A Brief Review on Instance Selection Based on Condensed Nearest Neighbors for Data Classification Tasks

Yasmany Fernández-Fernández, Diego H. Peluffo-Ordóñez, Ana C. Umaquinga-Criollo, Leandro L. Lorente-Leyva, and Elia N. Cabrera-Alvarez

Abstract The condensed nearest neighbor (CNN) classifier is one of the techniques used and known to perform recognition tasks. It has also proven to be one of the most interesting algorithms in the field of data mining despite its simplicity. However, CNN suffers from several drawbacks, such as high storage requirements and low noise tolerance. One of the characteristics of CNN is that it focuses on the selection of prototypes, which consists of reducing the set of training data. One of the goals of CNN seeks to achieve the reduction of information in such a way that the reduced information can represent large amounts of data to exercise decision-making on them. This paper mentions some of the most recent contributions to CNN-based unsupervised algorithms in a review that builds on the mathematical principles of condensed methods.

Keywords Prototypes · Nearest neighbor algorithms · Classification

Y. Fernández-Fernández (✉)
Universidad Politécnica Estatal del Carchi, Tulcán, Ecuador
e-mail: yasmany.fernandez@upec.edu.ec

Y. Fernández-Fernández · D. H. Peluffo-Ordóñez · A. C. Umaquinga-Criollo · L. L. Lorente-Leyva
SDAS Research Group, Ibarra, Ecuador
e-mail: dpeluffo@yachaytech.edu.ec

A. C. Umaquinga-Criollo
e-mail: acumaquinga@utn.edu.ec

L. L. Lorente-Leyva
e-mail: leandro.lorente@sdas-group.com

D. H. Peluffo-Ordóñez
Yachay Tech University, Urcuquí, Ecuador

Corporación Universitaria Autónoma de Nariño, Pasto, Colombia

A. C. Umaquinga-Criollo
Universidad Técnica del Norte, Ibarra, Ecuador

E. N. Cabrera-Alvarez
Universidad de Cienfuegos, Cienfuegos, Cuba
e-mail: elita@ucf.edu.cu

V. Bindhu et al. (eds.), *International Conference on Communication, Computing and Electronics Systems*, Lecture Notes in Electrical Engineering 733,
https://doi.org/10.1007/978-981-33-4909-4_23

1 Introduction

Instance selection methods represent an important approach in different areas of data science. In [1], some important elements are considered in topics related to instance selection. There are two important processes, namely training set selection and prototype selection.

The selection of a subset of data for another very large data set is summarized in the concept of "data condensation" [2]. This form of data reduction differs from the others and is integrated as one of the families of the instance selection methods. Mainly, data condensation approaches are studied based on the classification processes, particularly the k-nearest neighbor (KNN) methods which refer to obtain a consistent minimum set that classifies the entire original set. Figure 1 shows a simple representation of the KNN.

One of the first pioneering methods in the analysis in the data structure for the selection of instances was CNN [4]. The methods of condensation of data that are not related to the classification process are also known as methods of condensation of generic data, such condensation is performed through the so-called vector quantization (VQ), and example of this is the self-organization map and other ways of organizing the data as shown in Fig. 2.

1.1 Vector Quantization

Vector quantization (VQ) is a classic method that consists of approximating a continuous probability density function $p(x)$ of the vector input variable x by using a finite number of book-encoded vectors mi, $i = 1, 2, \ldots, k$; once these book-encoders have been chosen, the approximation of x implies finding the reference vector closest to x. An optimal location type of m minimizes to E where E is the rth power of the reconstruction error [5]:

$$E = \int \|x - m_c\|^r p(x) \mathrm{d}x \tag{1}$$

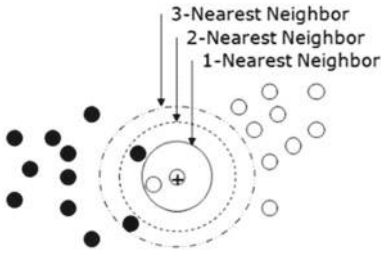

Fig. 1 K-nearest neighbor representation [3]

