Theoretical and Empirical Criteria for the Edited Nearest Neighbour Classifier

Ludmila I. Kuncheva
School of Computer Science, Bangor University
Dean Street, Bangor, Gwynedd, LL57 2NJ
United Kingdom
Email: l.i.kuncheva@bangor.ac.uk

Mikel Galar
Department of Automática y Computacíon
Universidad Pública de Navarra
31006 Navarra, Spain
Email: mikel.galar@unavarra.es

***Abstract*—We aim to dispel the blind faith in theoretical criteria for optimisation of the edited nearest neighbour classifier and its version called the Voronoi classifier. Three criteria from past and recent literature are considered: two bounds using Vapnik-Chervonenkis (VC) dimension and a probabilistic criterion derived by a Bayesian approach. We demonstrate the shortcomings of these criteria for selecting the best reference set, and summarise alternative empirical criteria found in the literature.**

I. Introduction

The nearest neighbour classifier is among the most intuitive, accurate, widely acclaimed and thoroughly studied classifiers. It has justly been nominated among the top ten algorithms in data mining [1].

Theoretical bounds on the error of the nearest neighbour classification rule (1-nn) have been continually developed and honed ever since the formal introduction of the rule [2], [3]. In parallel, a large body of research has followed the path of data editing for the nearest neighbour rule [4]–[6]. It will not be an exaggeration to say that there are over a hundred editing methods where prototypes are selected from the given training set [7], and a matching number of methods for prototype extraction (positioning of the prototypes in the data space, not necessarily as realistic data points) [8].

How to choose an editing method for the data set at hand? While domain knowledge, complexity-based measures and other data characteristics have been found useful in recommending the type of instance selection method [9], [10], the ultimate pair of criteria remain the classification accuracy and the data reduction rate. The hope that theory will guide the choice of editing method seems to be overly optimistic thus far as the overwhelming majority of such methods are heuristic-based.

Here we are interested in understanding better the reason behind the gap between the rich 1-nn theory and its limited impact on the design of methodology and algorithms for the edited 1-nn. Two fundamental questions need to be answered: 1. What insights do the theoretical results give us about the behaviour of the edited 1-nn? and 2. How can we use these insights to build better edited 1-nn classifiers? We argue that general results are of limited use for guiding the practical design of the edited 1-nn and propose a practical solution. The rest of the paper is organised as follows. Section II introduces the three theoretical bounds and criteria for the edited 1-nn and a version thereof called the Voronoi classifier. We present our argument about the weaknesses of these results in Section III. Section IV brings in alternative empirical criteria for the edited 1-nn and the Voronoi classifier, and Section V states our conclusions.

II. Theoretical bounds and criteria

Given is a labelled data set of N instances $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, called the reference set, where instances $\boldsymbol{x}_i$ live in some n-dimensional space $\mathbb{R}^n$. The corresponding set of labels is $Y = \{y_1, \ldots, y_N\}$, where label y_i takes values in the set $\Omega = \{\omega_1, \ldots, \omega_c\}$. Assuming a distance is specified for $\mathbb{R}^n$, the nearest neighbour classifier (1-nn) assigns an instance to the class of its nearest neighbour.

A. Bounds on the 1nn error

Asymptotic bounds of the 1-nn rule have been proven at its conception [3]. Denote by P_* the Bayes error for a given classification problem, and by P, the nearest neighbour error. Then the following holds

$$P_* \leq P \leq 2P_* - \frac{c}{c-1}P_*^2 \, .$$

The bound is proven for sample size $N \rightarrow \inf$.

It is more interesting to find out how 1-nn and the edited 1-nn behave when the data set is of finite cardinality. Nock and Sebban [11] derive a bound on the 1-nn error for a *fixed* reference set of cardinality N. They add a penalty term to the right-hand side of the above inequality reflecting the fact that the sample X is fixed. This term depends on the maximum variation of the likelihood between an element of X and its nearest neighbour (under some smoothness assumptions). Fukunaga and Hummels [12] estimate theoretically the expected bias of the finite-sample 1-nn as a product of two terms. The first term depends on the data size N and the feature space dimensionality n. Larger n increases the bias, while larger N decreases it. The second term accounts for the probability distributions but is dominated by the parameters of the metric used in the space. Their analysis concludes that, for low dimensionality n, increasing N is efficient in reducing the expected bias but for large n this is not the case. The expression for the 1-nn bias is only of theoretical interest because it assumes knowledge of the probability densities and requires high-dimensional integration.

