An Evidential K-Nearest Neighbor Classifier based on Contextual Discounting*

Orakanya Kanjanatarakul[1], Siwarat Kuson[2] and Thierry Denoeux[3]

[1] Faculty of Management Sciences,
Chiang Mai Rajabhat University, Thailand
[2] Faculty of Economics, Maejo University, Thailand
[3] Université de Technologie de Compiègne, CNRS,
UMR 7253 Heudiasyc, France
orakanyaa@gmail.com, ksiwarat@gmail.com, tdenoeux@utc.fr

Abstract. The evidential K nearest neighbor classifier is based on discounting evidence from learning instances in a neighborhood of the pattern to be classified. To adapt the method to partially supervised data, we propose to replace the classical discounting operation by contextual discounting, a more complex operation based on as many discount rates as classes. The parameters of the method are tuned by maximizing the evidential likelihood, an extended notion of likelihood based on uncertain data. The resulting classifier is shown to outperform alternative methods in partially supervised learning tasks.

Keywords: Belief functions, Dempster-Shafer theory, classification, machine learning, partially supervised learning, soft labels.

1 Introduction

Since its introduction in [2], the evidential K-nearest neighbor (EKNN) classifier has been used extensively and several variants have been developed (see, e.g., [8], [6], [7], [5] and [14] for some applications and recent developments). The EKNN classifier is based on the following simple ideas: (1) each neighbor of the pattern x to be classified is considered as a piece of evidence about the class of x, represented by a mass function; (2) each mass function is discounted based on its distance to x; and (3) the discounted mass functions induced by the K nearest neighbors of x are combined by Dempster's rule.

In [2], the parameters used to define the discount rate as a function of distance were fixed heuristically, and the method was shown to outperform other K-nearest neighbor rules. In [15], the authors showed that the performances of the method could be further improved by learning the parameters through minimizing the mean squares error (MSE) between pignistic probabilities and class indicator variables. In [4], the EKNN rule was extended to the case where

* This research was supported by the Center of Excellence in Econometrics at Chiang Mai University.

the class label of training patterns is only partially known, and described by a possibility distribution. However, the learning procedure defined in [15] cannot be straightforwardly extended to the partially labeled setting because (1) the discount rate defined in the procedure depends on the class of the neighboring pattern, and (2) combining arbitrary mass functions and computing pignistic probabilities has exponential complexity in the worst case.

In this paper, we revisit the EKNN classifier by exploiting some recent developments in the theory of belief functions: (1) The discounting operation is replaced by *contextual discounting* [9], allowing us to define one discount rate parameter per class even in the partially labeled case; and (2) instead of the MSE and pignistic probabilities, we propose to use the *conditional evidential likelihood* criterion [3, 11], which allows us to account for partial class labels in a natural way, and can be computed in linear time as a function of the number of classes.

The rest of this paper is organized as follows. The EKNN classifier and classical discounting operation are first recalled in Section 2. The Contextual-Discounting Evidential K-NN (CD-EKNN) classifier is then introduced in Section 3, and experimental results are reported in Section 4. Section 5 concludes the paper.

2 Background

In this section, we provide a reminder of the main notions needed in the rest of the paper. The EKNN classifier will first be recalled in Section 2.1, and the contextual discounting operation will be presented in Section 2.2.

2.1 Evidential K-NN classifier

Consider a classification problem with c classes in $\Omega = \{\omega_1, \ldots, \omega_c\}$, and a learning set $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n$ of n examples (x_i, y_i), where x_i is a p-dimensional feature vector describing example i, and $y_i \in \Omega$ is the class of that example. Let x be new pattern to be classified, and $\mathcal{N}_K(x)$ the set of its K nearest neighbors in $\mathcal{L}$, according to some distance d (usually, the Euclidean distance when the p features are numerical). In [2] and [15], it was assumed that each neighbor $x_j \in \mathcal{N}_K(x)$ induces a simple mass function $\widehat{m}_j$ defined as

