available: https://doi.org/10.6084/m9.figshare.3080389.v1

[50] R. J. Lyon, B. Stappers, S. Cooper, J. Brooke, and J. Knowles, "Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach," *Monthly Notices of the Royal Astronomical Society*, vol. 459, no. 1, pp. 1104–1123, 2016. [Online]. Available: https://doi.org/10.1093/mnras/stw656

[51] "Indian liver patient." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29

[52] R. Bhatt, "Planning-relax dataset for automatic classification of eeg signals." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Planning+Relax

[53] "Magic gamma telescope." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope

[54] "Eeg eye state." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State#

[55] "Climate model simulation crashes." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Climate+Model+Simulation+Crashes

[56] "Diabetic retinopathy debrecen data set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set

[57] B. Antal and A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy," *Knowledge-based systems*, vol. 60, pp. 20–27, 2014.

[58] "Spectf heart." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/SPECTF+Heart

[59] "Breast cancer wisconsin (diagnostic)." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[60] "Ionosphere." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Ionosphere

[61] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni, "Quantitative structure–activity relationship models for ready biodegradability of chemicals," *Journal of chemical information and modeling*, vol. 53, no. 4, pp. 867–878, 2013.

[62] "QSAR biodegradation." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/QSAR biodegradation

[63] "Miniboone particle identification." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/MiniBooNE+particle+identification

[64] J. P. Bridge, S. B. Holden, and L. C. Paulson, "Machine learning for first-order theorem proving," *Journal of automated reasoning*, vol. 53, no. 2, pp. 141–172, 2014.

[65] "First-order theorem proving." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/First-order+theorem+proving

[66] "Connectionist bench (sonar mines vs. rocks)." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+%28Sonar%2C+Mines+vs.+Rocks%29

[67] "Quality assessment of digital colposcopies." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Quality+Assessment+of+Digital+Colposcopies

[68] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Iberian conference on pattern recognition and image analysis*. Springer, 2017, pp. 243–250.

[69] "Musk 1 and 2." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Musk+%28Version+1%29

[70] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, "Result analysis of the nips 2003 feature selection challenge," in *Advances in neural information processing systems*, 2005, pp. 545–552.

[71] "Madelon." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Madelon

[72] "Gisette." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Gisette

[73] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[74] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[75] ——, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.

[76] P. Nemenyi, "Distribution-free multiple comparisons," in *Biometrics*, vol. 18, no. 2. International Biometric Soc 1441 I ST, NW, SUITE 700, Washington, DC 20005-2210, 1962, p. 263.

[77] D. Altman, D. Machin, T. Bryant, and M. Gardner, *Statistics with confidence: confidence intervals and statistical guidelines*. John Wiley & Sons, 2013.

[78] D. J. Biau, S. Kernéis, and R. Porcher, "Statistics in brief: the importance of sample size in the planning and interpretation of medical research," *Clinical orthopaedics and related research*, vol. 466, no. 9, pp. 2282–2288, 2008.

[79] P. Kadam and S. Bhalerao, "Sample size calculation," *International journal of Ayurveda research*, vol. 1, no. 1, p. 55, 2010.

[80] E. W. Weisstein, "Bonferroni correction," 2004. [Online]. Available: http://mathworld.wolfram.com/BonferroniCorrection.html

[81] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[82] "Wilcoxon signed rank test." [Online]. Available: https://www.mathworks.com/help/stats/signrank.html