B. Bounds on the edited 1-nn error

Methods for prototype selection and prototype extraction emerged early on, even without the prospect of large data. The task of a prototype selection method is to identify the minimum subset S of X with maximum classification accuracy for unseen data. Wilson [13] proposes to remove all examples which are misclassified by their three nearest neighbours, which clears up noise in the data, and obeys the following desirable asymptotic property. The nearest neighbour $\boldsymbol{x}'$ of instance $\boldsymbol{x}$ in the *edited* reference set S converges in probability to $\boldsymbol{x}$ when $N \to \inf$. This ensures that the edited 1-nn does not suffer any loss of accuracy *asymptotically*.

Estimating bounds on the finite-sample edited versions of 1-nn is more difficult. Denote by M be the number prototypes in the reference set selected from X (by any methodology or randomly sampled). Using the Vapnik-Chervonenkis (VC) dimension [14], Devroye et al. [15] and Devroye and Wagner [16] prove that for all $\epsilon > 0$ and all distributions,

$$\begin{aligned} Pr\left\{|P_N - \hat{P}_{N,M}| \geq \epsilon\right\} \quad &\leq \\ 8\left(\frac{Ne}{n+1}\right)^{(n+1)M(M-1)} &\times \exp\left\{\frac{-N\epsilon^2}{32}\right\}, \end{aligned} \tag{1}$$

where P_N is the classification error of 1-nn on a data set of N instances, and $\hat{P}_{N,M}$ is the classification error on the training data X using the M prototypes ($\hat{P}_{N,M}$ is termed sometimes 'apparent error rate' or 'empirical risk'). Equation (1) ascertains that the way to reduce the discrepancy between the true error P_N and $\hat{P}_{N,M}$ is increasing N.

C. Structural risk minimisation for the edited 1-nn

Karaçalı and Krim [17] give a proof that the VC dimension of the 1-nn classifier with M reference instances is exactly M. The VC dimension of a classifier C is the maximum number of instances that can be *shattered* by this classifier.[1] If the cardinality of the set is increased even by one instance, there exists a label assignment such that C will fail to label all instances correctly, for any parameter combination and any training of C.

Clearly, the 1-nn classifier with a reference set of M instances can label correctly these M instances. The proof is based on engineering labels for any set of $M+1$ instances so that no selection of a reference set of M instances will label all $M+1$ instances correctly.

Consider a family of labelling functions F. The structural risk minimisation (SRM) principle suggests that in order to improve the generalisation performance of a classifier from F, trained on a data set with N instances, the upper bound on the risk should be minimised. With probability at least $1-\delta$, the upper bound of the risk for a function $f \in F$ satisfies the following inequality [14]

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{h\log_2\left(\frac{2N}{h}+1\right) - \log_2\left(\frac{\delta}{4}\right)}{N}}, \tag{2}$$

where $R_{\text{emp}}(f)$ is the empirical risk, estimated as the classification error of f on the training set.

Applied to the edited 1-nn for a training set X of size N, SRM can be implemented in the following steps:

1) Create N families of functions: $F_1, F_2, \ldots, F_N$. Family F_i contains all 1-nn classifiers with a reference set of i instances selected from X. By the above argument, these families are arranged in the way of increasing VC-dimension.
2) For each i, find the function $f_i^* \in F_i$ which minimises the empirical risk.
3) Select as the final reference set the one corresponding to the smallest upper bound, i.e.,