$$\widehat{m}_j(\{\omega_k\}) = \beta_k(d_j) y_{jk}, \quad k = 1, \ldots, c \tag{1a}$$

$$\widehat{m}_j(\Omega_k) = 1 - \beta_k(d_j), \tag{1b}$$

where $y_{jk} = 1$ if $y_j = \omega_k$ and $y_{jk} = 0$ otherwise, $d_j = d(x, x_j)$ and β_k is a decreasing function, usually taken as $\beta_k = \alpha \exp(-\gamma_k d_j^2)$, where α is a coefficient in $[0, 1]$ and the γ_k's are strictly positive scale parameters. By pooling mass functions $\widehat{m}_j$ induced by the K nearest neighbors of x using Dempster's rule, we get the combined mass function $\widehat{m}$, which summarizes the evidence about the class of x based on its K nearest neighbors.

In [15], it was proposed to leave parameter α fixed and to learn parameter vector $\gamma = (\gamma_1, \ldots, \gamma_c)$ by minimizing the following error function,

$$C(\gamma) = \sum_{i=1}^{n} \sum_{k=1}^{c} (\widehat{Betp}_i(\omega_k) - y_{ik})^2, \tag{2}$$

where $\widehat{Betp}_i$ is the pignistic probability distribution computed from mass function $\widehat{m}_i$ obtained from the K nearest neighbors of x_i. Because this classifier is based on c learnable parameters γ_k, $k = 1, \ldots, c$, it will be later referred to as the γ_k-*EKNN classifier*.

The idea of applying the EKNN procedure to partially labeled data $\mathcal{L} = \{(x_i, m_i)\}_{i=1}^{n}$, where m_i is an arbitrary mass function that represents partial knowledge about the class of example x_i was already suggested in [2] and explored further in [4]. Indeed, mass function $\widehat{m}_j$ in (1) can be seen as the discounted version of the certain mass function $m_j(\{y_j\}) = 1$, with discount rate $1 - \beta_k(d_j)$ if $y_j = \{\omega_k\}$. The same discounting notion can be applied whatever the form of m_j, but the discount rate can no longer depend on y_j when it is unknown. Consequently, the extension is not straightforward. Also, the combination by Dempster's rule and the calculation of the pignistic probabilities in (2) have exponential complexities for arbitrary mass functions m_i, which makes the method less attractive unless c is very small. These issues will be addressed in Section 3, based on the notion of contextual discounting recalled hereafter.

2.2 Contextual discounting

Let m be a mass function on $\Omega = \{\omega_1, \ldots, \omega_c\}$ and β a coefficient in $[0, 1]$. The *discounting* operation [12] with discount rate $1 - \beta$ transforms m into the following mass function:

$$^{\alpha}m = \beta m + (1 - \beta) m_?, \tag{3}$$

where $m_?$ is the vacuous mass function defined by $m_?(\Omega) = 1$. This operation can be justified as follows [13]. Assume that m is provided by a source that may be reliable (R) or not ($\neg R$). If the source is reliable, we adopt its opinion as ours, i.e., we set $m(\cdot|R) = m$. If it is not reliable, then it leaves us in a state of total ignorance, i.e., $m(\cdot|\neg R) = m_?$. Furthermore, assume that we have the following mass function on $\mathcal{R} = \{R, \neg R\}$: $m_{\mathcal{R}}(\{R\}) = \beta$ and $m_{\mathcal{R}}(\mathcal{R}) = 1 - \beta$, i.e., our degree of belief that the source is reliable is equal to β. Then, combining the two mass functions $m(\cdot|R)$ (after deconditioning) and $m_{\mathcal{R}}$ yields precisely $^{\alpha}m$ in (3), after marginalizing on Ω.

