A generalized fuzzy k -nearest neighbor regression model based on Minkowski distance

Mahinda Mailagaha Kumbure¹ · Pasi Luukka¹

Received: 30 June 2021 / Accepted: 5 September 2021 / Published online: 25 September 2021
© The Author(s) 2021

Abstract

The fuzzy k -nearest neighbor (FKNN) algorithm, one of the most well-known and effective supervised learning techniques, has often been used in data classification problems but rarely in regression settings. This paper introduces a new, more general fuzzy k -nearest neighbor regression model. Generalization is based on the usage of the Minkowski distance instead of the usual Euclidean distance. The Euclidean distance is often not the optimal choice for practical problems, and better results can be obtained by generalizing this. Using the Minkowski distance allows the proposed method to obtain more reasonable nearest neighbors to the target sample. Another key advantage of this method is that the nearest neighbors are weighted by fuzzy weights based on their similarity to the target sample, leading to the most accurate prediction through a weighted average. The performance of the proposed method is tested with eight real-world datasets from different fields and benchmarked to the k -nearest neighbor and three other state-of-the-art regression methods. The Manhattan distance- and Euclidean distance-based FKNNreg methods are also implemented, and the results are compared. The empirical results show that the proposed Minkowski distance-based fuzzy regression (Md-FKNNreg) method outperforms the benchmarks and can be a good algorithm for regression problems. In particular, the Md-FKNNreg model gave the significantly lowest overall average root mean square error (0.0769) of all other regression methods used. As a special case of the Minkowski distance, the Manhattan distance yielded the optimal conditions for Md-FKNNreg and achieved the best performance for most of the datasets.

Keywords Fuzzy k -nearest neighbor · Regression · Minkowski distance · Machine learning · Performance measures

1 Introduction

In machine learning, a regression problem refers to estimating a real-valued continuous response (output) based on the values of one or more input variables. By determining the relationships between output and input variables, a regression method numerically predicts a target value. In the literature, various regression techniques have been introduced for a wide range of machine learning problems. Among them, k -nearest neighbor regression (KNNreg) (Benedetti 1977; Stone 1977; Turner 1977) has become

one of the most widely used regression techniques due to its simplicity and robustness (Buza et al. 2015). This method is an adapted version of the k -nearest neighbor (KNN) model that was initially introduced by Cover and Hart (1967) for applying classification problems. The main idea of the KNNreg is to predict the output value for a given test sample by averaging the output values of the nearest neighbor samples (Hu et al. 2014).

Though the KNN method has many significant advantages, it intuitively suffers from some weaknesses, for example, giving equal importance to all nearest neighbors (even if some of them are quite far from the test sample) in the classification process. To improve model and alleviate such issues, Keller et al. (1985) introduced the idea of using the degree of membership in the KNN method to propose its fuzzy version, called the fuzzy k -nearest neighbor (FKNN) classifier. Thanks to its capability of tackling uncertainty issues in the data, the FKNN model

✉ Mahinda Mailagaha Kumbure
mahinda.mailagaha.kumbure@lut.fi

Pasi Luukka
pasi.luukka@lut.fi

¹ School of Business and Management, LUT University,
53850 Lappeenranta, Finland

has proven promising for classification problems (Chen et al. 2013; Yu et al. 2002) compared to the classical KNN method. Although the FKNN classifier has received much attention in terms of classification, it has received less attention in the context of regression. This motivated us to establish the fuzzy k -nearest neighbor regression (FKNNreg) model in this research by modifying the original FKNN rule.

Typically, the distance metric is one of the main components of distance-based classifiers such as the KNN and FKNN methods (Rastin et al. 2021). Even though the Euclidean distance is the most common distance metric used in such methods to measure the similarity between two data samples, it is often not optimal for every problem domain (Cai et al. 2020; Nguyen et al. 2016). Several research papers have reported better results with a more general choice of distance metric (Chang et al. 2006; Dettmann et al. 2011; Jenicka and Suruliandi 2011; Kaski et al. 2001; Koloseni et al. 2012, 2013). Besides, the Euclidean distance has several drawbacks. For example, if two data samples have no feature values in common, they might have a shorter distance than the other sample pairs, including the same feature values (Shirkhorshidi et al. 2015). These facts encouraged us to examine the effectiveness of the Minkowski distance in the FKNN rule in the regression setting for low- and high-dimensional datasets.

The main goal of this study is to introduce the FKNNreg using the Minkowski distance metric and to examine its efficiency. The combination of the Minkowski distance metric and the FKNNreg has not been studied in the literature before. This led us to create the Minkowski distance-based fuzzy k -nearest neighbor regression (Md-FKNNreg) algorithm. A key advantage of this method is that the nearest neighbors are weighted by fuzzy weights considering their similarity to the test sample, leading to the most accurate prediction through a weighted average. Also, utilization of the Minkowski distance allows greater flexibility for obtaining more relevant neighboring samples close to the target sample.

Intuitively, most available regression models (e.g., multiple linear regression [MLR], least absolute shrinkage and selection operator [LASSO] regression) are based on assumptions regarding the distribution of the data. However, it is rarely confirmed that these assumptions apply to real-world problems. That is being said, an interesting fact about the KNNreg methods is that they do not explicitly make any assumptions about the underlying data (Yao and Ruzzo 2006) or model's components and simply use training data to make predictions. Another advantage is that they are, in general, relatively easy to implement and interpret and can potentially be applied even for non-linear problems (Hu et al. 2014). Moreover, support vector regression (SVR) is recognized as one of the well-known

methods applied for non-linear regression problems. However, its utilization is restricted in various disciplines due to the difficulty of selecting suitable parameters for the model (Liu et al. 2013). In this regard, FKNNreg methods could be better alternatives in the regression context, and the proposed new KNNreg method is found to be significant for non-linear regression problems.

To study the performance of the proposed Md-FKNNreg model, we conducted an experiment using real-world data from various applications. We compared the regression performance of the proposed variant with the KNNreg, Lasso, SVR, and multiple linear regression models. In addition, the Manhattan distance-based fuzzy k -nearest neighbor regression (Man-FKNNreg) and Euclidean distance-based fuzzy k -nearest neighbor regression (Euc-FKNNreg) methods were also implemented, and the results were compared. To evaluate the regression performance, we used root mean square error (RMSE) and the coefficient of determination (R^2) values as the evaluation metrics. We also tested whether there was a statistically significant difference between the regression results for the Md-FKNNreg and baseline methods.

The main contributions of this paper can be summarized as follows:

- (1) We propose a new regression approach based on the FKNN algorithm.
- (2) We introduce the Minkowski distance into the nearest neighbors search in the proposed algorithm and investigate its efficiency and robustness.
- (3) We demonstrate the performance of the proposed regression model on low- and high-dimensional real-world data coming from different domains.
- (4) We analyze, compare, and benchmark the regression results of the proposed method with select well-known state-of-the-art regression methods.

The remainder of this paper is organized as follows. Section 2 discusses the background information related to the present study. Section 3 briefly provides the theoretical underpinning of the KNNreg and FKNNreg models and the Minkowski distance measure. Section 4 proposes the Md-FKNNreg method. Section 5 introduces the data used and the experiment setting for the proposed method and presents and discusses the empirical results obtained with the proposed method and benchmarks. Section 6 summarizes the main findings and provides concluding remarks.

