Generalised max entropy classifiers

Fabio Cuzzolin[1]

Oxford Brookes University, UK
fabio.cuzzolin@brookes.ac.uk

Abstract. In this paper we propose a generalised maximum-entropy classification framework, in which the empirical expectation of the feature functions is bounded by the lower and upper expectations associated with the lower and upper probabilities associated with a belief measure. This generalised setting permits a more cautious appreciation of the information content of a training set. We analytically derive the Karush-Kuhn-Tucker conditions for the generalised max-entropy classifier in the case in which a Shannon-like entropy is adopted.

Keywords: Classification · Max entropy · Constrained optimisation.

1 Introduction

The emergence of new challenging real-world applications has exposed serious issues with current approaches to model adaptation in machine learning. Existing theory and algorithms focus on fitting the available training data, but cannot provide worst-case guarantees in mission-critical applications. Vapnik's statistical learning theory is useless for model selection, as the bounds on generalisation errors it predicts are too wide to be useful, and rely on the assumption that training and testing data come from the same (unknown) distribution. The crucial question is: what exactly can one infer from a training set?

Max entropy classifiers [19] provide a significant example, due to their simplicity and widespread application. There, the entropy of the sought joint (or conditional) probability distribution of data and class is maximised, following the *maximum entropy principle* that the least informative distribution which matches the available evidence should be chosen. Having picked a set of *feature functions*, selected to efficiently encode the training information, the joint distribution is subject to the constraint that their empirical expectation equals that associated with the max entropy distribution. The assumptions that (i) training and test data come from the same probability distribution, and that (ii) the empirical expectation of the training data is correct, and the model expectation should match it, are rather strong, and work against generalisation power.

A way around this issue is to adopt as models convex sets of probability distributions, rather than standard probability measures. Random sets, in particular, are mathematically equivalent to a special class of credal sets induced by probability mass assignments on the power set of the sample space. When random sets are defined on finite domain, they are often called *belief functions*

[20]. One can then envisage a robust theory of learning based on generalising traditional statistical learning theory in order to allow for test data to be sampled from a *different* probability distribution than the training data, under the weaker assumption that both belong to the same random set.

In this paper we make a step in that direction by generalising the max entropy classification framework. We take the view that a training set does not provide, in general, sufficient information to precisely estimate the joint probability distribution of class and data. We assume instead that a belief measure can be estimated, providing lower and upper bounds on the joint probability of data and class. As in the classical case, an appropriate measure of entropy for belief measures is maximised. In opposition to the classical case, however, the empirical expectation of the chosen feature functions is assumed to be *compatible* with lower and upper bounds associated with the sought belief measure. This leads to a constrained optimisation problem with inequality constraints, rather than equality ones, which needs to be solved by looking at the Karush-Kuhn-Tucker (KKT) conditions. Due to the concavity of the objective function and the convexity of the constraints, KKT conditions are both necessary and sufficient.

Related work. A significant amount of work has been conducted in the past on machine learning approaches based on belief theory. Most efforts were directed at developing clustering tools, including evidential clustering [4], evidential and belief C-means [15]. Ensemble classification [23], in particular, has been extensively studied. Concerning classification, Denoeux [5] proposed in a seminal work a k-nearest neighbor classifier based on belief theory. Relevantly to this paper, interesting work has been conducted to generalise the framework of decision trees to situations in which uncertainty is encoded by belief functions, mainly by Elouedi and co-authors [7], and Vannoorenberghe and Denoeux [22].

Paper outline. After reviewing in Section 2 max-entropy classification, we recall in Section 3 the necessary notions of belief theory. In Section 4 the possible generalisations of Shannon's entropy to the case of belief measures are reviewed. In Section 5 the generalised max-entropy problem is formulated, together with the associated Kush-Karun-Tucker conditions. It is shown that for several generalised measures of entropy the KKT conditions are necessary and sufficient for the optimalised of generalised max-entropy (Section 5.1). In Section 5.2 we derive the analytical expression of the system of KKT conditions for the case of a Shannon-like entropy for belief measures. Section 6 concludes the paper.