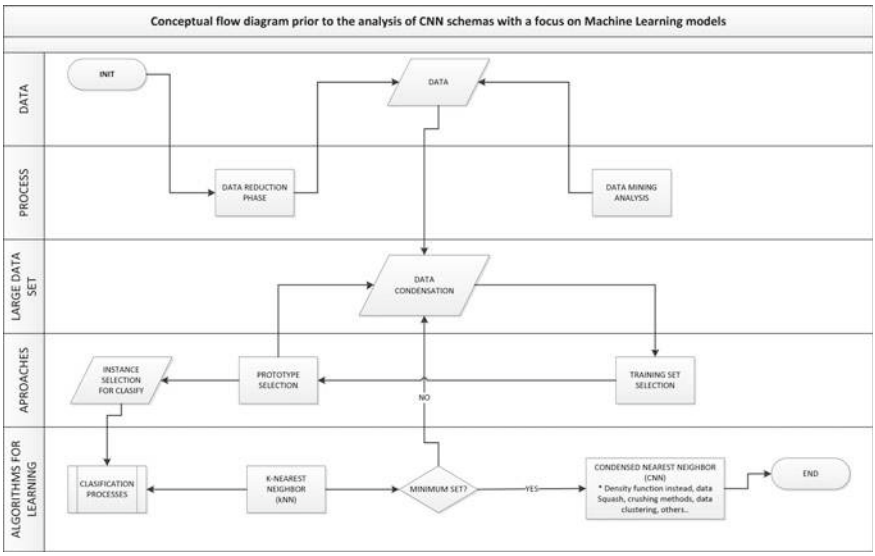

Fig. 2 CNN decision diagram for data reduction task

where dx represents the differential volume in the space x and the index $c = c\,(x)$ of the best match between the book-encoders (winner) is a function of the input vector:

$$\|x - m_c\| = \underbrace{\min}_{i}\{\|x - m_i\|\} \tag{2}$$

In general, a closed solution for the optimal location of m is not possible, so iterative approximation schemes can be used.

1.2 Condensed Methods

Generic data condensation methods are based on techniques that consider density; they consider the density function instead of minimizing the quantification error; that is, for a specific input set, the condensed output set [6] is established.

Other methods such as data squash or data clustering are used for sample selection. A crushing method seeks the compression of the data in such a way that a statistical analysis performed on the compressed data obtains the same result as with the original data. Clustering-based algorithms [7, 8] divide data into samples like each other and different from examples of data belonging to other groups [1].

Figure 3 is represented according to a distance function where the quality of the cluster could be measured according to the dimension of its diameter which is the maximum between two samples belonging to the same group.

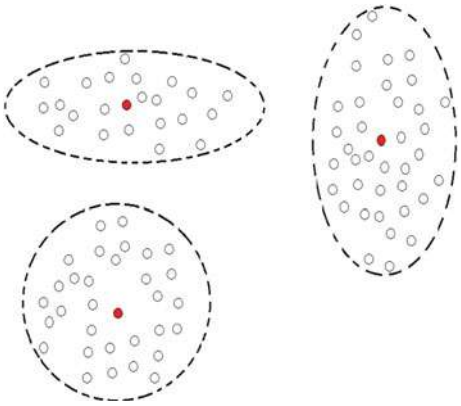

Fig. 3 Three clusters obtained from a set of two-dimensional data

1.3 Machine Learning and Feature Selection

In machine learning, a process known as feature selection consists of the selection of characteristics, attributes or selection of variable subsets for use in model building. In [2], two feature selection strategies are mentioned, the first based on feature ranking and the other based on best subset selection. In the case of the methods based on feature ranking, some statistical metrics are used, some of the simple complexity uses the correlation coefficient instead of other more complex used methods such as the Gini index, and this index can be used to quantify inequalities in variable distributions. Other feature ranking methods mentioned in the literature [9] are the bivariate and multivariate methods; these methods calculate the distance between the actual joint distributions of the characteristics of two or more variables and answer the question of what the joint distribution would be if these variables were independent, further. The joint distribution represents the probability distribution of existing case studies. Among the multivariate analysis, methods are the stepwise linear regression [10, 11] which has been used in cluster tasks and sample selection [12]; other slightly more complex algorithms include the use of machine learning and advanced statistics, for example, partial least squares regression [13] and sensitivity analysis [14]. Also, in performance analysis of virtual clusters [15] and architecture in wireless networks [16].

The second strategy based on subset selection has its focus on the selection of a subset for the selection of characteristics or attributes that have a significant effect on the prediction of a variable. The classic methods of data reduction and sample selection [17] mention its importance given the analysis of large amounts of data for each sample and the time consumed which may cause an over-adjustment of the model of training.