2 Related work

The KNNreg model has the potential to tackle linear and non-linear problems in an effective way (Cai et al. 2020) and performs especially well in a high-dimensional space.

Accordingly, the growing popularity of the KNNreg method can be seen in various fields, including renewable energy (Hu et al. 2014; Huang and Perry 2016; Zhou et al. 2020), physics research (Durbin et al. 2021), biological studies (Yao and Ruzzo 2006), transportation (Cai et al. 2020; Dell’Acqua et al. 2015), robotics (Chen and Lau 2016), and telecommunication (Adege et al. 2018). In addition, some studies have also employed the KNNreg model with other approaches to develop effective hybrid models for specific applications. For example, Chen and Hao (2017) proposed an integrated framework by employing support vector machine (SVM) and KNNreg for stock market prediction. Salari et al. (2015) also presented a novel hybrid approach with a combination of a genetic algorithm (GA), the KNNreg method, and artificial neural network (ANN) for classification problems. Cheng et al. (2019) utilized the same idea as the KNNreg to introduce a novel approach for missing value imputations. Furthermore, the simplicity and strength of the KNNreg algorithm have encouraged researchers to develop different enhanced variants (for examples, see Buza et al. 2015; Guillen et al. 2010; Nguyen et al. 2016; Song et al. 2017) and to construct mathematical estimations (Biau et al. 2012).

An ideal distance measure must have the ability to precisely detect the similarity between two samples while allowing the researchers to understand how to compare, classify, or cluster those samples. Therefore, such metrics have great potential to influence the outcomes of the models used (Bergamasco and Nunes 2019). Accordingly, some previous studies focused only on which similarity measure best fit the particular situation (for examples, see Rodrigues 2018; Moghtadaiee and Dempster 2015; Huo et al. 2021). The Minkowski distance is the most investigated measure among the frequently applied techniques for measuring the similarity between instances in machine learning-based applications (Bergamasco and Nunes 2019; Cordeiro and Makarenkov 2016; Gueorguieva et al. 2017). The Minkowski distance is the main focus of this research because it offers the opportunity to compute the distance between two instances in several different ways and holds several well-known distances as special cases, e.g., the Manhattan and Euclidean distances.

The concept of the fuzzy theory, originally introduced by Zadeh (1965), can operate under uncertainty and has advanced in many different ways in various applications (for examples, see Chen et al. 1990; Chen and Hsiao 2005; Chen and Chen 2007; Chen and Chang 2010; Chen et al. 2009; Horng et al. 2005; Zeng et al. 2019). The FKNN classifier (Keller et al. 1985) was derived from fuzzy theory and has been one of the most effective techniques in supervised machine learning tasks. Nikoo et al. (2018) applied the FKNN classifier to a regression application without modifying its original algorithm explicitly (i.e., it

was operated as a classification task). However, to the best of our knowledge, no one has attempted to utilize the FKNN model in the regression setting. Thus, the effectiveness of FKNNreg for machine learning applications requires further investigation.

3 Preliminaries

This section briefly discusses the KNNreg method, the FKNN method, and the Minkowski distance measure.

3.1 K-nearest neighbor regression

KNNreg (Benedetti 1977; Stone 1977; Turner 1977) is a simple, effective, and robust nonlinear regression method. The basic idea of KNNreg is to predict an output value to a given input sample based on a fixed number (k) of its nearest neighbors found from the input-output training samples. The k is a smoothing parameter, and its value controls the adaptability of the KNNreg method (Hu et al. 2014). KNNreg does not require an explicit training step besides the initial dataset’s inputs and outputs, which represent a unique property. The notion of the KNNreg model can be formally defined as follows.

Let $T = \{(X_i, y_i)\}_{i=1}^N$ be a training dataset with N samples, where $X_i = \{x_1^i, x_2^i, \dots, x_m^i\} \in \mathbb{R}^m$ is an input sample i from m -dimensional feature space, and its output value (response variable) is $y_i \in Y$, where $Y = \{y_1, y_2, \dots, y_N\}$ denotes the set of output values. For a given new data sample X , the goal is to learn the predictor function $h(X)$ from the training dataset such that $\hat{y} \approx h(X)$, where $\hat{y}$ is the estimated value for the output y of X . The KNNreg starts with measuring the distance (d) between the test sample X and each sample X_i in T . In this case, the Euclidean distance is the most commonly adopted distance metric, and its formulation for the distance between $X = \{x_1, x_2, \dots, x_m\}$ and X_i is presented by Eq. (1).

$$d(X, X_i) = \sqrt{\sum_{j=1}^m (x_j - x_j^i)^2}. \quad (1)$$

Next, the set of k nearest neighbors, $N_X^k = \{(X_i, y_i)\}_{i=1}^k$ for X , is found from the reordered training samples in T according to the increasing Euclidean distances. Finally, the output value y for X is estimated by taking the arithmetic mean of the output values ($y_1, y_2, \dots, y_k$) of the nearest neighbors (Song et al. 2017; Biau et al. 2012; Györfi et al. 2002) as follows:

$$\hat{y} = \frac{\sum_{j=1}^k y_j}{k}. \quad (2)$$

This is based on the assumptions that training samples in the N_X^k have similar output values to $h(X)$ (Kramer 2011) and also that all nearest neighbors in the N_X^k have equal importance in the prediction (Cover and Hart 1967).

3.2 Fuzzy k -nearest neighbor classification method

Unlike the KNN algorithm, the FKNN method uses the unbiased weighing scheme in the decision rule using the distances between the test sample and the nearest neighbor samples. Put it differently, the FKNN model computes a membership to the test sample for each class and makes the class decision according to the highest membership degree (Keller et al. 1985). These fuzzy memberships have excellent potential for accurate predictions (Kumbure et al. 2020). The membership degree of a given new sample X in a class i that is represented by the k nearest neighbors¹ is measured as follows:

$$u_i(y) = \frac{\sum_{j=1}^k u_{ij}(1/\|X - X_j\|^{2/(q-1)})}{\sum_{j=1}^k (1/\|X - X_j\|^{2/(q-1)})}, \quad (3)$$

where $q \in (1, +\infty)$ is the fuzzy strength parameter that controls the Euclidean distance $\|X - X_j\|^2$ between X and X_j to weigh the contribution of each nearest neighbor to the membership value. Also, u_{ij} is the membership of the sample X_j from the training data to the class i among the k nearest neighbors. Two methods are used to measure the u_{ij} : crisp memberships and fuzzy memberships. More details about these methods can be found in the work by Chen et al. (2011).

3.3 Minkowski distance

The Minkowski distance measure (also called L_p norm space) is a class of various distance functions that are formed by the parameter p . For two given samples X_i and X_j where $X_i = \{x_1^i, x_2^i, \dots, x_m^i\} \in \mathbb{R}^m$ and $X_j = \{x_1^j, x_2^j, \dots, x_m^j\} \in \mathbb{R}^m$, the Minkowski distance metric is defined as follows:

$$d_{Md}(X_i, X_j) = \left(\sum_{t=1}^m |x_t^i - x_t^j|^p \right)^{1/p} \quad \text{for } p \geq 1. \quad (4)$$

From this metric, we can specify different distance functions by changing the value of p . For example, we can obtain the Manhattan distance (also known as the city

block distance or L_1 norm) by setting $p = 1$ and the Euclidean distance, also referred to as L_2 norm (see also Eq. (1)) by setting $p = 2$.