2 Max-entropy classifiers

The objective of *maximum entropy classifiers* is to maximise the Shannon entropy of the conditional classification distribution $p(C_k|x)$, where $x \in X$ is the observable and $C_k \in \mathcal{C} = \{C_1, ..., C_K\}$ is the associated class.

Given a training set in which each observation is attached a class, namely: $\mathcal{D} = \{(x_i, y_i), i = 1, ..., N | x_i \in X, y_i \in \mathcal{C}\}$, a set M of *feature maps* is designed, $\phi(x, C_k) = [\phi_1(x, C_k), \cdots, \phi_M(x, C_k)]'$ whose values depend on both the object observed and its class. Each feature map $\phi_m : X \times \mathcal{C} \to \mathbb{R}$ is then a random

variable whose expectation is: $E[\phi_m] = \sum_{x,k} p(x, C_k)\phi_m(x, C_k)$. In opposition, the *empirical* expectation of ϕ_m is: $\hat{E}[\phi_m] = \sum_{x,k} \hat{p}(x, C_k)\phi_m(x, C_k)$, where $\hat{p}$ is a histogram constructed by counting occurrences of the pair (x, C_k) in the training set: $\hat{p}(x, C_k) = \frac{1}{N}\sum_{(x_i,y_i)\in\mathcal{D}} \delta(x_i = x \wedge y_i = C_k)$. The theoretical expectation $E[\phi_m]$ can be approximated by decomposing $p(x, C_k) = p(x)p(C_k|x)$ via Bayes' rule, and approximating the (unknown) prior of the observations $p(x)$ with the empirical prior $\hat{p}$, i.e., the histogram of observed values in the training set: $\tilde{E}[\phi_m] = \sum_{x,k} \hat{p}(x)p(C_k|x)\phi_m(x, C_k)$.

Definition 1. *Given a training set* $\mathcal{D} = \{(x_i, y_i), i = 1, ..., N|x_i \in X, y_i \in \mathcal{C}\}$ *related to problem of classifying* $x \in X$ *as belonging to one of the classes* $\mathcal{C} = \{C_1, ..., C_K\}$*, the* max entropy classifier *is the conditional probability* $p^*(C_k|x)$ *such that:* $p^*(C_k|x) \doteq \arg\max_{p(C_k|x)} H_s(P)$*, where* H_s *is the traditional Shannon entropy, subject to:* $\tilde{E}_p[\phi_m] = \hat{E}[\phi_m]\ \forall m = 1, ..., M$.

The constraint requires the classifier to be consistent with the empirical frequencies of the features in the training set, while seeking the least informative probability distribution that does so. The solution of the maximum entropy classification problem (Definition 1) is the so-called *log-linear model*: $p^*(C_k|x) = \frac{1}{Z_\lambda(x)} e^{\sum_m \lambda_m \phi_m(x, C_k)}$, where $\lambda = [\lambda_1, ..., \lambda_M]'$ are the Lagrange multipliers associated with the linear constraints $\tilde{E}_p[\phi_m] = \hat{E}[\phi_m]$, and $Z_\lambda(x)$ is a normalisation factor. The related classification function is: $y(x) = \arg\max_k \sum_m \lambda_m \phi_m(x, C_k)$, i.e., x is assigned the class which maximises the linear combination of the feature functions with coefficients λ.

3 Belief functions

Definition 2. *A* basic probability assignment *(BPA) [1] over a discrete set* Θ *is a function* $m : 2^\Theta \to [0, 1]$ *defined on* $2^\Theta = \{A \subseteq \Theta\}$ *such that:* $m(\emptyset) = 0$, $\sum_{A\subset\Theta} m(A) = 1$. *The* belief function *(BF) associated with a BPA* $m : 2^\Theta \to [0, 1]$ *is the set function* $Bel : 2^\Theta \to [0, 1]$ *defined as:* $Bel(A) = \sum_{B\subseteq A} m(B)$.

