A probabilistic approach to case-based inference

Martin Anthony* Joel Ratsaby†

Abstract

The central problem in case based reasoning (CBR) is to infer a solution for a new problem-instance by using a collection of existing problem-solution cases. The basic heuristic guiding CBR is the hypothesis that similar problems have similar solutions. Recently, some attempts at formalizing CBR in a theoretical framework have been made, including work by Hüllermeier who established a link between CBR and the probably approximately correct (PAC) theoretical model of learning in his 'case-based inference' (CBI) formulation. In this paper we develop further such probabilistic modelling, framing CBI it as a multi-category classification problem. We use a recently-developed notion of geometric margin of classification to obtain generalization error bounds.

*Department of Mathematics, The London School of Economics and Political Science, Houghton Street, London WC2A2AE, U.K., m.anthony@lse.ac.uk

†Electrical and Electronics Engineering Department, Ariel University, Ariel 40700, ISRAEL, ratsaby@ariel.ac.il

1 Introduction and related work

The basic problem in case based reasoning (CBR) is to infer a solution for a new problem-instance by using a collection of existing problem-solution cases [1]. (We will henceforth use 'problem' for 'problem instance'.) The basic heuristic that guides CBR is the hypothesis that similar problems have similar solutions (see [2], for example). The area of CBR research has had practical success and has been shown to be widely applicable [3]. The well known methodological framework of case-based reasoning divides CBR into four main steps (referred to as the R^4 framework): retrieve, reuse, refine and retain [2].

There have been a number of attempts to develop a sound theoretical basis for CBR. Significant recent work, due to Hüllermeier [4], makes a connection between CBR and the probably approximately correct (PAC) theoretical model of learning [5]. Hüllermeier defines case-based reasoning as a prediction process, which allows him to make the connection between CBR and the learning based on a sample. He calls this framework *case-based inference* (CBI) and it aims to solve the 'retrieve' and 'reuse' steps of the R^4 framework. Given a new problem to be solved, CBI aims just to produce a 'promising' set of solutions for use by the remaining two steps of the R^4 framework. The last two stages of the R^4 framework use not just the set of candidate solutions but also domain-knowledge, user input and further problem-solving strategies [2]. As noted in Section 5.4 of [6], these steps *adapt* the set of promising solutions into a solution that fits the existing problem.

In this paper, we continue work in the direction inspired by [2], probabilistically modelling case-based inference as a multi-category classification problem. We use a recently-developed notion of geometric margin of classification, called width, to obtain generalization error bounds. This notion has recently been used in [7] to exploit regularity in training samples for the problem of classification learning in finite metric spaces. The main results in the current paper are bounds on the error of case-based learning which involve the sample width.

Dubois and Prade [8] and Dubois *et al.* [9] attempted to provide a formal model of CBR which is based on fuzzy logic. The similarity between two

problems (or two solutions) is represented by a fuzzy relations. There is no learning process for determining these relations. Our model differs from theirs in that we learn from examples to produce a set of candidate solutions for input problems; and we do not employ fuzzy logic, but statistical learning under the PAC framework.

Ontañón and Plaza [10] introduce a model of knowledge transfer for case-based inference part of CBR. It produces, from retrieved cases (cases whose problems are similar to the given problem), a set of conjectures (or incomplete solutions) rather than actual solutions. A conjecture may require further adaptation, for instance using some domain specific rules, in order to produce a solution. (Their model can deal with cases where there is no clear distinction between a problem and solution.) Our model differs from theirs in that it produces a set of complete solutions for the input problem (rather than conjectures); and our model expands on the CBI framework of Hüllermeier, and hence we have two separate spaces, one for solutions and one for problems.

In Section 2, we start by describing Hüllermeier's framework of CBI, where the goal is to predict a 'credible' or promising set of solutions for a given input problem instance. We outline and explain our contribution and its connections with this framework. The key idea is that we model CBI as a supervised learning problem. Section 3 describes a probabilistic model that is the basis of our analysis. We redefine what is meant by a credible set in this context and we provide a mathematical formalism for measuring the success of a method for predicting credible sets. Section 4 presents some recent results on the generalization accuracy of learning multi-category classifiers defined on metric spaces, and provides results on which we draw for the conclusions of this paper. Section 5 describes in detail the important transformation of learning CBI to the problem of supervised learning. Section 6 provides bounds on the error of learning CBI. These bounds can serve as a guiding criterion for the design of successful algorithms.

One main contribution is to show how learning CBI over the wide spectrum of complex and unstructured CBR domains can be transformed to standard supervised learning problems. A further contribution is in showing how the large-width advantage (familiar from the branch of learning theory known as large-margin learning) can also be realised for learning CBI.