4 Proposed fuzzy k -nearest neighbor regression model using Minkowski distance

In this research, we focus on the fuzzy k -nearest neighbor regression. Given this, we define the FKNN method for regression together with the Minkowski distance. In this way, the novel regression method, Md-FKNNreg, is introduced. This method aims to achieve a reliable prediction for the predictor function by allowing the Minkowski distance to be adapted to the particular context with the optimal conditions. The procedure of the Md-FKNNreg method mainly includes four steps: measuring the distances, recognizing the nearest neighbors, computing the fuzzy weights, and making the prediction. The detailed process of this method is presented using the same notations in Sect. 3.1 as follows.

Step 1: Determine the Minkowski distance $d_{Md}(X, X_j)$ between X and X_j in T according to: $d_{Md}(X, X_j) = (\sum_{t=1}^m |x_t - x_t^j|^p)^{1/p}$.

Step 2: Find the set of k nearest neighbors N_X^k from the ranked training data samples according to increased Minkowski distances. Here, we used a grid-based search to find the optimal parameter p for the Minkowski distance and k that best fit a particular dataset.

Step 3: Calculate the fuzzy weight (w_j) for each nearest neighbor j using $d_{Md}(X, X_j)$ as follows:

$$w_j = \frac{1}{(1/d_{Md}(X, X_j))^{\frac{2}{q-1}}}, \quad \text{for } j = 1, 2, \dots, k, \quad (5)$$

where q is a fuzzy strength parameter, and $(\frac{2}{q-1})$ indicates the fuzziness exponent. The closer q is to 1, the larger the weights are. For distances over 1 unit, the larger q is, the smaller the weights are.

The purpose of these weights is to define a comprehensive linear predictor for the output value y such that $h(X) = W^T Y$, where $W = \{w_1, w_2, \dots, w_k\} \in \mathbb{R}^k$. The weighted value w_j ($0 \leq w_j \leq 1$) of the nearest neighbor X_j reflects its relative importance to Y .

Step 4: Predict the output value $\hat{y}$ for X by taking the weighted average (with the fuzzy weights) of the outputs y_j for $j = 1, 2, \dots, k$ in the N_X^k according to the following equation:

$$\hat{y} = \frac{\sum_{j=1}^k w_j y_j}{\sum_{j=1}^k w_j}. \quad (6)$$

¹ The set of nearest neighbors here is defined in terms of a classification problem, which means each nearest neighbor i contains m features values and a class label c_i (i.e., $X^i = \{x_1^i, x_2^i, \dots, x_m^i, c_i\}$). X is also shaped by similar characteristics.

It is clear that in this method, the Minkowski distance is used not only to find the nearest neighbors but also to measure the weights. Accordingly, the Minkowski distance plays a critical role in the proposed framework. Besides, the Md-FKNNreg method is intuitively adaptive to the number of nearest neighbors k and the p of the distance function to vary with iterations throughout its search in a particular situation. This characteristic allows the method to expand the search area to a broader domain. The steps of the Md-FKNNreg method discussed above are summarized in Algorithm 1 by introducing a pseudo-code to it. In addition, the pseudo-code for the grid search method used is presented in Algorithm 2.

Algorithm 1: Md-FKNNreg

Input: $\{X_i, y_i\}_{i=1}^N$ (training data set), X (test data sample), k ($1 \leq k \leq N$), p ($p \geq 1$)
Output: Predicted target value, $\hat{y}$ for X

```

1 Begin
2 for  $i \leftarrow 1$  to  $N$  do
3   Compute  $d_{Md}(X, X_i)$  between  $X$  and  $X_i$  using
   Eq. (4),
4   if  $i < k$  then
5     Add  $X_i$  to  $N_X^k$ 
6   else if  $X_i$  is closer to  $X$  than any of neighbors
   in  $N_X^k$  then
7     Drop the farthest neighbor from the set  $N_X^k$ 
   and add  $X_i$ 
8   end
9 end
10 for  $j \leftarrow 1$  to  $k$  do
11   Compute  $w_j = \frac{1}{(1/d_{Md}(X, X_j))^{\frac{p}{q-1}}}$ 
12 end
13 Estimate  $\hat{y}$  by taking weighted average such as

$$\hat{y} = \frac{\sum_{j=1}^k w_j y_j}{\sum_{j=1}^k w_j}$$

14 return  $\hat{y}$ 

```

Algorithm 2: Grid search

Input: p -values: $\{p_i\}_{i=1}^{n_p}$, k -values: $\{k_i\}_{i=1}^{n_k}$, n -runs for cross-validation
Output: Optimal p^* and k^*

```

1 Begin
2 for  $j \leftarrow 1$  to  $n$  do
3   Split the training data randomly into training
   and validation sets
4   for  $j \leftarrow 1$  to  $n_k$  do
5     for  $s \leftarrow 1$  to  $n_p$  do
6       Perform Md-FKNNreg model with  $k(j)$ 
   and  $p(s)$  and save the RMSE
7     end
8   end
9 end
10 for  $j \leftarrow 1$  to  $n_k$  do
11   for  $s \leftarrow 1$  to  $n_p$  do
12     Average the RMSE over  $n$ -runs
13   end
14 end
15 Estimate the minimum RMSE average and
   corresponding  $p^*$  and  $k^*$  values
16 return  $p^*$  and  $k^*$ 

```

In the KNNreg method, the prediction of the output for a new sample is made through a uniform weighting scheme (Cheng 1984). This means it makes the prediction

by taking the simple average of the outputs of the nearest neighbor samples and does not consider the distances between the new sample and its k nearest neighbors (i.e., all nearest neighbors have equal influence across the prediction) (Kramer 2011). In contrast, the FKNNreg uses an inverse weighting scheme that assigns higher weights to the closer training samples, allowing them more influence over the prediction. Moreover, the fuzzy strength parameter q controls how heavily the distance is weighted when determining the contribution of each nearest neighbor to the target sample (Keller et al. 1985). For example, when $q = 2$, the contributions of the nearest neighbors are weighted by the reciprocal of their distances to the target sample. Regarding the distance, the adopted distance metric plays a crucial role in achieving the best possible nearest neighbors and weighting them. Accordingly, the Minkowski distance is utilized in the proposed Md-FKNN method since it offers a more generalized nature² than the Euclidean distance and Manhattan distance. It has shown superior performance with supervised and unsupervised machine learning models compared to other distance measures (for examples, see Aggarwal et al. 2001; Ranmya and Sasikala 2019). Considering the above facts, overall, the proposed Md-FKNNreg is expected to produce significantly better results than the KNNreg and FKNNreg methods.

5 Experiment and results

This section presents the descriptions of the data sets used and the empirical procedures of the experiments conducted to investigate the regression performance of the proposed Md-FKNNreg model.

5.1 Data description

For our experiment, we selected eight real-world datasets that are freely available at the UCI Machine Learning repository (Dheeru and Taniskidou 2017) and at the Knowledge Extraction based on Evolutionary Learning (KEEL) repository (Alcala-Fdez et al. 2011). As summarized in Table 1, each of these datasets holds different characteristics in terms of the number of instances and features. Also, the related area of each of the datasets is provided in the “Domain” in Table 1.

² This can be defined in terms of the “metric space”: Minkowski metric space $\sim (\mathbb{R}^m, \mathbb{L}^p)$, Euclidean space $\sim (\mathbb{R}^m, \mathbb{L}^2)$, and Manhattan space $\sim (\mathbb{R}^m, \mathbb{L}^1)$.