The elements of the power set 2^Θ associated with non-zero values of m are called the *focal elements* of m. For each subset ('event') $A \subset \Theta$ the quantity $Bel(A)$ is called the *degree of belief* that the outcome lies in A, and represents the total belief committed to a set of outcomes A by the available evidence m. Dually, the *upper probability* of A: $Pl(A) \doteq 1 - Bel(\bar{A})$, $\bar{A} = \Theta \setminus A$, expresses the 'plausibility' of a proposition A or, in other words, the amount of evidence *not against* A [3]. The *plausibility function* $Pl : 2^\Theta \to [0, 1]$ thus conveys the same information as Bel, and can be expressed as: $Pl(A) = \sum_{B\cap A\neq\emptyset} m(B) \geq Bel(A)$.

Belief functions are mathematically equivalent to a special class of credal sets (convex sets of probability measures), as each BF Bel is associated with the set $\mathcal{P}[Bel] = \{P : P(A) \geq Bel(A)\}$ of probabilities dominating it. Its centre of mass is the *pignistic function* $BetP[Bel](x) = \sum_{A\ni x} m(A)/|A|$, $x \in \Theta$. Given a function $f : \Theta \to \mathbb{R}$, the *lower expectation* and *upper expectation* of f w.r.t. Bel are, respectively: $E_{Bel*}[f] \doteq \inf_{P\in\mathcal{P}[Bel]} E_P[f] = \sum_{A\subseteq\Theta} m(A) \inf_{x\in A} f(x)$, $E^*_{Bel}[f] \doteq \sup_{P\in\mathcal{P}[Bel]} E_P[f] = \sum_{A\subseteq\Theta} m(A) \sup_{x\in A} f(x)$.

4 Measures of generalised entropy

The issue of how to assess the level of uncertainty associated with a belief function [10] is not trivial, as authors such as Yager and Klir argued that there are several facets to uncertainty, such as *conflict* (or discord, dissonance) and *non-specificity* (also called vagueness, ambiguity or imprecision).

Some measures are directly inspired by Shannon's entropy of probability measures: $H_s[p] = -\sum_{x\in\Theta} p(x)\log p(x)$. While Nguyen's measure is a direct generalisation in which probability values are replaced by mass values [17]: $H_n[m] = -\sum_{A\in\mathcal{F}} m(A)\log m(A)$, where $\mathcal{F}$ is the list of focal elements of m, in Yager's entropy [24] probabilities are (partly) replaced by plausibilities: $H_y[m] = -\sum_{A\in\mathcal{F}} m(A)\log Pl(A)$. Hohle's *measure of confusion* [9] is the dual measure: $H_o[m] = -\sum_{A\in\mathcal{F}} m(A)\log Bel(A)$. All such measures only capture the 'conflict' portion of uncertainty. Other measures are designed to capture the *specificity* of belief measures, i.e., the degree of concentration of the mass assigned to focal elements. A first such measure was due to Klir, Dubois & Prade [6]: $H_d[m] = \sum_{A\in\mathcal{F}} m(A)\log|A|$, and can be considered as a generalization of Hartley's entropy ($H = \log(|\Theta|)$) to belief functions. A more sophisticated proposal by Pal [18]: $H_a[m] = \sum_{A\in\mathcal{F}} m(A)/|A|$, assesses the dispersion of the evidence and is linked to the pignistic transform. A final proposal based on the commonality function $Q(A) = \sum_{B\supset A} m(B)$ is due to Smets: $H_t = \sum_{A\in\mathcal{F}} \log(\frac{1}{Q(A)})$.

Composite measures, such as Lamata and Moral's $H_l[m] = H_y[m] + H_d[m]$ [14], as designed to capture both entropy and specificity. Klir & Ramer [13] proposed a 'global uncertainty measure' defined as: $H_k[m] = D[m] + H_d[m]$, where: $D(m) = -\sum_{A\in\mathcal{F}} m(A)\log[\sum_{B\in\mathcal{F}} m(B)\frac{|A\cap B|}{|B|}]$. Pal et al [18] argued that none of these composite measures is really satisfactory, as they do not admit a unique maximum and there is no sounding rationale for simply adding conflict and non-specificity measures together.