$$f^* = \arg\min_i R(f_i^*)\ .$$

D. A Bayesian view on prototype selection

Taking a Bayesian view on the prototype selection problem, Ferrandiz and Boullé [18] propose a prototype selection method called Eva. Their base classifier differs from the standard edited 1-nn in that each prototype is relabelled to the majority class in its Voronoi cell. We will refer to this classifier as the *Voronoi classifier*. The 'model' consists of the prototype set and the respective labels. Denote by N_m the number of instances from X in the Voronoi cell of prototype m, $m = 1,\ldots,M$. Among the N_m instances, denote by N_m^j the number of instances from class j. Under the assumption of uniform random priors, the likelihood of the model is maximised by choosing a set of prototypes which minimises the following expression

$$\begin{aligned} C = \log(N) + \log\binom{N+M-1}{M} \\ + \sum_{m=1}^{M}\left[\log\binom{N_m+c-1}{c-1} + \log\frac{N_m!}{N_m^1!N_m^2!\ldots N_m^c!}\right]\ . \end{aligned} \tag{3}$$

As with the other approaches, the criterion consists of a term accounting for the sizes of the original and edited data (N and M), and another, which accounts for the data distribution. Curiously, neither term is based on the training error rate. Instead, the data-distribution term accounts for the entropy in the Voronoi cells. Unfortunately, the criterion may prove computationally intractable due to the large values of the binomial coefficients even for small M and c.

III. How useful are the bounds and criteria?

The more general the bound, the wider is its validity. For example, distribution-free bounds apply to all problems. But by the same token, generality may hamper the usefulness for a specific problem. As always, a balance must be sought between generality and specificity.

[1] Instances are said to be *shattered* by a classifier model C with parameter set θ if, for any labelling of the instances, there exists a set of parameters θ^* such that C labels all instances correctly.

A. Bounds on the edited 1-nn error (1)

Unfortunately, by the author's own admission [16], p 209, the bound can be useless even for moderate N and small c, n and M. To be of any relevance, the RHS of (1) must be smaller than one. Denote the RHS by β. Let us fix the size of the reference set to $M = 2$ prototypes, the number of classes to $c = 2$ and the dimensionality to $d = 1$. Figure 1 shows $\log(\beta)$ as a function of the training set size N for several values of the discrepancy term ϵ.

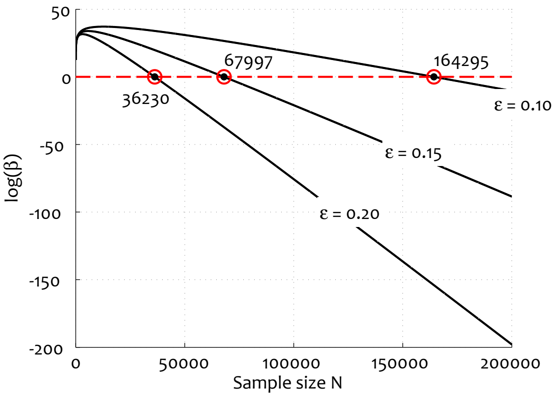


Fig. 1. Logarithm of the RHS (β) of (1) as a function of the sample size N ($n = 1$, $M = 2$). The displayed values at $\log(\beta) = 0$ are the minimum N beyond which the bound may become useful.

The example shows that, for the bound to be of any use (smaller than 1), an impractically large data set is needed even for a one-dimensional space and two prototypes. Hence, this bound is not useful for guiding the design of data editing methods.

B. The SRM bound (2)

The practical merit of a bound depends on how tight the bound is and whether or not it correlates with the behaviour of the generalisation error.

Figure 2 plots the second term of the right-hand side (RHS) of (2), called the 'VC confidence'. For the edited 1-nn classifier, $h = M$, where M is the size of the reference set. This plot will be the bound on $R(f)$ provided that the empirical risk $R_{\text{emp}}(f)$ is zero.[2]

The curve is not affected much by the value of δ. The curves for $\delta = \{0.001, 0.01, 0, 05\}$ practically coincide. The curve quickly shoots to above 1, indicating that the bound may be too loose to be useful.

C. The probabilistic criterion (3)

The computational infeasibility of criterion C for large N and M is only one of the problems. More importantly, its relationship with the generalisation error may be far too reliant on the arbitrary choice of the prior probabilities as demonstrated by the following argument.