Table 1 Summary of the datasets used in the experiment

Data set	Repository	Instances	Features	Domain
Stock	KEEL	950	9	Business
Airfoil	UCI	1503	5	Physics
AutoMPG	KEEL	392	6	Engineering
Baseball	KEEL	337	16	Sociology
Servo	UCI	167	4	IT
Laser	KEEL	993	4	Physics
Qsar Fish	UCI	908	6	Biology
Parkinson	UCI	5875	26	Medicine

5.2 Experimental setting

In each collected dataset, the data samples were divided into 40% for training, 40% for validation, and 20% for testing based on the works of Kumbure et al. (2019, 2020). Before data splitting, all the datasets were normalized into the unit interval of $[0, 1]$ to avoid data differences between small and large ranges. For cross-validation, we adopted the holdout method (Arlot and Celisse 2010), in which the training and validation datasets were randomly sampled 30 times, and mean performance measures were computed from the results. The examination of the proposed method using the data can be categorized into two phases, training & validation and testing. In the training and validation step, we trained the model by optimizing the parameter values for p in the Minkowski distance and the number of k nearest neighbors. Here, we used mean regression error to determine the optimal parameter values. To find the best possible values for the parameters, we deployed a grid search technique during the training and validation. Then, we evaluated the performance of the model with optimal parameters in the testing phase. The steps of this process are summarized by the flowchart in Fig. 1.

The proposed Md-FKNNreg, Man-FKNNreg and Euc-FKNNreg models were implemented using MATLAB 2019b software. The KNNreg was implemented from scratch. The SVR, LASSO, and MLR models were developed using the Statistics and Machine Learning Toolbox in MATLAB. The computer used was an Intel® Core™ i5 1.8GHz, 16GB RAM with Microsoft Windows 10 operating system.

5.3 Baseline models

We compared the performance of the developed Md-FKNNreg method with the classical KNNreg, Man-FKNNreg, and Euc-FKNNreg methods. In addition, we also selected three more commonly used regression techniques, namely SVR (Drucker et al. 1997), LASSO

regression (Tibshirani 1996; Wang et al. 2018), and MLR (Montgomery et al. 2012). The basic concepts of these methods are briefly presented.

SVR, a variant of SVM (Cortes and Vapnik 1995), is a non-linear kernel-based regression approach that performs the regression by constructing a hyperplane in a high-dimensional space. For a given test sample X , SVR develops a predictor function: $h(X) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) K(X, X_i) + b$ by mapping training samples onto the high-dimensional features space. Here, α_i and α_i^* are non-zero Lagrange multipliers, b is a bias constant, and K is the kernel function that represents the inner product of X and train sample X_i .

LASSO is a regularization-based³ linear regression model. The model is selected to minimize the objective function: $\sum_{i=1}^n (w_0 + \sum_{j=1}^m w_j x_j^i - y_i)^2 + \lambda \sum_{j=1}^m w_j$, where λ is the regularization parameter that is used to control the empirical error. As the λ value increases, LASSO changes more coefficients to zero (Wang et al. 2018).

MLR (also referred to as ordinary least squares regression) is one of the oldest and most frequently employed techniques for analyzing the relationship between the response variable and multiple input variables. The general form of the MLR model can be given by $h(X_i) = w_0 + \sum_{j=1}^m w_j x_j^i + \epsilon$, where $h(X_i)$ is the predictor function for the sample $X_i = x_1^i, x_2^i, \dots, x_m^i$, w_0 is the constant, w_j is the coefficient for the variable j , and ϵ is the error term ($\sim N(0, \sigma^2)$) of the model.

5.4 Parameter settings

The detailed parameter settings for the proposed method and benchmarks are presented in this sub-section. In the Md-FKNNreg algorithm, the value for p of the Minkowski distance was selected from $\{1, 1.5, \dots, 5\}$. The number of nearest neighbors k was chosen from the range $\{1, 2, \dots, 25\}$ for all nearest neighbor regression approaches. The value of the fuzzy strength parameter m was kept constant at $m = 1.5$ according to Arif et al. (2010) for both the Md-FKNNreg and FKNNreg models.

The kernel function is the most critical ingredient in the SVR model (Ali and Smith-Miles 2006) because it helps the model achieve robust mapping from training samples to the prediction. Accordingly, we tested the performance of the SVR model using three different kernels: *linear*, *Radial Basis Function (RBF)*, and *polynomial* based on Ali and Smith-Miles (2006). For the Lasso model, the regularization parameter λ was tuned from the range $\{0.001, 0.01, \dots, 100\}$ by following the experiments of Saccoccio et al. (2014). Here, we attempted to create a

³ In machine learning, regularization is a more sophisticated technique that is used to solve model overfitting problems.

Fig. 1 A workflow of the development and evaluation of Md-FKNNreg model

balance for the λ values due to the fact that low values of λ prefer predictor functions that achieve a small training error while larger values tend to obtain simple prediction functions (Wang et al. 2018). With the multiple regression model, we tested four different model types from the toolbox: *linear*, *interaction*⁴, *purequadratic*⁵, and *quadratic*⁶. Using these different types of models, we were able to generate more sophisticated MLR versions for competition in the comparison even though high-order terms of the features were not deployed with the nearest neighbor methods. The rest of the parameter values in the SVR, LASSO, and MLR models were set to the default values according to the toolbox specifications.

5.5 Evaluation metrics

To evaluate the prediction performance of the proposed regression method and benchmarks, we adopted two frequently applied measures: RMSE and R^2 . RMSE computes the square root of the average differences between the model's predictions and the true values. R^2 is the proportion of the variation in the response variable, which is "explained" by the regression model compared to the mean (Kurz-Kim and Loretan 2014). It is a statistical measure that implies how closely the data points in the response variable fit to the values of the regression model. In general, higher R^2 values and smaller RMSE values reflect better performance in the regression model (Pham 2019). The formulas used for both evaluation methods are defined as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2} \quad (7)$$

$$R^2 = \left(1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \right) \times 100\% \quad (8)$$

where n is the number of samples in the test data, $\hat{Y}_i$ indicates the predicted value, Y_i indicates the true value of the i th test sample, and $\bar{Y}$ is the average of the true values. As shown in Eq. (8), the percentage values of R^2 are considered.

When it is necessary to apply several models to a particular problem and pick the best one, the usual method is to use several evaluation metrics to measure the models' performance and select the best model with the highest performance. However, when trying to prove that one model outperforms another for a particular problem, we must use a statistical test of significance and validate the claim of improved performance (Borovicka et al. 2012).

Following Chen et al. (2011), we adopted the paired t test, one of the most commonly used statistical tests in machine learning. This analysis tested the null hypothesis that there is no significant difference between two mean RMSE rates at the 0.05 significant level. Here we considered the error samples from the holdout method (size of 30×1) for each regression model when the optimal parameter values were employed. The standard deviations were also computed.

5.6 Results and discussion

In this subsection, we first present the results of the proposed method compared with the baseline methods from the training and validation step. After that, optimal parameter values observed for each model are discussed. Then, the performances of the fitted models for the training and validation data are evaluated with the test datasets.

⁴ Includes an intercept, linear term for each variable, and all products of pairs of distinct variables.

⁵ Includes an intercept term and linear and squared terms for each variable.

⁶ Includes a constant term, linear and squared terms of each variable, and all products of pairs of distinct variables.