In the credal interpretation of belief functions, Harmanec and Klir's *aggregated uncertainty* (AU) [8] is defined as the maximal Shannon entropy of all the probabilities consistent with the given BF: $H_h[m] = \max_{P\in\mathcal{P}[Bel]}\{H_s[P]\}$. $H_h[m]$ is the minimal measure meeting a set of rationality requirements which include: symmetry, continuity, expansibility, subadditivity, additivity, monotonicity, normalisation. Similarly, Maeda and Ichihashi [16] proposed a composite measure $H_i[m] = H_h[m] + H_d[m]$ whose first component consists of the maximum entropy of the set of probability distributions consistent with m, and whose second part is the generalized Hartley entropy. As both H_h and H_i have high computational complexity, Jousselme et al [11] proposed an *ambiguity measure* (AM), as the classical entropy of the pignistic function: $H_j[m] = H_s[BetP[m]]$.

Jirousek and Shenoy [10] analysed all these proposal in 2016, assessing them versus a number of significant properties, concluding that only the Maeda-Ichihashi proposal meets all these properties. The issue remains still unsettled. In the following we will adopt a straighforward generalisation of Shannon's entropy, and a few selected proposals based on their concavity property.

5 Generalised max-entropy problem

Technically, in order to generalise the max-entropy optimisation problem (Definition 1) to the case of belief functions, we need to: (i) choose an appropriate measure of entropy for belief function as the objective function; (ii) revisit the constraints that the (theoretical) expectations of the feature maps are equal to the empirical ones computed over the training set.

As for (ii), it is sensible to require that the empirical expectation of the feature functions is bracketed by the lower and upper expectations associated with the sought belief function $Bel : 2^{X \times \mathcal{C}} \to [0, 1]$. In this paper we only make use of the 2-monotonicity of belief functions, and write:

$$\sum_{(x,C_k)} Bel(x, C_k)\phi_m(x, C_k) \leq \hat{E}[\phi_m] \leq \sum_{(x,C_k)} Pl(x, C_k)\phi_m(x, C_k) \tag{1}$$

$\forall m = 1, ..., M$, as we only consider probability intervals on singleton elements $(x, C_k) \in X \times \mathcal{C}$. Fully fledged lower and upper expectations (cfr. Section 3), which express the full monotonicity of BFs, will be considered in future work.

Going even further, should constraints of the form (1) be enforced on all possible subsets $A \subset X \times \mathcal{C}$, rather than just singleton pairs (x, C_k)? This goes back to the question of what information does a training set actually carry. More general constraints would require extending the domain of feature functions to set values – we will investigate this idea in the near future as well.

5.1 Formulation and Karush-Kuhn-Tucker (KKT) conditions

In the same classification setting of Section 2, the *maximum belief entropy classifier* is the joint belief measure $Bel^*(x, C_k) : 2^{X \times \mathcal{C}} \to [0, 1]$ which solves the following optimisation problem: $Bel^*(x, C_k) \doteq \arg\max_{Bel(x,C_k)} H(Bel)$ subject to the inequality constraints (1), where H is an appropriate measure of entropy for belief measures. As the above optimisation problem involves inequality constraints (1), as opposed to the equality constraints of traditional max entropy classifiers, we need to analyse the Karush-Kuhn-Tucker (KKT) [12] necessary conditions for a belief function Bel to be an optimal solution to the problem.

Definition 3. *Suppose that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $g_i : \mathbb{R}^n \to \mathbb{R}$ and $h_j : \mathbb{R}^n \to \mathbb{R}$ of a nonlinear optimisation problem $\arg\max_x f(x)$ subject to: $g_i(x) \leq 0$ $i = 1, ..., m$, $h_j(x) = 0$ $j = 1, ..., l$ are continuously differentiable at a point x^*. If x^* is a local optimum, under appropriate regularity conditions then there exist constants μ_i, $(i = 1, \ldots, m)$ and λ_j $(j = 1, \ldots, l)$, called* KKT multipliers*, such that the following conditions hold:*

1. Stationarity*: $\nabla f(x^*) = \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{l} \lambda_j \nabla h_j(x^*)$;*
2. Primal feasibility*: $g_i(x^*) \leq 0$ $\forall i = 1, \ldots, m$, and $h_j(x^*) = 0$, $\forall j = 1, \ldots, l$;*
3. Dual feasibility*: $\mu_i \geq 0$ for all $i = 1, \ldots, m$;*
4. Complementary slackness*: $\mu_i g_i(x^*) = 0$ for all $i = 1, \ldots, m$.*

Crucially, the KKT conditions are also sufficient whenever the objective function f is concave, the inequality constraints g_i are continuously differentiable convex functions, and the equality constraints h_j are affine[1].