2 Case-based inference (CBI)

In the Introduction, we mentioned that CBI infers as an output a set of candidate, or 'promising', solutions rather than solving the full CBR problem by predicting a single specific solution. This is at the basis of what Hüllermeier [4] calls *approximate reasoning.* We now describe his framework (using slightly different notation).

2.1 Hüllermeier's CBI framework

In the general set-up of Hüllermeier's case-based inference, there is a problem space, denoted by $\mathcal{X}$, and a solution space, denoted by $\mathcal{Y}$. We define $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. The problem space and solution space may be very general; in particular, not only finite-dimensional vector spaces (as those that are common in supervised learning) but also problems described by more complex structures like trees, graphs, or plans. Each of the spaces, $\mathcal{X}$, $\mathcal{Y}$, has a similarity function, $\mathrm{sim}_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to [0,1]$ and $\mathrm{sim}_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to [0,1]$, respectively. These are reflexive and symmetric; that is, $\mathrm{sim}_{\mathcal{X}}(x,x) = 1$, and $\mathrm{sim}_{\mathcal{X}}(x,x') = \mathrm{sim}_{\mathcal{X}}(x',x)$, and similarly for $\mathrm{sim}_{\mathcal{Y}}$. The goal of case-based inference in [4] can be described as follows.

Goal of CBI in [4]: Given a sample $\{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ (also referred to as a case-base), consisting of problem-solution pairs, and given a new problem instance x, produce for it a subset of solutions (subset of $\mathcal{Y}$) called a *credible set*, that contains some (possibly all) solutions for the problem x.

An underlying assumption is that there exists some unknown relationship between the level of similarity of pairs of problems and the similarity of their solutions. Hüllermeier[4] represents this by a *similarity profile* σ, mapping from $[0,1]$ to $[0,1]$ and defined by

$$\sigma(\alpha) := \inf_{x,x' \in \mathcal{X} : \mathrm{sim}_{\mathcal{X}}(x,x') = \alpha} \mathrm{sim}_{\mathcal{Y}}(y,y')$$

where (x,y), $(x',y') \in \mathcal{Z}$ are two problem-solution pairs. This function σ represents in a formal way the CBR assumption that similar problems have

similar solutions, since given any pair of problems that are similar by a value of α, their solutions must be at least similar by a level of $\sigma(\alpha)$.

Theoretically speaking, if one knows σ then, for a given problem x, one can produce a 'credible' set of solutions, which is defined as

$$C(x) := \bigcap_{i=1}^{m} \Gamma_\sigma \left(z_i, x\right) \tag{1}$$

where $\Gamma_\sigma(z_i, x) \subseteq \mathcal{Y}$ is given by

$$\Gamma_\sigma(z_i, x) = \{y : \text{sim}_\mathcal{Y}(y, y_i) \geq \sigma(\text{sim}_\mathcal{X}(x_i, x))\}\,.$$

Since σ is unknown, the aim in this framework is to learn a hypothesis function which approximates σ. This is called a 'similarity hypothesis' $h : [0, 1] \to [0, 1]$. Substituting h for σ in (1) yields the following *hypothesis set*:

$$C_h(x) := \bigcap_{i=1}^{m} \Gamma_h \left(z_i, x\right) \tag{2}$$

where

$$\Gamma_h(z_i, x) = \{y : \text{sim}_\mathcal{Y}(y, y_i) \geq h(\text{sim}_\mathcal{X}(x_i, x))\}\,.$$

If it is possible to guarantee that $h(\alpha) \leq \sigma(\alpha)$, for all $\alpha \in [0, 1]$ then $C(x) \subseteq C_h(x)$, which means that the hypothesis set must be a credible set since it contains the credible set of (1). In reality, one cannot guarantee this, and hence $C_h(x)$ may not be a credible set. But one can state probabilistic confidence levels that this holds if h is chosen from a suitable class of *empirical similarity* hypotheses. It is not intended here to describe the details of this class of hypothesis but we note that the class consists of hypotheses that are generated by the classic candidate-elimination algorithm (see algorithm Find-S of [11]). Hypotheses h in this class are piecewise-constant on $[0, 1]$ and the generalization heuristic of the Find-S algorithm is used to learn a good h.

To explain what a good h is, let us first define the notion of consistency. A hypothesis h is consistent with the sample if for all $1 \leq i, j \leq m$, $\text{sim}_\mathcal{X}(x_i, x_j) = \alpha$ implies that $\text{sim}_\mathcal{Y}(y_i, y_j) \geq h(\alpha)$.