Table 2 The results obtained for all methods in the training & validation step

Data set	Measure	Md-FKNNreg	Man-FKNNreg	Euc-FKNNreg	KNNreg (Biau et al. 2012)	SVR (Drucker et al. 1997)	LASSO (Tibshirani 1996)	MLR (Montgomery et al. 2012)
Stock	RMSE	0.0272	0.0272	0.0276	0.0308	0.0381	0.0854	0.0421
	STD	0.0022	0.0022	0.0023	0.0025	0.0062	0.0029	0.0026
Airfoil	RMSE	0.0932	0.0994	0.0942	0.1027	0.0957	0.1273	0.1113
	STD	0.0047	0.0049	0.0049	0.0046	0.0037	0.0031	0.0030
AutoMPG	RMSE	0.0769	0.0769	0.0812	0.0815	0.0843	0.0952	0.0779
	STD	0.0039	0.0039	0.0050	0.0047	0.0075	0.0045	0.0043
Baseball	RMSE	0.1235	0.1235	0.1254	0.1307	0.1326	0.1276	0.1293
	STD	0.0096	0.0096	0.0092	0.0091	0.0083	0.0072	0.0071
Servo	RMSE	0.1611	0.1611	0.1612	0.1549	0.1818	0.1861	0.1579
	STD	0.0221	0.0221	0.0185	0.0157	0.0296	0.0145	0.0150
Laser	RMSE	0.0407	0.0407	0.0431	0.0451	0.0396	0.0885	0.0438
	STD	0.0083	0.0083	0.0078	0.0084	0.0076	0.0054	0.0050
Qsar Fish	RMSE	0.0966	0.0969	0.0993	0.0981	0.0971	0.1022	0.1021
	STD	0.0036	0.0036	0.0035	0.0039	0.0037	0.0042	0.0042
Parkinson	RMSE	0.0583	0.0583	0.0626	0.0678	0.0810	0.1936	0.1865
	STD	0.0028	0.0028	0.0027	0.0022	0.0036	0.0016	0.0015

5.6.1 Results with the validation data samples

Table 2 summarizes the results of all the methods for each of the datasets from the training and validation step. In the table, we report the minimum RMSE and standard deviation (STD) for the RMSE as the performance measures. Notice that these mean RMSEs and standard deviations are the result of the holdout method with 30 repetitions.

From Table 2, it is apparent that the proposed Md-FKNNreg method outperformed all benchmarks for six datasets and had the second-best performance for the rest of the datasets. Also, the standard deviations of the proposed method were reasonable for all cases. In particular, the Md-FKNNreg method achieved significantly improved performance compared with the Euc-FKNNreg method even though the results of the two methods were comparable in some cases, for example, in the cases of Servo and Stock. Additionally, the Md-FKNNreg and Man-FKNNreg models produced the same results over six datasets. Moreover, the KNNreg and SVR models achieved the lowest mean RMSE results in the cases of Servo and Laser, respectively. Finally, neither the LASSO and nor MLR models offered comparative results for the datasets compared to the other models.

Figure 2 illustrates the R^2 values of each for the optimized regression models from the training and validation step for each dataset. Notice that the R^2 values revealed by the bar-heights in the graphs refer to the maximum of mean R^2 values obtained from the holdout cross-validation. The bar graphs in the figure are also displayed so that the blue

bar represents the highest R^2 while the other bars indicate the rest of the values. From Fig. 2, one can observe a similar indications about the regression performance of the proposed Md-FKNNreg method and benchmarks as from Table 2.

5.6.2 Evaluation of the optimal parameter values

During the training and validation, we evaluated the regression results by tuning the parameters of the Md-FKNNreg and benchmark models. Figure 3 displays the impacts of different combinations of the parameters k and p in the Md-FKNNreg method on the RMSE (and R^2) with the Stock dataset. Figure 3 demonstrates how these parameters maintain the good performance of the Md-FKNNreg method for the Stock data set.

Corresponding to the results in Table 2, the optimal parameter values observed for the proposed Md-FKNNreg and benchmark models are presented in Table 3. The table shows that $p = 1$ (i.e., the Manhattan distance) obtained the best performance with the Md-FKNNreg method for the majority of the datasets. We also applied the Manhattan distance-based FKNNreg and received the same results as for the Md-FKNNreg method with $p = 1$. This finding is consistent with the implications of Aggarwal et al. (2001), who showed that the Manhattan distance (L_1 norm) is the best option for high-dimensional applications. Additionally, it seems that relatively low k values (varying from 1 to 13) are better suited to the original KNNreg method, while high k values (ranging from 2 to 25) work

Fig. 2 Observed R^2 values of each model for each dataset

Fig. 3 The averages of the RMSE and R^2 values for the Md-FKNNreg method for different parameter combinations (p , k) in the training and validation step with the Stock dataset

Table 3 Optimal parameter values of each model for each dataset

Model	Parameters	Stock	Airfoil	AutoMPG	Baseball	Servo	Laser	Qsar fish	Parkinson
Md-FKNNreg	(k, p)	(9, 1)	(5, 4.5)	(16, 1)	(8, 1)	(8, 1)	(4, 1)	(13, 1.5)	(4, 1)
Man-FKNNreg	k	9	2	16	8	8	4	15	4
Euc-FKNNreg	k	9	11	25	13	6	4	25	5
KNNreg (Biau et al. 2012)	k	2	1	7	13	9	3	10	3
SVR (Drucker et al. 1997)	kernel	RBF	RBF	RBF	linear	polynomial	RBF	RBF	RBF
LASSO (Tibshirani 1996)	λ	0.001	0.001	0.010	0.001	0.001	0.001	0.001	0.001
MLR (Montgomery et al. 2012)	Model	Quadratic	Quadratic	Interaction	Linear	Pure-quadratic	Quadratic	Linear	Linear

better for its fuzzy versions. Moreover, the RBF kernel appears to be the most suitable for the SVR model because it achieved high performance for all datasets except for Baseball and Servo data. Regarding the LASSO model having the most instances of the least-squares estimations, we believe that this is because of the lowest $\lambda = 0.001$ shows the optimum. The optimal type of MLR model varied depending on the particular datasets. By taking the parameters and other results together, it is evident that the fuzzy variants have more potential for linear and non-linear problems. For instance, the case with the Baseball data could be considered a linear problem since both SVR and MLR hold linear parameters, but high performance was achieved with the Md-FKNNreg, Man-FKNNreg, and Euc-FKNNreg methods.

5.6.3 Results with the test data samples

This subsection presents the regression results of the Md-FKNNreg and baseline models with the testing data samples that were initially split from the original datasets. In this testing phase, we used the optimized parameter values and training data samples stored from the cross-validation step to evaluate the models' performances with the previously unseen test data samples.

Table 4 summarizes the results with the mean RMSEs and the standard deviations (STD) of the proposed Md-FKNNreg and benchmarks models for the selected datasets. In addition, the average computational time (Com. time, in seconds) of each method in the testing phase is also reported.