Theorem 1. *If either $H_t, H_n, H_d, H_s[Bel]$ or $H_s[Pl]$ is adopted as measure of entropy, the generalised max entropy optimisation problem has concave objective function and convex constraints. Therefore, the KKT conditions are sufficient for the optimality of its solution(s).*

Concavity of the entropy objective function. It is well known that Shannon's entropy is a concave function of probability distributions, represented as vectors of probability values[2]. Furthermore: any linear combination of concave functions is concave; a monotonic and concave function of a concave function is still concave; the logarithm is a concave function.

As shown by Smets [21], the transformations which map mass vectors to vectors of belief (and commonality) values are linear, as they can be expressed in the form of matrices. In particular, $\boldsymbol{bel} = BfrM\boldsymbol{m}$, where $BfrM$ is a matrix whose (A, B) entry is: $BfrM(A, B) = 1$ if $B \subseteq A$, 0 otherwise, and $\boldsymbol{bel}, \boldsymbol{m}$ are vectors collecting the belief (mass) values of all events $A \subseteq \Theta$. The same can be said of the mapping $\boldsymbol{q} = QfrM\boldsymbol{m}$ between a mass vector and the associated commonality vector. As a consequence, belief, plausibility and commonality are all linear (and therefore concave) functions of a mass vector.

Using this matrix representation, it is easy to conclude that several of the entropies defined in Section 4 are indeed concave. In particular, Smets' specificity measure $H_t = \sum_A \log(\frac{1}{Q(A)})$ is concave, as a linear combination of concave functions. Nguyen's entropy $H_n = -\sum_A m(A)\log(m(A)) = H_s[m]$ is also concave, as the Shannon's entropy of a mass assignment. Dubois and Prade's measure $H_d = \sum_A m(A)\log(|A|)$ is also concave with respect to m, as a linear combination of mass values. Direct applications of Shannon's entropy function to Bel and Pl: $H_{Bel}[m] = H_s[Bel] = \sum_{A\subseteq\Theta} Bel(A)\log(\frac{1}{Bel(A)})$, $H_{Pl}[m] = H_s[Pl] = \sum_{A\subseteq\Theta} Pl(A)\log(\frac{1}{Pl(A)})$ are also trivially concave, due to the concavity of the entropy function and to the linearity of the mapping from m to Bel, Pl. Drawing conclusions on the other measures is less immediate, as they involve products of concave functions (which are not, in general, guaranteed to be concave).

Convexity of the interval expectation constraints. As for the contraints (1) of the generalised max entropy problem, we first note that (1) can be decomposed into the following pair of constraints: $g^1_m(m) \doteq \sum_{x,k} Bel(x, C_k)\phi_m(x, C_k) - \hat{E}[\phi_m] \leq 0$, $g^2_m(m) = \sum_{x,k} \phi_m(x, C_k)[\hat{p}(x, C_k) - Pl(x, C_k)] \leq 0$ for all $m = 1, ..., M$. The first inequality constraint is a linear combination of linear functions of the sought mass assignment $m^* : 2^{\mathbb{X}\times\mathcal{C}} \to [0, 1]$ (since Bel^* results from applying a matrix transformation to m^*). As $\boldsymbol{pl} = 1 - J\boldsymbol{bel} = 1 - JBfrM\boldsymbol{m}$, constraint g^2_m is also a linear combination of mass values. Hence, as linear function, constraints g^1_m and g^2_m are both concave and convex.