In [9], the authors generalized the discounting operation using the notion of *contextual discounting*. In the corresponding refined model, $m(\cdot|R)$ and $m(\cdot|\neg R)$ are defined as before, but our beliefs about the reliability of the source are now defined given each state in Ω, i.e., we have c conditional mass functions defined by $m_{\mathcal{R}}(\{R\}|\omega_k) = \beta_k$ and $m_{\mathcal{R}}(\mathcal{R}|\omega_k) = 1 - \beta_k$, for $k = 1, \ldots, c$. Combining

$m(\cdot|R)$ with mass functions $m_{\mathcal{R}}(\cdot|\omega_k)$ after deconditioning yields the following discounted mass function,

$$^{\boldsymbol{\beta}}m(A) = \sum_{B \subseteq A} m(B) \left(\prod_{\omega_k \in A \setminus B} (1 - \beta_k) \prod_{\omega_l \in \overline{A}} \beta_l \right) \tag{4}$$

for all $A \subseteq \Omega$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_c)$, and a product of terms is equal to 1 if the index set is empty. The associated contour function is

$$^{\boldsymbol{\beta}}pl(\omega_k) = 1 - \beta_k + \beta_k pl(\omega_k), \quad k = 1, \ldots, c, \tag{5}$$

where pl is the contour function corresponding to m.

3 Contextual-discounting Evidential *K*-NN classifier

An alternative to the γ_k-EKNN classifier based on contextual discounting will first be defined in Section 3.1, and learning the parameters in this model will be addressed in Section 3.2.

3.1 Extending the EKNN classifier to partially labelled data

As the EKNN classifier is based on discounting, it can be readily generalized using contextual discounting. More precisely, let us assume that we have a partially labeled learning set $\mathcal{L} = \{(x_i, m_i)\}_{i=1}^n$. (The fully supervised case is recovered when all mass functions m_i are certain). Let x be a pattern to be classified, and x_j one of its K nearest neighbors. In [4], it was proposed to generalize (1) by discounting each neighbor mass function m_j with discount rate $1 - \beta(d_j) = 1 - \alpha \exp(-\gamma d_j^2)$. We then have two learnable parameters: coefficient α and a single scale parameter γ. This rule will later be referred to as the (α, γ)-*EKNN classifer*.

In this paper, we propose to use contextual discounting (4) instead of classical discounting. The resulting rule, called *Contextual Discounting Evidential K-nearest neighbor* (CD-EKNN) is based on c coefficients $\beta_k(d_j)$ defined by

$$\beta_k(d_j) = \alpha \exp(-\gamma_k d_j^2), \quad k = 1, \ldots, c, \tag{6}$$

and there are $c + 1$ learnable parameters $\alpha \in [0, 1]$ and $\gamma_k \geq 0$, $k = 1, \ldots, c$.

Whereas the discounted mass function $\widehat{m}_j$ may have a complicated expression, its contour function can be obtained from (5) as

$$\widehat{pl}_j(\omega_k) = 1 - \beta_k(d_j) + \beta_k(d_j) pl_j(\omega_k), \quad k = 1, \ldots, c, \tag{7}$$

and the combined contour function after pooling the evidence of the K nearest neighbors is

$$\widehat{pl}(\omega_k) \propto \prod_{x_j \in \mathcal{N}_K(x)} \left[1 - \beta_k(d_j) + \beta_k(d_j) pl_j(\omega_k)\right], \quad k = 1, \ldots, c. \tag{8}$$

We note that $\widehat{pl}$ can be computed, up to a multiplicative constant, in time proportional to the number K of neighbors and the number of c of classes. The contour function is all we need to make decisions and, as we will see in the next section, to train the classifier by maximizing the evidential likelihood criterion.