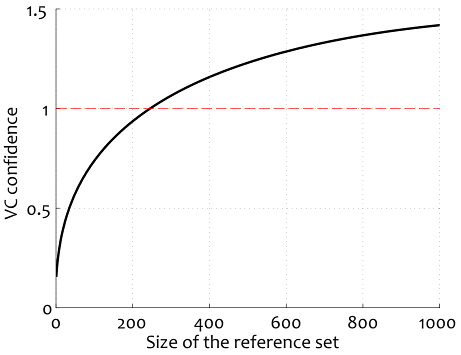


Fig. 2. The bound on $R(f)$ for the edited 1-nn and consistent reference sets, $N = 1000$ instances and $\delta = 0.001$.

Suppose that the training set X is sampled from the distribution of the problem. A *model* $Q_M =< Z, M >$ consists of a set Z of M prototypes, $Z \subseteq X$ together with their class labels $\zeta = \{\zeta_1, \ldots, \zeta_M\}$. The class label ζ_i of prototype $\boldsymbol{z}_i \in Z$ is obtained as the majority class of the instances from X falling in $\boldsymbol{z}$'s Voronoi cell. The class of the model is the value of M (the cardinality of the prototype set) which also happens to be the VC-dimension of the Voronoi classifier as explained above. This parameter governs the balance between underleaning and overfitting, and is therefore crucial for the success of the classifier. Once M is chosen, the concrete realisation of the model is the selection of Z from X. As the overfitting issue is accounted for through M, the task is to select the most accurate model of this class as evaluated on the training set.

Formally, we want to choose a model class and then a realisation, which maximise the likelihood of the model given the data set, that is:

$$P(Q_M|X) = \frac{P(X|Q_M)P(Q_M)}{P(X)}.$$

Note that $P(X)$ does not have any effect on the choice of the model parameters, hence it can be ignored to give

$$P(Q_M|X) \propto P(X|Q_M)P(Q_M) \qquad (4)$$

Thus the function to be maximised is a product of the likelihood of X, given Q_M and the prior for Q_M. With the Bayesian approach, we have the freedom to choose interpretations for both terms and specify conditions under which these interpretations hold. We can choose to think of $P(X|Q_M)$ as the classification accuracy of 1-nn with the set of labelled prototypes Q_M estimated on X. This is a reasonable interpretation because the likelihood of the model containing the whole data set itself ($Q_M =< X, N >$) will be maximum.

Q_M consists of two parameters: the model class M and the prototype set Z. The prior for the model is

$$P(Q_M) = P(Z, M) = P(Z|M)P(M).$$

Ferrandiz and Boullé [18] take $P(M)$ to be uniform

$$P(M) = \frac{1}{N}, \quad M = 1, \ldots, N.$$

[2] Reference sets for which $R_{\text{emp}}(f) = 0$ are called 'consistent'.

This is called a non-informative prior and does not have any effect on the model choice. Knowing the problem, however, we know that the curve of the testing accuracy as a function of M is likely to have a "belly" shape. Small values of M lead to under-training while large values of M lead to overfitting, with $M = N$ giving a zero resubstitution error but learning all the noise. Then it is more intuitive to choose a binomial distribution for M, $B(N, p)$:

$$P(M) = \binom{N}{M} p^M (1-p)^{(N-M)}, \;\; M = 1, \ldots, N,$$

for some probability of success p. In absence of a further insight, we can set $p = 0.5$, giving

$$P(M) = \frac{1}{2^N}\binom{N}{M}, \;\; M = 1, \ldots, N.$$

The prior $P(Z|M)$ is more difficult to rationalise. Any set of prototypes may turn out to be the best. Therefore, a uniform prior probability is justified. In our set up, there are $\binom{N}{M}$ ways to select M instances out of a set of N, hence

$$P(Z|M) = \frac{1}{\binom{N}{M}}.$$

Based on the above, consider the following two cases

1) Uniform priors for M and for Z, given M [18].
2) A binomial prior for M and a uniform prior for Z, given M.

Case 1.: If both priors are uniform,

$$P(Z, M) = \frac{1}{N\binom{N}{M}}.$$

This means that reference sets of relatively smaller or relatively larger cardinality will have higher likelihood of being chosen as solutions for the same likelihood $P(X|Q_M)$ compared to sets of medium cardinality, which sounds counter-intuitive.