The results of Table 4 show that the Md-FKNNreg method outperformed all benchmarks for all datasets, verifying the effectiveness of the proposed Md-FKNNreg

Table 4 The results obtained for all methods in the testing step

Data set	Measure	Md-FKNNreg	Man-FKNNreg	Euc-FKNNreg	KNNreg (Biau et al. 2012)	SVR (Drucker et al. 1997)	LASSO (Tibshirani 1996)	MLR (Montgomery et al. 2012)
Stock	RMSE mean	0.0294	0.0294	0.0302	0.0311	0.0406	0.0762	0.0407
	STD	0.0016	0.0016	0.0017	0.0017	0.0041	0.0007	0.0011
	Com. time	0.0230	0.0171	0.0200	0.0207	0.0009	0.0002	0.0082
Airfoil	RMSE mean	0.0963	0.0966	0.1002	0.1036	0.0986	0.1342	0.1182
	STD	0.0046	0.0039	0.0046	0.0048	0.0022	0.0010	0.0013
	Com. time	0.0428	0.0306	0.0264	0.0233	0.0009	0.0001	0.0030
AutoMPG	RMSE mean	0.0687	0.0687	0.0719	0.0728	0.0707	0.0824	0.0725
	STD	0.0024	0.0024	0.0023	0.0028	0.0036	0.0015	0.0020
	Com. time	0.0098	0.0066	0.0075	0.0066	0.0004	0.0002	0.0020
Baseball	RMSE mean	0.1184	0.1184	0.1239	0.1316	0.1448	0.1329	0.1350
	STD	0.0080	0.0080	0.0082	0.0070	0.0064	0.0043	0.0050
	Com. time	0.0068	0.0052	0.0060	0.0051	0.0004	0.0001	0.0003
Servo	RMSE mean	0.1120	0.1120	0.1231	0.1167	0.1577	0.1602	0.1197
	STD	0.0210	0.0210	0.0237	0.0164	0.0218	0.0038	0.0107
	Com. time	0.0062	0.0060	0.0062	0.0054	0.0014	0.0003	0.0051
Laser	RMSE mean	0.0435	0.0435	0.0441	0.0527	0.0483	0.0986	0.0509
	STD	0.0094	0.0094	0.0124	0.0106	0.0158	0.0026	0.0058
	Com. time	0.0237	0.0168	0.0192	0.0175	0.0004	0.0002	0.0030
Qsar Fish	RMSE mean	0.0902	0.0905	0.0917	0.0942	0.0943	0.0976	0.0973
	STD	0.0021	0.0021	0.0027	0.0020	0.0026	0.0010	0.0009
	Com. time	0.0249	0.0058	0.0201	0.0046	0.0005	0.0001	0.0003
Parkinson	RMSE mean	0.0566	0.0566	0.0608	0.0666	0.0786	0.1915	0.1844
	STD	0.0025	0.0025	0.0027	0.0027	0.0028	0.0005	0.0031
	Com. time	0.2296	0.2229	0.2971	0.2815	0.0178	0.0039	0.0101
Overall	RMSE mean	0.0769	0.0770	0.0807	0.0837	0.0917	0.1217	0.1023

method for regression problems. Compared with the Man-FKNNreg results, the Md-FKNNreg model achieved somewhat higher accuracy for the Airfoil and Qsar Fish datasets. Additionally, the Md-FKNNreg showed significantly better performance than the Euc-FKNNreg method across almost all datasets. This reveals that introducing the Minkowski distance in the learning part can result in finding more reasonable nearest neighbors, leading to better performance compared to the Manhattan and Euclidean distances. Considering the KNNreg and SVR models, even though these achieved the lowest errors in some cases during the training and validation, the Md-FKNNreg method outperformed them on all testing cases.

This proves that the proposed Md-FKNNreg model has more power to overcome over-fitting issues than both KNNreg and SVR models. Moreover, the lowest regression performance for the testing data, similar to the validation data, occurred for both the LASSO and MLR models.

Based on the testing times, it is clear that the computational complexity of the FKNNreg methods was relatively high compared with the other methods used. This might be because of the inclusion of the weight generation process based on the inverse of the distances and the fuzzy strength parameter. Additionally, the proposed Md-FKNNreg method required more time (in seconds) than the Euc-FKNNreg and Man-FKNNreg methods to deliver the

results since it includes an additional computation with the parameter p .

To validate the test results statistically, a paired t test was applied, and the observed results for the P values and test statistics are presented in Table 5. The t test results demonstrate whether there is a statistically significant difference between mean RMSEs of the compared regression methods. From the evidence in the table, it is apparent that

the Md-FKNNreg method yielded statistically significantly better performance than the benchmarks in terms of the lowest error. In particular, we observed that the proposed method could not produce statistically significant results for either the Servo dataset (compared with the Euc-FKNNreg and KNNreg methods) or the Laser dataset (compared with the Euc-FKNNreg and SVR methods). It should be noted here that we did not find a statistically

Table 5 Paired t test results on the performance of the Md-FKNNreg method vs. five benchmarks for the test datasets

Data set	Paired t with Md-FKNNreg	P value	Test statistic
Stock	Euc-FKNNreg	0.0463	Significant
	KNNreg (Biau et al. 2012)	1.0715e−04	Significant
	SVR (Drucker et al. 1997)	3.8055e−20	Significant
	LASSO (Tibshirani 1996)	1.7732e−76	Significant
	MLR (Montgomery et al. 2012)	1.8274e−38	Significant
Airfoil	Euc-FKNNreg	0.0014	Significant
	KNNreg (Biau et al. 2012)	4.2915e−08	Significant
	SVR (Drucker et al. 1997)	0.0131	Significant
	LASSO (Tibshirani 1996)	3.3463e−50	Significant
	MLR (Montgomery et al. 2012)	2.3700e−36	Significant
AutoMPG	Euc-FKNNreg	2.7216e−06	Significant
	KNNreg (Biau et al. 2012)	8.6347e−08	Significant
	SVR (Drucker et al. 1997)	0.0141	Significant
	LASSO (Tibshirani 1996)	2.1512e−34	Significant
	MLR (Montgomery et al. 2012)	1.6638e−08	Significant
Baseball	Euc-FKNNreg	0.0101	Significant
	KNNreg (Biau et al. 2012)	5.3335e−09	Significant
	SVR (Drucker et al. 1997)	1.9285e−20	Significant
	LASSO (Tibshirani 1996)	3.1952e−12	Significant
	MLR (Montgomery et al. 2012)	1.1241e−13	Significant
Servo	Euc-FKNNreg	0.0600	Not significant
	KNNreg (Biau et al. 2012)	0.3348	Not significant
	SVR (Drucker et al. 1997)	2.2466e−11	Significant
	LASSO (Tibshirani 1996)	6.2441e−18	significant
	MLR (Montgomery et al. 2012)	0.0778	Not significant
Laser	Euc-FKNNreg	0.8251	Not significant
	KNNreg (Biau et al. 2012)	7.6651e−04	Significant
	SVR (Drucker et al. 1997)	0.1544	Not significant
	LASSO (Tibshirani 1996)	8.6754e−38	Significant
	MLR (Montgomery et al. 2012)	4.9310e−04	Significant
Qsar Fish	Euc-FKNNreg	0.0155	Significant
	KNNreg (Biau et al. 2012)	3.0113e−10	Significant
	SVR (Drucker et al. 1997)	9.4806e−09	Significant
	SVR (Drucker et al. 1997)	1.3310e−24	significant
	MLR (Montgomery et al. 2012)	5.3260e−24	Significant
Parkinson	Euc-FKNNreg	3.7463e−08	Significant
	KNNreg (Biau et al. 2012)	1.9184e−21	Significant
	SVR (Drucker et al. 1997)	2.0103e−38	Significant
	LASSO (Tibshirani 1996)	2.2655e−93	Significant
	MLR (Montgomery et al. 2012)	6.3721e−81	Significant

significant difference between Md-FKNNreg and Man-FKNNreg results for any dataset. In addition to finding no difference between the test results of Md-FKNNreg and Man-FKNNreg for six datasets, the t test produced no evidence of a significant difference between these methods for the Airfoil and Qsar Fish cases.