3.2 Learning

To learn the parameters $\theta = (\alpha, \gamma_1, \ldots, \gamma_c)$ of the CD-EKNN classifier defined in Section 3.1, we propose to maximize the *evidential likelihood* function introduced in [3]. Before, we introduce the evidential likelihood for this model, let us recall the expression of the "classical likelihood" in the case of fully supervised data $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^n$. Let $\widehat{pl}_i$ the contour function computed for instance i based on its K nearest neighbors using (8), and let $\widehat{p}_i$ be the probability distribution obtained from $\widehat{pl}_i$ by normalization. The conditional likelihood (given feature vectors $x_1, \ldots, x_n$) after observing the true class labels $y_1, \ldots, y_n$ is

$$L_c(\theta) = \prod_{i=1}^{n} \prod_{k=1}^{c} \widehat{p}_i(\omega_k)^{y_{ik}}. \tag{9}$$

In the partially supervised learning case, the learning set is of the form $\mathcal{L} = \{(x_i, m_i)\}_{i=1}^n$, where m_i is a mass function that represents our partial knowledge of the class of x_i. An extension of the likelihood function for such uncertain data was introduced and justified in [3]. Basically, the term $\prod_{k=1}^c \widehat{p}_i(\omega_k)^{y_{ik}}$ in (9) is replaced by the expected plausibility $\sum_{k=1}^c \widehat{p}_i(\omega_k) pl_i(\omega_k)$. The *evidential likelihood* is then defined as

$$L_e(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{c} \widehat{p}_i(\omega_k) pl_i(\omega_k), \tag{10}$$

We note that the evidential likelihood (10) boils down to the classical likelihood (9) when all mass functions m_i are certain, i.e., when $pl_i(\omega_k) = y_{ik}$ for all i and k. The evidential log-likelihood $\log L_e(\theta)$ can be maximized using an iterative optimization procedure such as Newton's method.

4 Numerical Experiments

In this section, we present some results with one simulated and two real datasets, in which label uncertainty was simulated by corrupting labels with noise and representing uncertainty using suitable mass functions. The simulated data were generated from $c = 2$ Gaussian distributions with densities $\mathcal{N}(\mu_k, \sigma_k^2 I)$, where $\mu_1 = (0, 0)^T$, $\mu_2 = (1, 0)^T$, $\sigma_1^2 = 0.1I$, $\sigma_2^2 = 2I$, and I is the identity matrix. Each simulated dataset had 100 vectors from each class. The real data were the Ionosphere data ($n = 351$ instances, $p = 34$ features and $c = 2$ classes) and the Sonar data ($n = 204$, $p = 60$, $c = 2$), both from the UCI Machine Learning Repository[4].

[4] Available at `http://archive.ics.uci.edu/ml`.

Figure 1 shows the leave-one-out error rates as functions of the number K of neighbors, in two learning situations: with true class labels (Figures 1(a), 1(c) and 1(e)), and with uncertain (soft) class labels (Figures 1(b), 1(d) and 1(f)). To generate the uncertain labels m_i, we proceeded as in [1] and [11]: for each instance i, a number p_i was generated from a beta distribution with mean $\mu = 0.5$ and variance 0.04. Then, with probability p_i, the class label y_i of instance i was replaced by y_i' picked randomly from Ω. Otherwise, we set $y_i' = y_i$. Contour function pl_i was then defined as $pl_i(\{y_i'\}) = 1$ and $pl_i(\{\omega\}) = p_i$ for all $\omega \neq y_i'$. This procedure guarantees that the soft label pl_i is all the more uncertain that the label with maximum plausibility has the more chance of being incorrect.

For each dataset and each learning situation, we considered four classifiers: (1) the (α, γ)-EKNN rule based on classical discounting and criterion (10); (2) the CD-EKNN rule with c scale parameters $\gamma_1, \ldots, \gamma_c$ trained with criterion (10); (3) the original γ_k-EKNN rule recalled in Section 2.1, trained with criterion (2); and (4) the voting K-NN rule. As the γ_k-EKNN and voting K-NN classifiers can only handle fully supervised data with certain labels, we used the noisy labels y_i' with these classifiers, instead of the soft labels m_i.