In all the approaches seen so far in a very simple way, the importance of selecting a suitable sample has been evidenced to reduce computational cost and time among other aspects. From now on, the various efforts made to obtain results using the

CNN method [18] with the prototype approach that facilitates the machine learning approach [19] will be more rigorously required.

The rest of this paper is structured as follows: Sect. 2 presents the theoretical background and overview referring to the main problem by the CNN method. Section 3 describes more practically by introducing the idea of the use of metrics in unsupervised learning and its relationship with CNN. Finally, the conclusions are presented in Sect. 4.

2 Theoretical Background and Overview

In practical problems, one of the most important elements to handle is the elimination of noise, redundancies, useless instances and therefore the selection of prototypes, constituting the first step for any practical application.

2.1 Problem Definition

It is desired to isolate the smallest set of instances that could predict the class with the same or greater precision than the original set [20]:

Lemma 2.1.1. *Let X_p be an instance where $X_p = \left(X_{p1}, X_{p2}, ..., X_{pm}, X_{pc}\right)$, with $X_p \in c$ given by X_{pc} and a $X_{pi} \in R^m$ being the value of the ith feature of the p_{th} sample. A training set TR, and also the N instances X_p and a validation set TS with t instances X_p, is obtained. $S \subset TR$ is the subset of the selected samples that resulted from applying an instance selection algorithm.*

Summarizing Lemma 2.1.1., *the objective of an instance selection method is to obtain a subset $S \subset T$ such that S does not contain unnecessary instances* [21]*:*

$$\text{Acc}(S) \cong \text{Acc}(X) \tag{3}$$

where Acc(X) is the qualifier of the training set X.

2.2 Prototype-Based Approach on Unsupervised Learning

Models based on prototype analysis represent several appealing concepts such as the explicit representation of observations, data or typical representatives that exhibit some relation to psychology and neuroscience.

In Sect. 1, the relationship between condensation methods and vector quantization was approached in a very simple way, and this subsection discusses how to prototype

selection matches the instance selection approach with a competitive perspective in unsupervised learning [18].

The vector quantization mathematical statement is formulated in terms of a function that represents costs and generally guides the computation of prototype vectors. A prototype-based representation [22] of a given set of P is defined in Lemma 2.2.1.

Lemma 2.2.1. *Assign the representation of a set of P feature vectors* $\{x^{\mu} \in \mathbb{R}^n\},\ \mu = 1, 2, ..., P$ *that represent a particular input values.*

A popular approach considers the assignment of any data point to the closest prototype, the so-called winner in the set $W = \left\{w^1, w^2, \ldots, w^K\right\}$ *in terms of a predefined distance measure.*

Using the Euclidean metric in feature space with:

$$d^2(x, y) = (x - y)^2 \text{ for } x, y \in \mathbb{R}^N \tag{4}$$

Having the quantization error [3] as the corresponding cost function:

$$H_{\mathrm{VQ}} = \sum_{i=1}^{P} \frac{1}{2} d^2\left(w^*(x^{\mu}), x^{\mu}\right) \tag{5}$$

where $w^*(x^{\mu}) \in \mathbb{R}^N$ denote the closest prototype using a Euclidean metric $x^{\mu} \in \mathbb{R}^n$:

$$d\left(w^*(x^{\mu}), x^{\mu}\right) \leq d\left(w^j, x^{\mu}\right) \text{ for all } j = 1, 2, \ldots, K \tag{6}$$

The quantization error quantifies the fidelity with which the set of prototypes represent data.

2.3 The Condensed Nearest Neighbor Rule (CNN Rule)

An in-depth study on the pillars that support the CNN method [23] and that will be specified below:

Let $\left(X_1^{'}, Y_1^{'}\right) \ldots \left(X_m^{'}, Y_m^{'}\right)$ be a sequence that depends somehow on the data D_n, and let g_n be the 1-nearest neighbor rule with $\left(X_1^{'}, Y_1^{'}\right) \ldots \left(X_m^{'}, Y_m^{'}\right)$ where m is previously set. One way to find the data is to find the subset of the size m data, for the remained minimal $n - m$ data is confirmed by the error with the I-NN rule (this is known as Hart's rule).