[1] More general sufficient conditions can be given in terms of *invexity* [2] requirements.
[2] http://projecteuclid.org/euclid.lnms/1215465631

5.2 Belief max-entropy classifier for Shannon's entropy

For the Shannon-like entropy Condition 1. (stationarity), applied to the sought optimal BF $Bel^* : 2^{X \times \mathcal{C}} \rightarrow [0,1]$, reads as: $\nabla H_{Bel}(Bel^*) = \sum_{m=1}^{M} \mu_m^1 \nabla g_m^1(Bel^*) + \mu_m^2 \nabla g_m^2(Bel^*)$. The components of ∇H_{Bel} are the partial derivatives of the entropy with respect to the mass values $m(\overline{B})$, for all $\overline{B} \subseteq \Theta$. They read as:

$$\frac{\partial H_{Bel}}{\partial m(\overline{B})} = \frac{\partial}{\partial m(\overline{B})} \sum_{A \supseteq \overline{B}} \left[-(\sum_{B \subseteq A} m(B)) \log(\sum_{B \subseteq A} m(B)) \right] = -\sum_{A \supseteq \overline{B}} [1 + \log Bel(A)].$$

As for $\nabla g_m^1(Bel^*)$ we have: $\frac{\partial g_m^1}{\partial m(\overline{B})} = \frac{\partial}{\partial m(\overline{B})} \sum_{(x,C_k) \in \Theta} Bel(x, C_k) \phi_m(x, C_k) - \hat{E}[\phi_m] = \frac{\partial}{\partial m(\overline{B})} \sum_{(x,C_k) \in \Theta} m(x, C_k) \phi_m(x, C_k) - \hat{E}[\phi_m]$ which is equal to $\phi_m(x, C_k)$ for $\overline{B} = \{(x, C_k)\}$, 0 otherwise[3]. As for the second set of constraints: $\frac{\partial g_m^2}{\partial m(\overline{B})} = \frac{\partial}{\partial m(\overline{B})} \sum_{(x,C_k) \in \Theta} \phi_m(x, C_k)[\hat{p}(x, C_k) - Pl(x, C_k)]$ which, recalling that $Pl(x, C_k) = \sum_{B \cap \{(x,C_k)\} \neq \emptyset} m(B)$, becomes equal to $= -\sum_{(x,C_k) \in \overline{B}} \phi_m(x, C_k)$.

Assembling all our results, the KKT stationarity conditions for the generalised, belief-theoretical maximum entropy problem amount to, for all $\overline{B} \subset X \times \mathcal{C}$:

$$\begin{cases} -\sum_{A \supseteq \overline{B}} [1 + \log Bel(A)] = \sum_{m=1}^{M} \phi_m(\overline{x}, \overline{C}_k)[\mu_m^1 - \mu_m^2], |\overline{B} = \{(\overline{x}, \overline{C}_k)\}| = 1, \\ -\sum_{A \supseteq \overline{B}} [1 + \log Bel(A)] = \sum_{m=1}^{M} \mu_m^2 \sum_{(x,C_k) \in \overline{B}} \phi_m(x, C_k), |\overline{B}| > 1. \end{cases} \tag{2}$$

The other conditions are, $\forall m = 1, ..., M$, (1) (primal feasibility), $\mu_m^1, \mu_m^2 \geq 0$ (dual feasibility), and complementary slackness: $\mu_m^1 \sum_{(x,C_k) \in \Theta} Bel(x, C_k) \phi_m(x, C_k) - \hat{E}[\phi_m] = 0$, $\mu_m^2 \sum_{(x,C_k) \in \Theta} \phi_m(x, C_k)[\hat{p}(x, C_k) - Pl(x, C_k)] = 0$.

6 Conclusions

In this paper we proposed a generalisation of the max entropy classifier entropy in which the assumptions that test and training data are sampled by a same probability distribution, and that the empirical expectation of the feature functions is 'correct' are relaxed in the formalism of belief theory. We also studied the conditions under which the associated KKT conditions are necessary and sufficient for the optimality of the solution. Much work remains: (i) providing analytical model expressions, similar to log-linear models, for the Shannon-like and other major entropy measures for belief functions; (ii) analysing the case in which the full lower and upper expectations are plugged in; (iii) comparing the resulting classifiers; (iv) analysing a formulation based on the least commitment principle, rather than max entropy, for the objective function to optimise; finally, (v) relaxing the constraint that feature functions be defined on singleton pairs (x, C_k), in a further generalisation of this important framework.