As can be seen from Figures 1(a), 1(c) and 1(e), the original γ_k-EKNN and CD-EKNN rules have similar performances in the fully supervised case, and they perform better than the (α, γ)-EKNN rule. On the simulated data, the (α, γ)-EKNN rule does not even outperform the voting K-NN rule (Figure 1(a)), whereas it performs much better on the Sonar data (Figure 1(e)).

When applied to data with soft labels, the CD-EKNN classifier clearly has the best performances. In contrast, the γ_k-EKNN and voting K-NN classifiers, which use noisy labels, perform poorly. This result confirms similar findings reported in [1], [3] and [11] for parametric classifiers. The CD-EKNN classifier also performs better than the (α, γ)-EKNN rule, except on the Sonar data, for which they achieve similar error rates.

5 Conclusions

The EKNN classifier introduced in [2] and perfected in [15] has proved very efficient for fully supervised classification. Because it applies different discount rates to neighbors from different classes, the method cannot be readily extended to the partially supervised learning situation, in which we only have uncertain information about the class of learning instances. Also, it is not clear how the MSE criterion used in [15] could be generalized in the case of partially labeled data. In this paper, we have proposed a solution to this problem by replacing classical discounting with contextual discounting introduced in [9]. The underlying idea is that the reliability of the information from different neighbors depends on the class of the pattern to be classified. We also replaced the MSE by the conditional likelihood, which has already been generalized to uncertain data in [3]. The resulting CD-EKNN classifier was shown to perform very well with partially supervised data, while performing as well as the original EKNN classifier with fully supervised data.