Case 2.: If $M \sim B(N, 0.5)$,

$$P(Z, M) = \frac{1}{2^N}.$$

As this prior does not depend on M, it can be ignored in (4), leaving.

$$P(Q_M|X) \propto P(X|Q_M).$$

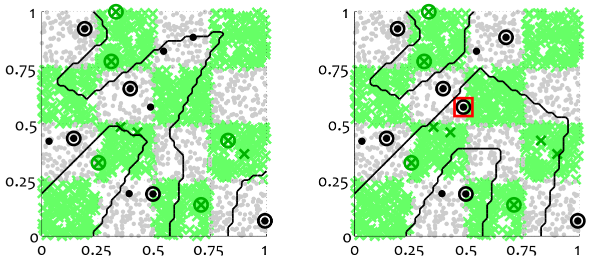
This expression is useless as a criterion because it will always favour the trivial model $Q_M =< X, M^* >$, where M^* is the cardinality of the smallest consistent set.

Of course we can engineer a prior which will place more weight on the medium values of M so as to reflect the anticipated shape of the generalisation error as a function of M. The problem, however, is that this (*arbitrary*) choice of the model prior and its parameters will determine the optimal value of M. While this will be a criterion dressed in a theoretical guise, it will be no different to a criterion where the trade-off between under- and over-training is set up by eye or estimated by a cross-validation protocol.

D. A toy example

Consider a toy example of two classes distributed in a 4×4 checker board pattern as shown in Figure 3. Eighteen instances are sampled randomly from a uniform distribution in the unit square as the data set X along with their labels ($N = 18$). *All* $2^{18} - 1 = 262144$ possible non-empty subsets of X were generated with a view to explore the criteria values as predictors of the errors of the edited 1-nn and the Voronoi classifier.

Figure 3 shows the classification boundaries and the error rates for the *best* reference sets from X for the edited 1-nn (10 prototypes) and the Voronoi classifier (11 prototypes). The Voronoi classifier has a slightly lower error because of its ability to relabel the prototypes as the majority of instances in their respective Voronoi cells. The difference from the original labelling is indicated by a square marker.



(a) 1-nn classifier
$\hat{P}_{18,10} = 34.2\%$

(b) Voronoi classifier
$\hat{P}_{18,11} = 32.3\%$

Fig. 3. The classification boundaries for the edited 1-nn (a) and the Voronoi classifier (b) for the toy example data set. The prototype relabelled into the opposite class for the Voronoi classifier is marked with a square.

The top two rows of plots in Figure 4 show the best training error for $M = 1, \ldots, 18$, the best testing error (not necessarily corresponding to the best training error). The shaded plots in the bottom row show the respective bound/criterion curves. It can be seen that the criterion/bound *do not correlate* with the testing error.

IV. Empirical criteria for the edited 1-nn

The theoretical bounds reviewed above consist of two terms, either additive or multiplicative. One of the terms accounts for the sizes of the original and the edited data, and the other one, for the estimated error.

A. Leave-one-out error estimate

The term accounting for the data sizes is meant to guard against potential overfitting reflected in the low *training* error rate. However, the error term could be calculated on a separate validation set or through cross-validation, which makes it a proxy for the generalisation error itself. This renders the size-penalising term redundant. If the data size does not allow for a separate validation set, the leave-one-out (LOO) protocol offers a reasonable alternative [19]–[26]. In this case, every training instance is used to compute the error, but those in the reference set, $\boldsymbol{z} \in Z$, are labelled using $Z \setminus \{\boldsymbol{z}\}$.

B. Weighted additive penalty

A simple alternative to the theoretical criterion for choosing an edited 1-nn or the Voronoi classifier can be devised using an additive weighted penalty term

$$C_{\rm wp} = (1-\alpha)\hat{P}(Q_M) + \alpha\,\frac{M}{N}, \qquad (5)$$

where $0 < \alpha < 1$.