For any two hypotheses h and h', h is said to be *stronger* than h' if $h(\alpha) \geq h'(\alpha)$ for all $\alpha \in [0,1]$. This implies that for any problem instance x the corresponding set $C_h(x)$ is contained in $C_{h'}(x)$, that is, the set corresponding to h is *more specific* than the set that corresponds to h'. Then, the goal of learning in [4] is as follows.

Goal of learning in [4]: Find the strongest h that is consistent with the sample.

Hüllermeier's Algorithm 1 (see [4]) achieves this goal by producing a consistent hypothesis which is piecewise constant over the subintervals of a partition of $[0,1]$ which is fixed before the learning starts. The algorithm converges to the strongest hypothesis of this kind. Hüllermeier obtains a bound on the probability that $C_h(x)$ does not contain a solution of x. The bound is of the form $O(1/m)$ and is linearly proportional to the size of the partition. This means that as m increases, there is a higher probability that $C_h(x)$ is a credible set.

2.2 Our contribution

Learning is at the core of the above-mentioned inference framework. It is responsible for generating a set C_h that with high probability contains solutions for future problem instances. While the learning approach taken by Hüllermeier is sensible, it leads to sets that are constrained to take a special form as defined by (2) and based on hypotheses h that are piecewise constant over a partition which is chosen in advance, based on heuristic domain knowledge. The resulting mapping that outputs a credible solution set from an input problem instance is very particular and hence potentially introduces an inductive bias [11], which is a known cause for less accurate learning [5], meaning that there is more chance that the learnt mapping produces a set which is not credible.

To circumvent that, in this paper we extend the CBI model such that no *a priori* inductive bias is placed through a choice of a particular class of mappings. We represent the CBI as a multi-category classification problem where the class of hypotheses is different from the one used by Hüllermeier.

Our class of hypotheses models an extremely rich non-parametric class of mappings from problem space to subsets of the solution space, while respecting the CBR assumption that similar problems have similar solutions. As we describe in Sections 4 and 5 we consider vector-valued functions which represent distances from a given input problem instance x to general labeled subsets S of the space. The fact that these sets can be any subsets of the metric space makes the hypothesis class extremely rich. In this paper we do not offer any particular algorithm to search over this space but we provide theorems that apply to *any* hypothesis in this class: hence they apply to *any* learning algorithm over this class (the inductive bias enters from the choice of a supervised learning algorithm). If, for example, we use as sets S the sample points that correspond to the case-base (referred to as auxiliary samples in Section 5) then the mapping h is based on a multi-category nearest-neighbor rule.

The overall goal of CBI remains as in Hüllermeier's framework, but the goal of learning is different from that of Hüllermeier. We have provided generalization error bounds in Section 6 and the goal of learning becomes that of producing by any means and with any algorithm hypotheses that give low value to these bounds. Our error bounds can provide criteria for an algorithm to minimize. In particular, this motivates the use of algorithms that seek to maximize sample width.

In comparison to CBR, there are two important points to emphasize here: first, CBR is fundamentally based on "lazy learning" [12] where the inference, or generalization, is done at the problem-solving time. Our learning approach for CBI is left open as we do not offer any particular algorithm to obtain the hypothesis that is used for inferring the credible set given an input problem instance. In particular, this hypothesis can be based on lazy learning, for example the nearest-neighbor rule (or any of its generalizations), or on any other supervised non-lazy learning algorithm. We emphasize again that the learning bounds that we obtain in this paper (Theorems 6.1 and 6.2) apply to any hypotheses regardless of the type of learning algorithm that produces it. As far as these bounds are concerned, it is only the performance of h on the sample that counts and not on the algorithmic procedure that obtained it.

The second point is the similarity that exists between our learning approach and the non-parametric similarity based approach of CBR: we show that there is an important notion of *width* of a hypothesis that can be used in a learning algorithm to select simpler and more accurate hypotheses automatically and based only on the sample. The width is actually defined based on discriminant functions (Section 4) that involve the computation of distance functions (which are the opposite analog of similarity functions). They measure the distance between the input problem instance and some other sets S in the metric space which in general can be any subsets of the space, and in particular can be the elements of the case-base. The fact that just calculation of distances is sufficient to give better hypotheses (and hence improved prediction of credible solution sets) makes our learning potentially of interest to the CBR research community.

We do not in the paper define one algorithmic scheme, but rather a general approach based on the modelling of CBI as the learning of two related multi-category classification problems (through the construction of two auxiliary samples). These problems can be solved by any supervised learning algorithm: we do not stipulate any particular learning algorithms, but provide generalization error bounds that apply to any algorithm.