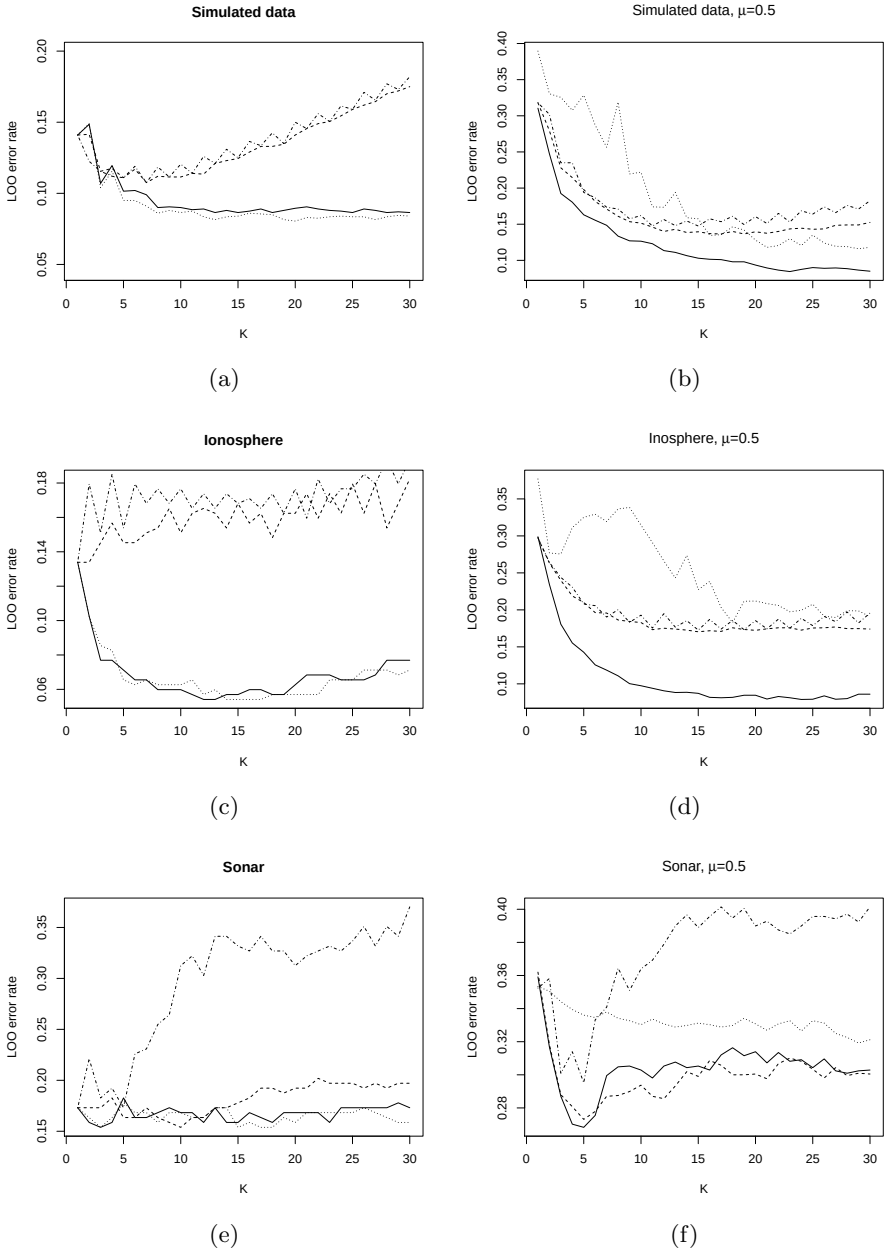

Fig. 1. Leave-one-out error rates vs. number K of neighbors for fully supervised (a, c, e) and partially supervised (b,d,f) datasets. The methods are: the CD-EKNN classifier (solid lines), the (α, γ)-EKNN classifier (dashed lines), the original γ_k-EKNN classifier (dotted lines) and the voting K-NN rule (dash-dotted lines).

In contrast with the original EKNN classifier, which assigns masses only to singletons and the whole frame of discernment, the CD-EKNN classifier generates more general mass functions, as a result of applying the contextual discounting operation. In future work, it will be interesting to study how masses assigned to various subsets of classes can be interpreted, and to find out if this richer information can be exploited for, e.g., classifier combination. Beyond discounting, other contextual mass correction mechanisms such as introduced in [10] could also be investigated.

References

1. E. Côme, L. Oukhellou, T. Denœux, and P. Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition*, 42(3):334–348, 2009.
2. T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.
3. T. Denœux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):119–130, 2013.
4. T. Denœux and L. M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, 122(3):47–62, 2001.
5. N. Guettari, A. S. Capelle-Laizé, and P. Carré. Blind image steganalysis based on evidential k-nearest neighbors. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2742–2746, Sept 2016.
6. C. Lian, S. Ruan, and T. Denœux. An evidential classifier based on feature selection and two-step classification strategy. *Pattern Recognition*, 48:2318–2327, 2015.
7. C. Lian, S. Ruan, and T. Denœux. Dissimilarity metric learning in the belief function framework. *IEEE Transactions on Fuzzy Systems*, 24(6):1555–1564, 2016.
8. Z.-G. Liu, Q. Pan, and J. Dezert. A new belief-based K-nearest neighbor classification method. *Pattern Recognition*, 46(3):834–844, 2013.
9. D. Mercier, B. Quost, and T. Denœux. Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information Fusion*, 9(2):246–258, 2008.
10. F. Pichon, D. Mercier, E. Lefèvre, and F. Delmotte. Proposition and learning of some belief function contextual correction mechanisms. *International Journal of Approximate Reasoning*, 72:4–42, 2016.
11. B. Quost, T. Denœux, and S. Li. Parametric classification with soft labels using the evidential em algorithm: linear discriminant analysis versus logistic regression. *Advances in Data Analysis and Classification*, 11(4):659–690, Dec 2017.
12. G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.
13. P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.
14. Z.-G. Su, T. Denoeux, Y.-S. Hao, and M. Zhao. Evidential K-NN classification with enhanced performance via optimizing a class of parametric conjunctive t-rules. *Knowledge-Based Systems*, 142:7–16, 2018.
15. L. M. Zouhal and T. Denœux. An evidence-theoretic k-NN rule with parameter optimization. *IEEE Trans. on Systems, Man and Cybernetics C*, 28(2):263–271, 1998.