6 Conclusions

This paper proposed a new generalized regression model based on the FKNN rule and investigated its effectiveness on different regression problems from various domains. The Minkowski distance metric was introduced into the nearest neighbor search in the proposed algorithm to examine how it improves accuracy. Accordingly, the proposed method was named the Md-FKNNreg model. The effectiveness of the Md-FKNNreg method was evaluated in comparative experiments with the standard nearest neighbor and three well-known regression methods, namely SVR, LASSO, and MLR. In addition, Man-FKNNreg and Euc-FKNNreg methods were implemented, and their results were compared. For the experiments, we used eight real-world datasets that are freely available from machine learning repositories. The regression performance of each model for those datasets were discussed in terms of the RMSE, R^2 and standard deviation. The results showed that the Md-FKNNreg method outperformed the benchmarks and is a suitable choice for regression problems. In our experiments, Md-FKNNreg gave the lowest overall average RMSE of 0.0769.

The results of the experiments showed that the proposed Md-FKNNreg method achieved statistically significantly higher performance than the other methods in almost all cases in terms of the RMSE means. Additionally, we found that the Minkowski distance with $p = 1$ yielded the optimal Md-FKNNreg model, which then achieved the best performance for the majority of the datasets. In other words, the Man-FKNNreg showed promising results for the regression at large, supporting the indications in the study by Aggarwal et al. (2001).

However, it should be noted that the computational complexity of the proposed method is relatively high compared with the Euc-FKNNreg and KNNreg methods because an additional calculation with the Minkowski distance is included in the learning algorithm. Despite that, this research has offered some insight into further investigations. For example, it would be interesting to see how the Md-FKNNreg method adapts and performs in regression applications in which KNNreg was previously utilized (e.g., Hu et al. 2014; Cai et al. 2020; Yao and Ruzzo 2006; Huang and Perry 2016; Zhou et al. 2020; Durbin et al.

2021; Dell'Acqua et al. 2015). Furthermore, future research directions may also test the effect of combining Md-FKNNreg with other efficient variants, such as SVM (Chen and Hao 2017) and ANN (Salari et al. 2015) in a hybrid framework.

Funding Open access funding provided by LUT University (previously Lappeenranta University of Technology (LUT)).

Availability of data and materials The data used in the manuscript are freely available from the UCI and KEEL repositories.

Code availability The code is available in <https://github.com/MahindaMK/A-generalized-fuzzy-k-nearest-neighbor-regression-model-based-on-Minkowski-distance>.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adege AB, Yayeh Y, Berie G, Lin H, Yen L, Li YR (2018) Indoor localization using k-nearest neighbor and artificial neural network back propagation algorithms. In: 27th Wireless and Optical Communication Conference (WOCC), pp 1–2
- Aggarwal CC, Hinneburg A, Keim DA (2001) On the surprising behavior of distance metrics in high dimensional space. Database Theory ICDT 2001. Springer, Berlin, pp 420–434
- Alcala-Fdez J, Fernandez A, Luengo J, Derrac J, García S, Sánchez L, Herrera F (2011) Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J Mult-Valued Log Soft Comput 17:255–287, <https://sci2s.ugr.es/keel/datasets.php>
- Ali S, Smith-Miles KA (2006) A meta-learning approach to automatic kernel selection for support vector machines. Neurocomputing 70:173–186
- Arif M, Akram MU, Minhas FA (2010) Pruned fuzzy k-nearest neighbor classifier for beat classification. J Biomed Sci Eng 3:380–3899
- Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Stat Surv 4:40–79
- Benedetti JK (1977) On the nonparametric estimation of regression functions. J R Stat Soc Series B 39:248–253