We assume each space has a metric $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ associated with it (which, therefore, in particular, satisfies the triangle inequality). We also assume that each of the two metric spaces has a finite diameter $\mathrm{diam}(\mathcal{X}) := \max_{x,x' \in \mathcal{X}} d_{\mathcal{X}}(x, x') < \infty$, $\mathrm{diam}(\mathcal{Y}) = \max_{y,y' \in \mathcal{Y}} d_{\mathcal{Y}}(y, y') < \infty$. Note, however, that the metric spaces need not be finite and could consist of elements which are complex or highly-structured, as is typical in CBR. Our idea is based on learning 'hypotheses' (multi-category classifiers) on each of the metric spaces $\mathcal{X}$ and $\mathcal{Y}$ separately, and by taking account of the sample-width, an idea we introduced in [13, 7] and applied in [14, 15]. This leads to favoring more regular (or 'smooth') hypotheses, as much as the complexity of the sample permits. The fact that the learning approach favors such simpler hypotheses is entirely compatible with the underlying assumption of CBR that similarity in problem space implies similarity in solution space.

3 A probabilistic set-up for CBI

In this section we extend the CBI model of [4] described in Section 2. The underlying assumption that similar problems must have similar solutions is represented here through an automatic preference for, albeit not limited to, smooth hypotheses that map similar problems to similar solutions (discussed in Section 4). Automatic means that the complexity of the learnt hypothesis is dictated by the given case-base rather than by some heuristic choice. Compared to the learning approach in [4], and in addition to the advantage of not assuming any particular class of mappings (see the previous section), our approach also delivers rigorous error-bounds that depend on the particular sample (the case-base) and are therefore more useful in practice. Such bounds are referred to in the literature as sample-dependent error bounds.

We now describe our learning approach.

3.1 The probabilistic framework

In this framework, examples of problem-solutions pairs (all being positive examples, meaning that each pair consists of a problem and a solution to it) are drawn according to an unknown probability distribution $Q(Z) := Q(X, Y)$. We assume Q is multi-modal; that is, it takes the form of a weighted sum with a finite number of terms as follows:

$$\begin{aligned} Q(Z) &= \sum_{k\in[K]} Q_{Z|M}(Z\,|k)Q_M(k) \qquad (3)\\ &= \sum_{k\in[K]} Q_{Y|M}(Y\,|k)Q_{X|Y,M}(X|Y,k)Q_M(k), \end{aligned}$$

where M is a random variable representing the mode whose possible values are in a set $[K] := \{1, 2, \ldots, K\}$. The mode-conditional distribution $Q_{Z|M}(Z|k)$ is defined on $\mathcal{Z}$, and $Q_{Y|M}(Y|k) := \sum_{x\in\mathcal{X}} Q_{Z|M}((x, Y)|k)$ is a mode-conditional distribution defined on $\mathcal{Y}$ with $Q_{X|Y,M}$ a conditional distribution on $\mathcal{X}$. We henceforth refer to the support of the mode-conditional distribution $Q_{Y|M}$ in $\mathcal{Y}$ as a *mode-region*. (The support of a probability distribution is the smallest closed set whose complement has measure zero.)

For any probability distribution P on $\mathcal{Y}$, denote by $\text{supp}(P) \subseteq \mathcal{Y}$ the support of P. We assume that there exists a $\tau > 0$ such that Q belongs to a family $\mathcal{Q}_\tau$ of probability distributions that satisfy the following properties on $\mathcal{Y}$:

(A) For $k \neq k'$, we have $\text{supp}\left(Q_{Y|M}\left(Y|k\right)\right) \bigcap \text{supp}\left(Q_{Y|M}\left(Y|k'\right)\right) = \emptyset$.

(B) For any y, y' in the support of the marginal distribution of Q on $\mathcal{Y}$ such that $d_{\mathcal{Y}}(y, y') \leq \tau$, there exists $k \in [K]$ such that $y, y' \in \text{supp}\left(Q_{Y|M}(Y|k)\right)$.

(C) For any $\alpha \in (0, 1)$, there is $m_0^Q(\alpha)$ such that if a sequence of $m \geq m_0^Q(\alpha)$ elements of $\mathcal{Z}$, $\xi^{(m)} = \{(x_i, y_i)\}_{i=1}^m$, is drawn according to the product probability measure Q^m, then, with probability at least $1 - \alpha$, the following holds: for any y_{i_1}, y_{i_2} in the sample which belong to the same mode-region, there is a sequence $y_{j_1}, y_{j_2},, \ldots, y_{j_N}$ in the sample and in that same mode-region such that $d_{\mathcal{Y}}(y_{i_1}, y_{j_1}) \leq \tau$, $d_{\mathcal{Y}}(y_{j_l}, y_{j_{l+1}}) \leq \tau$ and $d_{\mathcal{Y}}(y_{j_N}, y_{i_2}) \leq \tau$, $1 \leq l \leq N - 1$.