If:

$$\hat{L}_n = \left(\frac{1}{n}\right) \sum_{i=1}^{n} I_{\{g_n(X_i) \neq Y_i\}} \tag{7}$$

And:

$$L_n = P\{g_n(X) \neq Y | D_n\} \tag{8}$$

Then, we have the following:

Lemma 2.3.1. $\forall\, \varepsilon > 0$,

$$P\left\{|L_n - \hat{L}_n| \geq \varepsilon\right\} \leq 8e^{-\frac{n\varepsilon^2}{32}} \left(\frac{ne}{d+1}\right)^{(d+1)m(m-1)} \tag{9}$$

where $\hat{L}_n$*is about the estimate error probability.*

Observe that:

$$\hat{L}_n = \left(\frac{1}{n}\right) \sum_{i=1}^{n} I_{\left\{(X_j, Y_j) \notin \bigcup_{i=1}^{m} B_i \times \left\{Y_i^{'}\right\}\right\}} \tag{10}$$

where B_i is the Voronoi cell of $X_i^{'}$ corresponding to $X_1^{'} \ldots X_m^{'}$, where $B_i \subset R^d$ is the closer partition to $X_i^{'}$ than to any other $X_j^{'}$:

$$L_n = P\left\{(X, Y) \notin \bigcup_{i=1}^{m} B_i \times \left\{Y_i^{'}\right\} | D_n\right\} \tag{11}$$

Using simple upper bound:

$$|L_n - \hat{L}_n| \leq \underbrace{\text{Sup}}_{A \in A_m} |v_n(A) - v(A)| \tag{12}$$

where v denotes the measure of (X, Y), v_n is some measure and A_m refer a set of all subsets of $R^d \times \{0, 1\}$ of the form $\bigcup_{i=1}^{m} B_i \times \{y_i\}$ where $B_1, \ldots, B_m$ are Voronoi's cells corresponding to $x_1, \ldots, x_m, x_i \in R^d$, $y_i \in \{0, 1\}$.

Using the Vapnik–Chervonenkis inequality [24]:

$$s(A_m, n) \leq s(A, n)^m \tag{13}$$

Such that A is the class of sets $B_1 \times \{y_1\}$ and each set in A intercepts in at most $m - 1$ hyperplanes. Then:

$$s(A, n) \leq \underbrace{Sup}_{n_0, n_1 : n_0 + n_1 = n} \left(\prod_{j=0}^{1} \left(\frac{n_j e}{d+1}\right)\right)^{(d+1)(k-1)} \leq \left(\frac{n_j e}{d+1}\right)^{(d+1)(k-1)} \tag{14}$$

where n_j denotes the points $R^d \times \{j\}$ and the result follows from the Vapnik–Chervonenkis.

Other condensate rules based on CNN were also presented in [25, 26].

Table 1 New approaches based on traditional CNN methods

Method	Short description	References
Extended nearest neighbor	Used for pattern recognition	[10]
The fast-condensed nearest neighbor algorithm	Reuse Voronoi's concepts	[18]
Hierarchy extreme learning machine, for instance, selection	Fuzzy c-means utilizes condensed nearest neighbor (CNN) to make a preliminary selection of training samples	[7]
A modified firefly algorithm for image classification	Used in image classification task	[8]
Nonparametrically regression algorithm with instance selection	Provide flexible forms of prediction	[11]

An approach to the CNN algorithm [27, 28] can be as follows:

Algorithm 1

1. $\boldsymbol{T} \leftarrow \emptyset$
2. Do
3. $\forall \boldsymbol{x} \in \boldsymbol{X}(\boldsymbol{inrandomorder})$
4. Find $\boldsymbol{x}' \in \boldsymbol{T}$ such that

$$\boldsymbol{x} - \boldsymbol{x}' = \underbrace{\boldsymbol{min}}_{\boldsymbol{x}^{\boldsymbol{w}} \in \boldsymbol{T}} \boldsymbol{x} - \boldsymbol{x}^{\boldsymbol{w}}$$

5. If $\boldsymbol{Class}(\boldsymbol{x}) \neq \boldsymbol{Class}(\boldsymbol{x}^{\boldsymbol{w}})$ insert x to T
6. While T does not change

Several investigations have been carried out to interpret, extend and enhance the traditional CNN algorithm [29, 30]. In Table 1, some novel variants of implementation and application of the CNN method are shown.