The weighted additive penalty criterion has been used for 1-nn editing through random search [27], [28], genetic algorithms [19], tabu search [29], memetic algorithms [22], [30], cooperative coevolution [31], ensemble-based instance selection [32], [33] or meta-learning [34]. The only difference between these approaches is the value of α. The most commonly used value is $\alpha = 0.5$, e.g., [20]–[26]), even though earlier proposals considered α values lower than 0.1, putting more emphasis on accuracy [19], [29], [35]. Depending on the specific purpose of the model, values between 0.1 and 0.5 have also been considered [21], [31], [32], [36]. In fact, an exhaustive empirical analysis on the effect of this value is still missing in the literature, since, with few exceptions [10], [37], α is set by eye of through preliminary, non-reported experiments . Likewise, there is no insight as to how different data types may benefit from different α values.

Curiously, the penalty criterion is seldom used with the training error. Typically a penalty term is added to the LOO error.

C. Results for the toy example

To illustrate the behaviour of the empirical criteria, we use again the toy example. In addition to the training error and the "true" best generalisation error for a given M, we calculated the leave-one-out (LOO) error. Based on the minimum LOO error, we chose the corresponding model (M) and report the *best* error for this model class. The third row of plots in Figures 4 displays the LOO error for the 1-nn and the Voronoi classifier. The best solution according to the LOO error is circled. The results indicate that LOO error is a useful criterion for determining the cardinality of the reference set M as it points at a near-optimal classifier.

Finally, the weighted penalty criterion was put to the test. We examined an array of values for α in equation (5). The results are reported in Table I. We calculated criterion (5) with the *training error* as the first term (columns M and P), and then with the *LOO error* as the first term (columns M^* and O^*). The top row is calculated for $\alpha = 0$, which is exactly the LOO criterion.

The table reveals that:

• The LOO criterion is sufficiently good on its own. Varying α did not offer any improvement for this example. This is not to say that no improvement can be achieved by using the weighted penalty criterion. However, the price to pay for a possible improvement is the fine-tuning of α, possibly by an internal cross-validation experiment, adding to the already expensive LOO calculation.

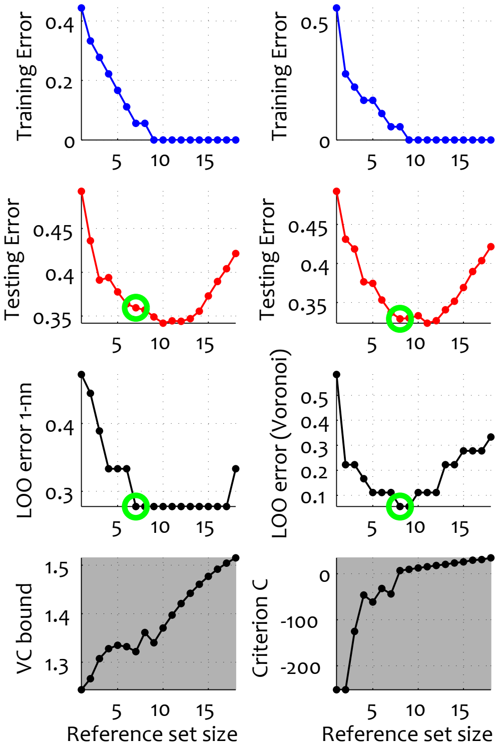


Fig. 4. Best training error, best testing error, the leave-one-out error estimate and the bound/criterion curves for the 1-nn and the Voronoi classifier on the toy example. The best solution corresponding to the minimum-cardinality set with the lowest LOO error is indicated with a circle.

TABLE I

Testing error (%) for the best reference sets for α found through the two criteria for the toy example.

	1-nn (true best 34.2)				Voronoi (true best 32.2)			
α	M	P	M^*	P^*	M	P	M^*	P^*
0.0	–	–	7	35.95	–	–	8	32.90
0.1	9	34.90	7	35.95	9	33.00	8	32.90
0.2	9	34.90	7	35.95	9	33.00	8	32.90
0.4	7	35.95	4	39.40	4	37.65	2	43.10
0.5	2	43.60	1	49.25	3	41.85	2	43.10
0.6	2	43.60	1	49.25	2	43.10	2	43.10
0.8	1	49.25	1	49.25	2	43.10	2	43.10
0.9	1	49.25	1	49.25	1	49.25	1	49.25

Notes: M and P are calculated when the training error is taken as the error term in (5); M^* and P^* are calculated when the LOO error is taken as the error term in (5). The row for $\alpha = 0$ is the LOO criterion.