Condition (A) says that the mode regions are disjoint (non-overlapping). Condition (B) implies that mode regions must be at least distance τ apart. Thus both conditions imply that cases drawn fall into non-overlapping 'clusters' that are at least distance τ apart in the solution space. Condition (C) implies the following deterministic condition: any pair y, y' in a mode region (not necessarily sample points) is 'τ-connected', that is, there exists a sequence of points $\{y_i\}_{i=1}^N$ in the mode-region that satisfies $d_{\mathcal{Y}}(y_i, y_{i+1}) \leq \tau$, $d_{\mathcal{Y}}(y, y_1) \leq \tau$ and $d_{\mathcal{Y}}(y', y_N) \leq \tau$. Condition (C) essentially says that the mode conditional distribution of points is 'smooth', to the extent that for any pair of random points, no matter how far apart they are in a mode region, there is a high enough probability density to ensure that with high probability there will be points drawn in between them that are not too far apart.

The above conditions imply that, for a given x, if a solution y to x is in a mode region k, then it is acceptable to predict the whole region k as its credible solution set. For, if this support region is small, then any solution contained in it is not too distant from y and is therefore a good candidate for a solution for x. And if the region is not small, then condition (C) captures

the notion that the mode conditional distribution of points is 'smooth' (not discontinuous), and the mode does not contain outlier solution instances, and therefore may likely serve as a credible set. This is a natural constraint on a mode-conditional distribution since, without it, a mode-region could further be split into multiple (smaller) modes in which case the true number of modes K would be higher and Q would be different. (The intuition is that this indicates K as being as small as possible.) Thus a mode region may serve as a credible set of candidate solutions which can be further processed by the third and fourth stages of the R^4 model to produce a solution for x. (We note again that in this paper we only deal with the CBI part which concerns the first two steps of the R^4 model.) Thus, in this set-up, we infer a whole mode region k as the inferred credible set for any problem x.

Learning CBI amounts to learning to map an x to a mode that, with high confidence (in a sense to be defined shortly), contains a solution y, and then predict the corresponding mode region as a credible set for x. We assume that τ is known to the learner but that the number K of modes of Q is unknown.

Relating to Condition (C), it is intuitively plausible that for m larger than some finite threshold value, the condition will hold. Related ideas have been studied in the context of percolation theory (see [16], for instance). In particular, the following related problem has been studied. Given a parameter τ, and a random sample from a given distribution, if the graph G_τ has as vertices the points of the sample and two vertices are connected if their distance is at most τ, is there a high probability that G_τ is connected? This has been studied in particular when the distribution is uniform on the d-dimensional unit cube.

Before continuing to describe our model, let us define two probability functions that we refer to in subsequent sections,

$$\begin{aligned} P_{\mathcal{X}}(X = x, M = k) : &= \sum_{y \in \mathcal{Y}} Q_{Z|M}((x, y)|k) Q_M(k) \\ P_{\mathcal{Y}}(Y = y, M = k) &:= \sum_{x \in \mathcal{X}} Q_{Z|M}((x, y)|k) Q_M(k). \end{aligned} \tag{4}$$

3.2 Inference by hypothesis

Given a randomly drawn problem $X \in \mathcal{X}$ the inference task is to produce a set of credible solutions for X. This inference can be represented by a mapping from $\mathcal{X}$ to $2^{\mathcal{Y}}$ in terms of a pair of functions $h_1 : \mathcal{X} \to [K]$ and $h_2 : \mathcal{Y} \to [K]$ which map the problem and solution spaces into a finite set $[K] := \{1, 2, \ldots, K\}$ of natural numbers. We henceforth write

$$h(z) : \quad = \quad [h_1(x), h_2(y)] \tag{5}$$

and refer to the vector-valued function $h : \mathcal{Z} \to [K]^2$ as a *hypothesis* for case-based inference. Note that $[K]$ is the same set in (3) that defines the modes of Q. We later show that h is learned based on a sample whose labels are equal to the mode-values (up to some permutation). Thus while Q, and hence $[K]$, are unknown, the above conditions on Q ensure that information about the set $[K]$ is available in the random sample, from which it is possible to learn h as a mapping into $[K]^2$.