3 Results and Discussion

A small review of the process of selecting instances has shown the high potential of sample selection techniques. Its application is valid in all areas and sub-areas of the modern world. The prototyping approach given by machine learning contributes too many investigations to reduce the computational cost of processes and the tasks of classifying huge amounts of data. Stopping in the analysis of the condensed nearest neighbor (CNN) algorithm [31], it represents a cognitive and theoretical element that means the basis of other evolutionary models.

The CNN algorithms use one nearest neighbor rule to iteratively decide if a sample should be removed or not [4].

3.1 Metric Considerations and Visual Scheme for the CNN

Many unsupervised algorithms perform unsupervised learning of distance metrics using information from the data itself or from the dimension where they are represented. In the selection of instances, the measurement of the distance between instances or the metric used is of crucial importance.

To formalize, denote the vectors x and y to those that represent the attributes of two instances x and y (classes are excluded).

A widely used metric is the Minkowski metric, which is defined as:

$$d = \sqrt[p]{\sum_{j=1}^{m} d_j^p} \tag{15}$$

where d_j is defined for continuous attributes such as $d_j = \left|x_j - y_j\right|$.

For some values of p, the Minkowski distance corresponds to a special metric as reflected in Table 2.

There are other important metrics such as the Mahalanobis distance based on the location of multivariate outliers to indicate an unusual combination between one or more variables.

A simple definition to this problem [10] is defined by:

$$d(\text{Mahalanobis}) = \left[(x_B - x_A)^{\mathrm{T}} * C^{-1} * (x_B - x_A)\right]^{0.5} \tag{19}$$

where:

x_A and x_B are a pair of objects and C is the sample covariance matrix.

The following figure shows some examples of sample selection using the Euclidean and the Mahalanobis distance using the CNN algorithm and comparing some values for the n-neighbors:

Figure 4 shows the importance of the selection and use of metrics at the time of clustering, as indicated by the classic methods of selection of instances, and the

Table 2 Minkowski metrics for different p values

Minkowski variant metric	The p value	Metric
Manhattan distance	1	$d = \sum_{j=1}^{m} d_j^p$ (16)
Euclidean distance	2	$d = \sqrt{\sum_{j=1}^{m} d_j^p}$ (17)
Chebyshev distance	∞	$d = \overbrace{\underbrace{\text{Max}}_{j=1}}^{n} \left\|d_j\right\|$ (18)