- Bergamasco LCC, Nunes FLS (2019) Intelligent retrieval and classification in three-dimensional biomedical images—a systematic mapping. *Comput Sci Rev* 31:19–38
- Biau G, Devroye L, Dujmović V, Krzyżak A (2012) An affine invariant k-nearest neighbor regression estimate. *J Multivar Anal* 112:24–34
- Borovicka T, Jirina MJ, Kordik P, Jirina M (2012) Selecting representatives data sets *Advances in Data Mining Knowledge Discovery and Applications*. Rijeka, Croatia, pp 43–70
- Buza K, Nanopoulos A, Nagy G (2015) Nearest neighbor regression in the presence of bad hubs. *Knowl Based Syst* 86:250–260
- Cai L, Yu Y, Zhang S, Song Y, Xiong Z, Zhou T (2020) A sample-rebalanced outlier-rejected k-nearest neighbor regression model for short-term traffic flow forecasting. *IEEE Access* 8:22686–22696
- Chang H, Yeung DY, Cheung WK (2006) Relaxational metric adaptation and its application to semi-supervised clustering and content-based image retrieval. *Pattern Recognit* 39:1905–1917
- Chen SM, Chang YC (2010) Multi-variable fuzzy forecasting based on fuzzy clustering and fuzzy rule interpolation techniques. *Inf Sci* 180:4772–4783
- Chen S, Chen L (2007) A fuzzy hierarchical clustering method for clustering documents based on dynamic cluster centers. *J Chin Inst Eng* 30:169–172
- Chen Y, Hao Y (2017) A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Syst Appl* 80:340–355. <https://doi.org/10.1016/j.eswa.2017.02.044>
- Chen SM, Hsiao HR (2005) A new method to estimate null values in relational database systems based on automatic clustering techniques. *Inf Sci* 169:47–69
- Chen J, Lau HYK (2016) Learning the inverse kinematics of tendon-driven soft manipulators with k-nearest neighbors regression and gaussian mixture regression. In: 2nd International conference on control, automation and robotics (ICCAR), pp 103–107
- Chen HL, Liu DY, Yang B, Wang SJ (2011) An adaptive fuzzy k-nearest neighbor method based on parallel particle swarm optimization for bankruptcy prediction. In: *Lecture Notes in computer science* 6634 LNAI (PART 1), pp 249–264
- Chen SM, Ke JS, Chang JF (1990) Knowledge representation using fuzzy petri nets. *IEEE Trans Knowl Data Eng* 2:311–319
- Chen SM, Wang NY, Pan JS (2009) Forecasting enrollments using automatic clustering techniques and fuzzy logical relationships. *Expert Syst Appl* 36:11070–11076
- Chen HL, Huang CC, Yu XG, Xu X, Sun X, Wang G, Wang SJ (2013) An efficient diagnosis system for detection of parkinson's disease using fuzzy k-nearest neighbor approach. *Expert Syst Appl* 40(1):263–271
- Cheng PE (1984) Strong consistency of nearest neighbor regression function estimators. *J Multivar Anal* 15:63–72
- Cheng CH, Chan CP, Sheu YJ (2019) A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. *Eng Appl Artif Intell* 81:283–299
- Cordeiro R, Makarenkov V (2016) Applying subclustering and lp distance in weighted k-means with distributed centroids. *Neurocomputing* 173:700–707. <https://doi.org/10.1016/j.neucom.2015.08.018>
- Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20:273–297
- Cover T, Hart P (1967) Nearest neighbor pattern classification. *IEEE Trans Inf Theory* 13:21–27
- Dell'Acqua P, Bellotti F, Berta R, De Gloria A (2015) Time-aware multivar. nearest neighbor regression methods for traffic flow prediction. *IEEE Trans Intell Transp Syst* 16:3393–3402
- Dettmann E, Becker C, Schmeiser C (2011) Distance functions for matching in small samples. *Comput Stat Data Anal* 55:1942–1960
- Dheeru D, Taniskidou EK (2017) Uci machine learning repository. <http://archive.ics.uci.edu/ml>
- Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. *Neural Inf Proc Syst* 9:155–161
- Durbin M, Wonders MA, Flaska M, Lintereur AT (2021) K-nearest neighbors regression for the discrimination of gamma rays and neutrons in organic scintillators. *Nucl Instrum Methods Phys Res A* 987:164826
- Gueorguieva N, Valova I, Georgiev G (2017) M&mfc: Fuzzy c-means clustering with mahalanobis and minkowski distance metrics. *Proc Comput Sci* 114:224–233. <https://doi.org/10.1016/j.procs.2017.09.064>
- Guillen A, Herrera LJ, Rubio G, Pomares H, Lendasse A, Rojas I (2010) New method for instance or prototype selection using mutual information in time series prediction. *Neurocomputing* 73:2030–2038
- Györfi L, Kohler M, Krzyżak A, Walk H (2002) A distribution free theory of nonparametric regression. Springer, New York
- Hornig YJ, Chen SM, Chang YC, Lee CH (2005) A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques. *IEEE Trans Fuzzy Syst* 13:216–228
- Hu C, Jain G, Zhang P, Schmidt C, Gomadam P, Gorka T (2014) Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery. *Appl Energy* 129:49–55
- Huang J, Perry M (2016) A semi-empirical approach using gradient boosting and k-nearest neighbors regression for GEFCom2014 probabilistic solar power forecasting. *Int J Forecast* 32:1081–1086
- Huo J, Ma Y, Lu C, Li C, Duan K, Li H (2021) Mahalanobis distance based similarity regression learning of nirs for quality assurance of tobacco product with different variable selection methods. *Spectrochimica Acta A Mol Biomol Spectrosc* 251:119364
- Jenicka S, Suruliandi A (2011) Empirical evaluation of distance measures for supervised classification of remotely sensed image with modified multivariate local binary pattern. In: *International conference on emerging trends in electrical and computer technology (ICETECT)*, pp 762–767
- Kaski S, Sinkkonen J, Peltonen J (2001) Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Trans Neural Netw Learn Syst* 12(4):936–947
- Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbor algorithm. *IEEE Trans Syst Man Cybern Syst* 15:580–585
- Koloseni D, Lampinen J, Luukka P (2012) Optimized distance metrics for differential evolution based nearest prototype classifier. *Expert Syst Appl* 39(12):10564–10570
- Koloseni D, Lampinen J, Luukka P (2013) Differential evolution based nearest prototype classifier with optimized distance measures for the features in the data set. *Expert Syst Appl* 40(10):4075–4081
- Kramer O (2011) Dimensionality reduction by unsupervised K-nearest neighbor regression. In: *Proceedings of the 10th International Conference on Machine Learning and Applications, ICMLA*, pp 275–278
- Kumbure MM, Luukka P, Collan M (2019) An enhancement of fuzzy k-nearest neighbor classifier using multi-local power means. In: *Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)*, Atlantis Press, pp 83–90
- Kumbure MM, Luukka P, Collan M (2020) A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean. *Pattern*

- Recognit Lett 140:172–178. <https://doi.org/10.1016/j.patrec.2020.10.005>
- Kurz-Kim JR, Loretan M (2014) On the properties of the coefficient of determination in regression models with infinite variance variables. *J Econ* 181:15–24
- Liu X, Beyrend-Dur D, Dur G, Ban S (2013) Effects of temperature on life history traits of *Eodiaptomus japonicus* (copepoda: Calanoida) from lake biwa (japan). *Limnology* 15:85–97
- Moghtadaiee V, Dempster AG (2015) Determining the best vector distance measure for use in location fingerprinting. *Pervasive Mob Comput* 23:59–79. <https://doi.org/10.1016/j.pmcj.2014.11.002>
- Montgomery DC, Peck EA, Vining GG (2012) Introduction to linear regression analysis. John Wiley & Sons, Hoboken
- Nguyen B, Morell C, Baets BD (2016) Large-scale distance metric learning for k-nearest neighbors regression. *Neurocomputing* 214:805–814
- Nikoo MR, Kerachian R, Alizadeh MR (2018) A fuzzy knn-based model for significant wave height prediction in large lakes. *Oceanologia* 60:153–168
- Pham H (2019) A new criterion for model selection. *Mathematics* 7:1215
- Ranmya R, Sasikala T (2019) An efficient minkowski distance-based matching with merkle hash tree authentication for biometric recognition in cloud computing. *Soft Comput* 23:13423–13431
- Rastin N, Jahromi MZ, Taheri M (2021) A generalized weighted distance k-nearest neighbor for multi-label problems. *Pattern Recognit* 114:107526. <https://doi.org/10.1016/j.patcog.2020.107526>
- Rodrigues EO (2018) Combining minkowski and chebyshev: new distance proposal and survey of distance metrics using k-nearest neighbours classifier. *Pattern Recognit Lett* 110:66–71
- Saccoccio M, Wan TH, Chen C, Ciucci F (2014) Optimal regularization in distribution of relaxation times applied to electrochemical impedance spectroscopy: ridge and lasso regression methods - a theoretical and experimental study. *Electrochim Acta* 147:470–482
- Salari N, Shohaimi S, Najafi F, Nallappan M, Karishnarajah I (2015) Time-aware multivar. nearest neighbor regression methods for traffic flow prediction. *IEEE Trans Intell Transp Syst* 16:3393–3402
- Shirshorshidi AS, Aghabozorgi S, Wah TY (2015) A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLOS ONE* 10(12):1–20
- Song Y, Liang J, Lu J, Zhao X (2017) An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing* 251:26–34
- Stone CJ (1977) Consistent nonparametric regression. *Ann Stat* 5:595–645
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol* 58:267–288
- Turner T (1977) Exploratory data analysis. Addison Wesley, Reading
- Wang S, Ji B, Zhao J, Liu W, Xu T (2018) Predicting ship fuel consumption based on LASSO regression. *Transp Res D Transp* 65:817–824
- Yao Z, Ruzzo W (2006) A regression-based k nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinformatics* 7:S11
- Yu S, De Backer S, Scheunders P (2002) Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for hyperspectral satellite imagery. *Pattern Recognit Lett* 23(1):183–190
- Zadeh LA (1965) Fuzzy sets. *Inf and Control* 8:338–353
- Zeng S, Chen SM, Teng MO (2019) Fuzzy forecasting based on linear combinations of independent variables, subtractive clustering algorithm and artificial bee colony algorithm. *Inf Sci* 484:350–366
- Zhou Y, Huang M, Pecht M (2020) Remaining useful life estimation of lithium-ion cells based on k- nearest neighbor regression with differential evolution optimization. *J Clean Prod* 249:119409. <https://doi.org/10.1016/j.jclepro.2019.119409>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.