Given a hypothesis h of this new type and a problem x, the credible solution *set* $C(x)$ predicted by h can now be defined as

$$C(x) := C_h(x) = \{y \in \mathcal{Y} : \; h_2(y) = h_1(x)\}$$

or, equivalently,

$$C_h(x) = h_2^{-1}(h_1(x)).$$

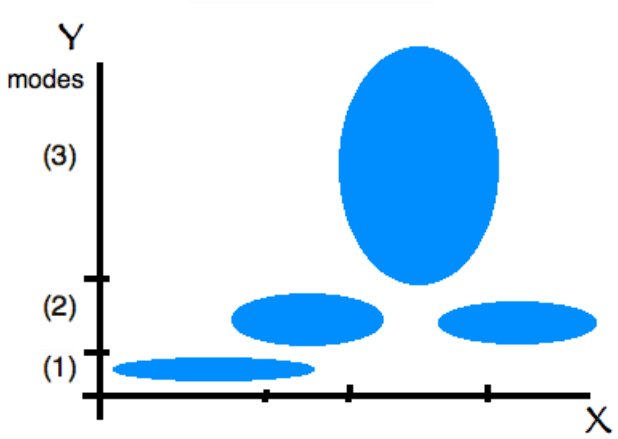
(We continue to use the notation of Hüllermeier, but have replaced his definitions by our new ones.) In other words, if $x \in \mathcal{X}$ has $h_1(x) = k$ then $C(x)$ is a set of solutions that are *classified* by h_2 as k. Thus, inference in this CBI setting amounts to classifying x into one of a finite number of solution regions.

In Section 4 we discuss how to learn h by learning the two classifiers h_1 and h_2. We learn each separately based on a labeled sample. Given a sample of cases, we prefer a simpler h that has 'smoother' component mappings h_1 and h_2. Being smooth means that the learning process prefers hypotheses h whose h_1 maps similar ($d_{\mathcal{X}}$-close) problems x, x' to the same $k \in [K]$. For similar problems, h predicts the same credible set. Thus the CBR assumption that similar problems map to similar credible solutions holds in our model.

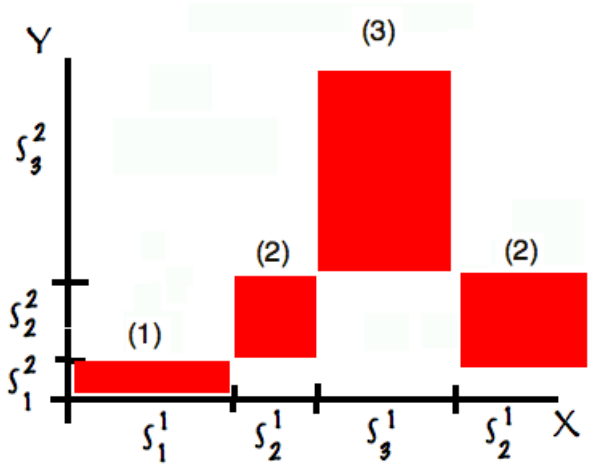
In Section 5 we show that training samples for each of h_1 and h_2 can be constructed in such a way that the labels are the values of the corresponding modes of Q. So learning h amounts to learning the mode-regions and, thus, given a problem x the learnt hypothesis h predicts the mode region (that contains a solution y of x) to be the credible solution set for x. The intuition is that if h is sufficiently 'accurate' then, with a large confidence, the predicted credible set consists of solutions y to x. More importantly, as explained above, the conditions on Q ensure that the mode region (which is the predicted credible set) has other solutions that are close to y or, at least, typical elements of the region that contains y.

Learning mode-regions is reminiscent of identifying clusters in unsupervised learning and clustering research; for instance, unsupervised learning of mixture distributions or non-parametric statistical density estimation where subgroups of the data are identified based on the modes of the estimated density [17]. As in our framework, these areas of research also assume the existence of some unknown underlying multi-modal probability distribution.

Figure 1 shows an example of a distribution Q and hypothesis h. For illustrative purposes, we have assumed that the metric spaces $\mathcal{X}$ and $\mathcal{Y}$ are one-dimensional. There are three modes $Q_{Y|M}(Y|k)$, $k = 1, 2, 3$ with non-overlapping supports in $\mathcal{Y}$ (obeying condition (A)). Associated with them are mode-conditional distributions $Q_{Z|M}(Z|k)$, $k = 1, 2, 3$, where the support of $Q_{Z|M}(Z|2)$ splits into two regions in $\mathcal{Z}$. In this example, when Q is projected on $\mathcal{X}$ there is overlap between the modes (which is permitted by the above conditions). This means that a problem may have multiple solutions, even in different mode regions. The component hypotheses h_1 and h_2 partition $\mathcal{X}$ and $\mathcal{Y}$, respectively, into regions that are labeled with values in the set $[K] = [3] = \{1, 2, 3\}$. We denote these regions by $S_k^{(1)}$ and $S_k^{(2)}$, $1 \leq k \leq 3$. Given an x, if $x \in S_k^{(1)}$ then h predicts a credible solution set $C_h(x) = S_k^{(2)}$. Note that it is possible that dissimilar problems have similar solutions. For instance, consider two different problems x in the left region of $S_2^{(1)}$ and x' in the right region of $S_2^{(1)}$. Both have similar solutions y, $y' \in S_2^{(2)}$. Our learning approach is applicable in general to probability distributions Q with mode regions that are not necessarily circular as in this example and the mapping h from $\mathcal{X}$ to sets of $\mathcal{Y}$ can be arbitrarily complex rather than box-shaped as in this example.