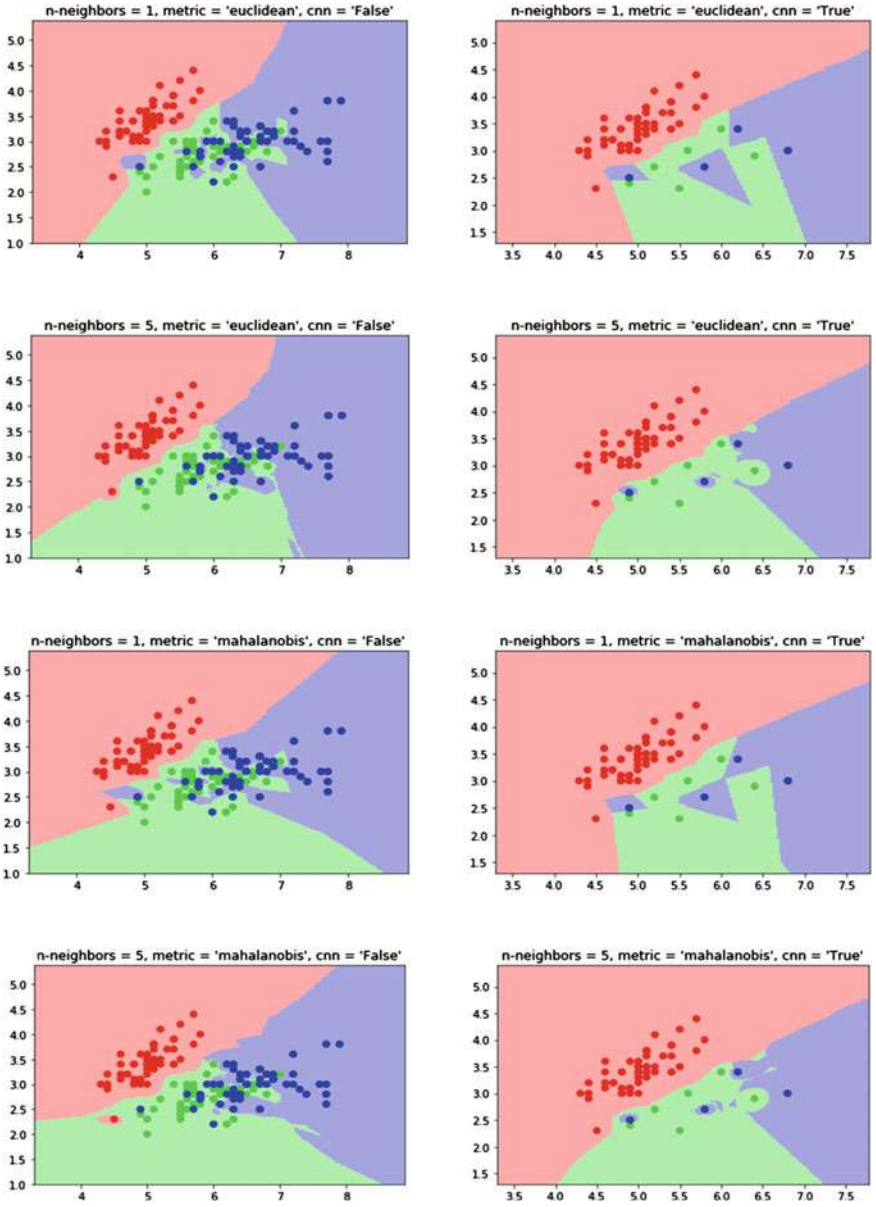

Fig. 4 Sample selection considering the Euclidean and Mahalanobis distance

fact of resorting to a sample that is sufficiently representative of a large population constitutes a difficult job. In this case, the example presented in Fig. 4 shows how the red, green and blue points are selected reflecting their color in a determined area according to the Euclidean and Mahalanobis metrics but using the CNN algorithm (squares on the right in Fig. 4) or simply using the aforementioned metrics (left squares in Fig. 4). As can be seen, using the CNN algorithm in combination with one of the two metrics achieves a clearer and more precise level of the reduced instances.

4 Conclusion

The beginning of the history of instance selection algorithms can be placed in the CNN algorithm (condensed nearest neighbor rule) whose contribution is due to Hart in 1968. The algorithm in its simplest state leaves in S a subset of T such that each element of T is closer to an element of S of the same class than to an element of S of a different class. From this idea, various variants have been formulated with an elegant mathematical profile that has allowed the reduction of computational costs in various modern problems given its simplicity.

Finally, the aim of this work has been to show some theoretical elements about the importance of the sample selection process and the condensed nearest neighbor method collected in the effort of several authors who have tried to theorize in complex aspects of the real world to give solutions to problems of today's world.

Acknowledgements The authors are greatly grateful for the support given by the SDAS Research Group. https://sdas-group.com/.

References

1. García S, Luengo J, Herrera F (2015) Data preprocessing in data mining. Springer International Publishing, Cham
2. Nisbet R, Elder J, Miner G (2009) Handbook of statistical analysis and data mining applications. Elsevier
3. Liu B (2011) Web data mining: exploring hyperlinks, contents, and usage data, 2nd edn. Springer, Heidelberg, New York
4. Hart P (1968) The condensed nearest neighbor rule (Corresp.). IEEE Trans Inf Theory 14:515–516. https://doi.org/10.1109/TIT.1968.1054155
5. Kohonen T (1990) The self-organizing map. Proc IEEE 78:1464–1480. https://doi.org/10.1109/5.58325
6. Girolami M, He C (2003) Probability density estimation from optimally condensed data samples. IEEE Trans Pattern Anal Mach Intell 25:1253–1264
7. Tang B, He H, Zhang S (2020) MCENN: a variant of extended nearest neighbor method for pattern recognition. Pattern Recogn Lett S0167865520300143. https://doi.org/10.1016/j.patrec.2020.01.015
8. Dey N (2020) Applications of firefly algorithm and its variants: case studies and new developments. Springer, Singapore