• Due to its ability to filter noise by relabelling prototypes, the Voronoi classifier (a variant of the edited 1-nn) had an edge over the baseline edited 1-nn in our example. The training of the classifier is slightly different but the operation is exactly the 1-nn classifier, hence it does not bear any extra cost.

V. Conclusion

This study looked at existing theoretical results related to the edited nearest neighbour classifier and their practical use. We found that the existing bounds and criteria are either too loose to be useful or unrelated to the generalisation error of the data at hand. Our argument is illustrated on a toy problem. Instead of drawing upon the theoretical bounds and criteria when designing 1-nn editing methods, we recommend using empirical criteria. In particular, the leave-one-out estimate of the error using a given reference set seems to be justly chosen in many works.

As a byproduct, we showed the advantage of a variant of 1-nn, called the Voronoi classifier, over the baseline model.

Given the wealth of literature on the 1-nn theory and applications, the hope is that future research will bring theory and practice closer together, especially in the context of large data. Large data offers interesting prospectives. For example, instead of running internal cross-validation for evaluating α or M, independent testing sets could be sampled.

The code and data to reproduce this work are available at *https://github.com/mikelgalar/Kuncheva_Galar_ICDM2015*.

Acknowledgement

This work was supported by The Leverhulme Trust grant #RPG-2015-188, José Castillejo grant CAS14/00148 and project TIN2013-40765-P.

References

[1] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[2] E. Fix and J. L. Hodges, "Discriminatory analysis : Non parametric discrimination : Small sample performance," USAF School of Aviation Medicine, Randolph Field,Texas, Tech. Rep. Project 21 - 49 - 004 (11), 1952.

[3] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Information Theory*, vol. 13, no. 1, pp. 21–27, January 1967.

[4] D. Wilson and T. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine Learning*, vol. 38, pp. 257–286, 2000.

[5] B. V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, California: IEEE Computer Society Press, 1990.

[6] J. A. Olvera-López, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, and J. Kittler, "A review of instance selection methods," *Artificial Intelligence Review*, vol. 34, no. 2, pp. 133–143, 2010.

[7] S. García, J. Derrac, J. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 417–435, 2012.

[8] I. Triguero, J. Derrac, S. García, and F. Herrera, "A taxonomy and experimental study on prototype generation for nearest neighbor classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 1, pp. 86–100, 2012.

[9] H. Brighton and C. Mellish, "Advances in instance selection for instance-based learning algorithms," *Data Mining and Knowledge Discovery*, vol. 6, no. 2, pp. 153–172, 2002.

[10] E. Leyva, Y. Caises, A. González, and R. Pérez, "On the use of meta-learning for instance selection: An architecture and an experimental study," *Information Sciences*, vol. 266, pp. 16–30, 2014.

[11] R. Nock and M. Sebban, "An improved bound on the finite-sample risk of the nearest neighbor rule," *Pattern Recognition Letters*, vol. 22, no. 3–4, pp. 407–412, 2001.

[12] K. F. D. M. Hummels, "Bias of nearest neighbor error estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 1, pp. 103–112, 1987.

[13] D. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, pp. 408–421, 1972.

[14] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[15] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[16] L. Devroye and T. J. Wagner, "Distribution-free performance bounds with the resubstitution error estimate," *IEEE Transactions on Information Theory*, vol. 25, no. 2, pp. 208–210, 1979.

[17] B. Karaçalı and H. Krim, "Fast minimization of structural risk by nearest neighbor rule," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 127–137, 2003.

[18] S. Ferrandiz and M. Boullé, "Bayesian instance selection for the nearest neighbor rule," *Machine Learning*, vol. 81, pp. 229–256, 2010.

[19] L. Kuncheva, "Fitness functions in editing k-nn reference set by genetic algorithms," *Pattern Recognition*, vol. 30, pp. 1041–1049, 1997.