(a) Example of a distribution Q on $\mathcal{X} \times \mathcal{Y}$. It has K modes on $\mathcal{Y}$, $Q_{Y|M}(Y|k)$, $k = 1, \ldots, K = 3$.



(b) a hypothesis $h : \mathcal{Z} \to [K]^2$, with classification regions $S_k^{(1)}$, in $\mathcal{X}$, and $S_k^{(2)}$ in $\mathcal{Y}$, $k = 1, \ldots, K$, with , $K = 3$.

Figure 1: (a) Circular regions are mode regions of Q. Regions of different mode value may overlap with respect to $\mathcal{X}$ but not on $\mathcal{Y}$. (b) Rectangular regions are sets of problems and their credible solutions that are inferred by h. There are three such sets: the k^{th} set is labeled (k) and is defined as $S_k^{(1)} \times S_k^{(2)} = \{(x, y) : h_1(x) = h_2(y) = k\}$, $k = 1, \ldots, K$, with $K = 3$.

In the learning model that we introduce in Section 5 the number of modes K is not assumed to be known. The value of K is estimated based on a training sample of problem-solution pairs and on knowing the value of τ (which is given as domain knowledge). The estimate of K may be as large as the sample size m.

3.3 Error of h

We define the error of a hypothesis h as the probability that for a randomly drawn problem-solution pair $Z = (X, Y) \in \mathcal{Z}$, h mis-predicts Z, that is, h predicts a bad credible solution set $C_h(X)$ for X. This means that $Y \notin C_h(X)$. We therefore denote the error of h as

$$\text{err}\,(h) := Q\,(Y \notin C_h(X))\,. \tag{6}$$

Since the two components of h are classifiers, then the event that h mis-predicts (X, Y) implies that the two component classifiers disagree on the category of the solution. We can represent this as follows: denote by $M \in [K]$ the mode from which the random (X, Y) is drawn. Then the probability of mis-predicting is

$$\begin{aligned} Q\,(\{(X,Y) : Y \notin C_h(X)\}) &= Q\,(\{(X,Y) : h_1(X) \neq h_2(Y)\}) \\ &= \sum_{k\in[K]} Q_{Z|M}\,(\{(X,Y) : h_1(X) \neq h_2(Y)\}\,|k)\,Q_M(k) \\ &\leq \sum_{k\in[K]} Q_{Z|M}\,(\{(X,Y) : h_1(X) \neq k \text{ or } h_2(Y) \neq k\}\,|k)\,Q_M\,(k) \end{aligned}$$

which is bounded from above by

$$\begin{aligned}
&\sum_k Q_{Z|M}\left(\{(X,Y): h_1(X)\neq k\}\,|k\right) Q_M(k) + \sum_k Q_{Z|M}\left(\{(X,Y): h_2(Y)\neq k\}\,|k\right) Q_M(k) \\
&= \sum_k \sum_{y\in\mathcal{Y}} Q_{Z|M}\left(\{(X,y): h_1(X)\neq k\}\,|k\right) Q_M(k) \qquad (7)\\
&\qquad + \sum_k \sum_{x\in\mathcal{X}} Q_{Z|M}\left(\{(x,Y): h_2(Y)\neq k\}\,|k\right) Q_M(k) \\
&= \sum_k \sum_{x:h_1(x)\neq k} \sum_{y\in\mathcal{Y}} Q_{Z|M}\left((x,y)\,|k\right) Q_M(k) + \sum_k \sum_{y:h_2(y)\neq k} \sum_{x\in\mathcal{X}} Q_{Z|M}\left((x,y)\,|k\right) Q_M(k) \\
&= \sum_k \sum_{x:h_1(x)\neq k} P_{\mathcal{X}}(X=x, M=k) + \sum_k \sum_{y:h_2(y)\neq k} P_{\mathcal{Y}}\left(Y=y, M=k\right) \\
&= P_{\mathcal{X}}\left(h_1(X)\neq M\right) + P_{\mathcal{Y}}\left(h_2(Y)\neq M\right). \qquad (8)
\end{aligned}$$