9. Chen Y, Liu Y, Ning J, Nie L, Zhu H, Chu H (2017) A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. Stat Methods Med Res 26:914–930. https://doi.org/10.1177/0962280214562146
10. Stephanie (2017) Mahalanobis distance: simple definition, examples. In: statistics how to. https://www.statisticshowto.com/mahalanobis-distance/. Accessed 19 July 2020
11. Gong C, Wang P, Su Z (2020) An interactive nonparametric evidential regression algorithm with instance selection. Soft Comput. https://doi.org/10.1007/s00500-020-04667-4
12. Silhavy P, Silhavy R, Prokopova Z (2017) Evaluation of data clustering for stepwise linear regression on use case points estimation. Adv Intell Syst Comput 575:491–496. https://doi.org/10.1007/978-3-319-57141-6_52
13. Biancolillo A, Næs T (2019) The sequential and orthogonalized PLS regression for multiblock regression. In: Data handling in science and technology. Elsevier, pp 157–177
14. Barraza N, Moro S, Ferreyra M, de la Peña A (2019) Mutual information and sensitivity analysis for feature selection in customer targeting: a comparative study. J Inf Sci 45:53–67. https://doi.org/10.1177/0165551518770967
15. Smys S, Bala GJ (2012) Performance analysis of virtual clusters in personal communication networks. Cluster Comput 15:211–222. https://doi.org/10.1007/s10586-012-0209-8
16. Jyothirmai P, Raj J, Smys S (2017) Secured self organizing network architecture in wireless personal networks. Wireless Pers Commun 96:5603–5620. https://doi.org/10.1007/s11277-017-4436-4
17. Xu X, Li S, Liang T, Sun T (2020) Sample selection-based hierarchical extreme learning machine. Neurocomputing 377:95–102. https://doi.org/10.1016/j.neucom.2019.10.013
18. Ros F, Guillaume S (2020) Sampling techniques for supervised or unsupervised tasks. Springer International Publishing, Cham
19. Cerruela-García G, de Haro-García A, Toledano JP-P, García-Pedrajas N (2019) Improving the combination of results in the ensembles of prototype selectors. Neural Netw 118:175–191. https://doi.org/10.1016/j.neunet.2019.06.013
20. Brighton H, Mellish C (2002) Advances in instance selection for instance-based learning algorithms. Data Min Knowl Disc 6:153–172. https://doi.org/10.1023/A:1014043630878
21. Garcia S, Derrac J, Cano JR, Herrera F (2012) Prototype selection for nearest neighbor classification: taxonomy and empirical study. IEEE Trans Pattern Anal Mach Intell 34:417–435. https://doi.org/10.1109/TPAMI.2011.142
22. Biehl M, Hammer B, Villmann T (2016) Prototype-based models in machine learning: prototype-based models in machine learning. WIREs Cogn Sci 7:92–111. https://doi.org/10.1002/wcs.1378
23. Devroye L, Györfi L, Lugosi G (1996) A probabilistic theory of pattern recognition. Springer, New York, NY
24. Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK (1989) Learnability and the Vapnik-Chervonenkis dimension. J ACM 36:929–965. https://doi.org/10.1145/76359.76371
25. Gates W (1972) The reduced nearest neighbor rule
26. Fukunaga K, Mantock JM (1984) Nonparametric Data Reduction. IEEE Trans Pattern Anal Mach Intell PAMI-6:115–118. https://doi.org/10.1109/TPAMI.1984.4767485
27. Ullmann J (1974) Automatic selection of reference data for use in a nearest-neighbor method of pattern classification (Corresp.). IEEE Trans Inform Theory 20:541–543. https://doi.org/10.1109/TIT.1974.1055252
28. Ritter G, Woodruff H, Lowry S, Isenhour T (1975) An algorithm for a selective nearest neighbor decision rule (Corresp.). IEEE Trans Inform Theory 21:665–669. https://doi.org/10.1109/TIT.1975.1055464
29. TOMEK I (1976) Two modifications of CNN. IEEE Trans Syst, Man, Cybern SMC-6:769–772. https://doi.org/10.1109/TSMC.1976.4309452
30. Swonger CW (1972) Sample set condensation for a condensed nearest neighbor decision rule for pattern recognition 511–519
31. Gowda K, Krishna G (1979) The condensed nearest neighbor rule using the concept of mutual nearest neighborhood (Corresp.). IEEE Trans Inform Theory 25:488–490. https://doi.org/10.1109/TIT.1979.1056066