[3] If we could define feature functions over non singletons subsets $A \subseteq \Theta$, this would simply generalise to $\phi(\overline{B})$ for all $\overline{B} \subseteq \Theta$.

References

1. Augustin, T.: Modeling weak information with generalized basic probability assignments. In: Data Analysis and Information Systems, 101–113, Springer, 1996.
2. Ben-Israel, A. et al: What is invexity? J. Austral. Math. Soc. Ser. B 28:1–9, 1986.
3. Cuzzolin, F.: Three alternative combinatorial formulations of the theory of evidence. Intelligent Data Analysis 14(4):439–464, 2010.
4. Denoeux, T. and Masson, M.-H.: EVCLUS: Evidential Clustering of Proximity Data. IEEE Trans Syst Man Cybern B 34(1):95-109, 2004.
5. Denœux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans Syst Man Cybern 25(5):804-813, 1995.
6. Dubois, D., Prade, H.: Properties of measures of information in evidence and possibility theories. Fuzzy Sets Syst 100:35–49, 1999.
7. Elouedi, Z., Mellouli, K., Smets, P.: Belief decision trees: theoretical foundations. Int J Approx Reason 28(23):91–124, 2001.
8. Harmanec, D., Klir, G.J.: Measuring total uncertainty in Dempster-Shafer theory: A novel approach. Int J Gen Syst 22(4):405–419, 1994.
9. Hohle, U.: Entropy with respect to plausibility measures. In: Proceedings of the 12th IEEE Symposium on Multiple-Valued Logic., pp. 167–169, 1982.
10. Jirousek, R., Shenoy, P.P.: Entropy of belief functions in the Dempster-Shafer theory: A new perspective. In: Proceedings of BELIEF, pp. 3-13, 2016.
11. Jousselme, A.L. et al: Measuring ambiguity in the evidence theory. IEEE Trans Syst Man Cybern A 36(5):890–903, 2006.
12. Karush, W.: Minima of functions of several variables with inequalities as side constraints. MSc Dissertation, Dept. of Mathematics, Univ. of Chicago, 1939.
13. Klir, G.J.: Measures of uncertainty in the Dempster-Shafer theory of evidence. In: Advances in the Dempster-Shafer theory of evidence, pp. 35–49, 1994.
14. Lamata, M.T., Moral, S.: Measures of entropy in the theory of evidence. Int J Gen Syst 14(4):297–305, 1988.
15. Liu, Z. et al: Belief C-means: An extension of fuzzy c-means algorithm in belief functions framework. Pattern Recognit Lett 33(3):291–300, 2012.
16. Maeda, Y., Ichihashi, H.: An uncertainty measure with monotonicity under the random set inclusion. Int J Gen Syst 21(4):379–392, 1993.
17. Nguyen, H.: On entropy of random sets and possibility distributions. In: The Analysis of Fuzzy Information, pp. 145–156, 1985.
18. Pal, N.R., Bezdek, J.C., Hemasinha, R.: Uncertainty measures for evidential reasoning ii: A new measure of total uncertainty. Int J Approx Reason 8:1–16, 1993.
19. Pietra, S.D., et al: Inducing features of random fields. IEEE Trans Pattern Anal Mach Intell 19(4):380–393, 1997.
20. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, 1976.
21. Smets, P.: The application of the matrix calculus to belief functions. Int J Approx Reason 31(1-2):1–30, 2002.
22. Vannoorenberghe, P. et al: Handling uncertain labels in multiclass problems using belief decision trees. In: Proceedings of IPMU, 2002.
23. Xu, L. et al: Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans Syst Man Cybern 22(3):418–435, 1992.
24. Yager, R.R.: Entropy and specificity in a mathematical theory of evidence. Int J Gen Syst 9:249–260, 1983.