[20] J. R. Cano, F. Herrera, and M. Lozano, "Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study," *Evolutionary Computation, IEEE Transactions on*, vol. 7, no. 6, pp. 561–575, 2003.

[21] J. Derrac, S. García, and F. Herrera, "Ifs-coco: Instance and feature selection based on cooperative coevolution with nearest neighbor rule," *Pattern Recognition*, vol. 43, no. 6, pp. 2082–2105, 2010.

[22] S. García, J. R. Cano, and F. Herrera, "A memetic algorithm for evolutionary prototype selection: A scaling up approach," *Pattern Recognition*, vol. 41, no. 8, pp. 2693–2709, 2008.

[23] A. de Haro-García and N. García-Pedrajas, "A divide-and-conquer recursive approach for scaling up instance selection algorithms," *Data Mining and Knowledge Discovery*, vol. 18, no. 3, pp. 392–418, 2009.

[24] J. Derrac, C. Cornelis, S. García, and F. Herrera, "Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection," *Information Sciences*, vol. 186, no. 1, pp. 73–92, 2012.

[25] N. García-Pedrajas, A. de Haro-García, and J. Pérez-Rodríguez, "A scalable approach to simultaneous evolutionary instance and feature selection," *Information Sciences*, vol. 228, pp. 150–174, 2013.

[26] J. Derrac, I. Triguero, S. Garcia, and F. Herrera, "Integrating instance selection, instance weighting, and feature weighting for nearest neighbor classifiers by coevolutionary algorithms," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 5, pp. 1383–1397, 2012.

[27] L. Kuncheva and J. Bezdek, "On prototype selection: Genetic algorithms or random search?" *IEEE Transactions on Systems, Man, and Cybernetics*, vol. C28, no. 1, pp. 160–164, 1998.

[28] J. C. Bezdek and L. I. Kuncheva, "Some notes on twenty one 21 nearest prototype classifiers," in *Advances in Pattern Recognition*, ser. LNSC, F. J. Ferri, J. M. Iesta, A. Amin, and P. Pudil, Eds., 2000, vol. 1876, pp. 1–16.

[29] V. Cerverón and F. Ferri, "Another move towards the minimum consistent subset: A tabu search approach to the condensed nearest neighbor rule," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 31, no. 3, pp. 408–413, 2001.

[30] J. Derrac, S. García, and F. Herrera, "Stratified prototype selection based on a steady-state memetic algorithm: a study of scalability," *Memetic Computing*, vol. 2, no. 3, pp. 183–199, 2010.

[31] N. García-Pedrajas, J. A. R. del Castillo, and D. Ortiz-Boyer, "A cooperative coevolutionary algorithm for instance selection for instance-based learning," *Machine Learning*, vol. 78, no. 3, pp. 381–420, 2010.

[32] C. García-Osorio, A. de Haro-García, and N. García-Pedrajas, "Democratic instance selection: A linear complexity instance selection algorithm based on classifier ensemble concepts," *Artificial Intelligence*, vol. 174, no. 5-6, pp. 410–441, 2010.

[33] A. de Haro-García, N. García-Pedrajas, and J. A. R. del Castillo, "Large scale instance selection by means of federal instance selection," *Data & Knowledge Engineering*, vol. 75, pp. 58–77, 2012.

[34] E. Leyva, A. González, and R. Pérez, "Knowledge-based instance selection: A compromise between efficiency and versatility," *Knowledge-Based Systems*, vol. 47, pp. 65–76, 2013.

[35] T. Nakashima and H. Ishibuchi, "Ga-based approaches for finding the minimum reference set for nearest neighbor classification," in *The 1998 IEEE International Conference on Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence*, 1998, pp. 709–714.

[36] N. García-Pedrajas and A. de Haro-García, "Boosting instance selection algorithms," *Knowledge-Based Systems*, vol. 67, pp. 342–360, 2014.

[37] N. García-Pedrajas, J. Perez-Rodríguez, and A. de Haro-García, "OligoIS: Scalable instance selection for class-imbalanced data sets," *IEEE Transactions on Cybernetics*, vol. 43, no. 1, pp. 332–346, 2013.