The first and second term in (8) are the probability of misclassifying a labeled example $(X, M) \in \mathcal{X} \times [K]$ and the probability of misclassifying a labeled example $(Y, M) \in \mathcal{Y} \times [K]$ by the classifier h_1 and h_2, respectively. We denote these misclassification probabilities by

$$\mathrm{err}(h_1) := P_{\mathcal{X}}\left(h_1(X)\neq M\right)$$

and

$$\mathrm{err}(h_2) := P_{\mathcal{Y}}\left(h_2(Y)\neq M\right)$$

and therefore have

$$\mathrm{err}(h) \leq \mathrm{err}(h_1) + \mathrm{err}(h_2). \qquad (9)$$

In splitting the error of h into a sum of two errors we assumed that the mode set $[K]$ is fixed and is known to the learner. The errors (9) are implicitly dependent on the set $[K]$. In Section 5, we loosen this assumption and treat K as an unknown so that when a case Z is drawn randomly according to $Q(Z)$ the mode value k is not disclosed to the learner as part of the information in the sample. It is therefore necessary to produce auxiliary labeled samples that contain this mode information. We do that in Section 5.1.

We now proceed to present new results on learning multi-category classification on metric spaces which we subsequently use for the analysis of CBI learning in Section 6.

4 Multi-category classification on a metric space

In this section we consider classification learning on a metric space. Our aim here is to provide a bound on the error of each of the individual component hypotheses of Section 3; that is, on each of the two terms on the right side of (9). At this point, we consider a general metric space $\mathcal{X}$. (We will then apply the results to the case in which that metric space is $\mathcal{X}$ or $\mathcal{Y}$ in the CBI framework.)

For a given $x \in \mathcal{X}$, by a K-category classifier h we mean a function $h : \mathcal{X} \to [K] = \{1, \ldots, K\}$: every element $x \in \mathcal{X}$ has one definite classification according to h. (Note: here, h is not the vector-valued hypothesis defined in Section 3.)

We can associate with h the regions $S_k^{(h)} := \{x :\in \mathcal{X} : h(x) = k\}$, $k \in [K]$, where we drop the superscript and write S_k when it is clear that h is the classifier. Note that these regions are disjoint, $S_k \bigcap S_{k'} = \emptyset$ for $k \neq k'$ and their union equals $\mathcal{X}$. We define the distance between a point x and a set $S \subseteq \mathcal{X}$ based on the metric $d_{\mathcal{X}}$ as follows,

$$\operatorname{dist}(x, S) := \inf_{x' \in S} d_{\mathcal{X}}(x, x').$$

As in [7] we define the notion of *width* of a classifier h at a point x as follows:

$$w_h(x) := \min_{k \neq h(x)} \operatorname{dist}(x, S_k).$$

The width $w_h(x)$ measures how 'definite' the classification of x is according to h since the further x is from the 'border' (the set of closest points to x that are not in $S_{h(x)}$), the higher the width and the more definite the classification. Note that the width $w_h(x)$ is always non-negative. For a labeled point (x, l), $l \in [K]$, we define a real-valued *discriminant function* [17] which we denote by $f_h : \mathcal{X} \times [K] \to \mathbb{R}$ and which is defined as follows:

$$f_h(x, l) := \min_{k \neq l} \operatorname{dist}(x, S_k) - \operatorname{dist}(x, S_l).$$

Note that if $x \in S_l$ then by definition $x \notin S_k$ for every $k \neq l$ and so we have

$$f_h(x, l) = w_h(x).$$

If $x \notin S_l$ then it must be that $x \in S_k$ for some $k \neq l$ and hence

$$f_h(x, l) = -\text{dist}\,(x, S_l)\,.$$

For a fixed h and $k \in [K]$ define the real-valued function $g_k^{(h)} : \mathcal{X} \to \mathbb{R}$ as

$$g_k^{(h)}(x) = f_h(x, k)$$

where we will drop the superscript for brevity and write g_k whenever the dependence on h can be left implicit. We denote by $g^{(h)}$ the vector-valued function $g^{(h)} : \mathcal{X} \to \mathbb{R}^K$ given by

$$g^{(h)}(x) := [g_1^{(h)}(x), \ldots, g_K^{(h)}(x)].$$

We refer to $g^{(h)}$ as the *margin function* of the classifier h. Note that for a fixed h and $x \in \mathcal{X}$ there is only a single component $g_k^{(h)}$ of $g^{(h)}$ which is non-negative, and its value equals the width $w_h(x)$, while the remaining components are all negative.

Thus we can express the decision of the classifier h in terms of g as follows:

$$h(x) = \text{argmax}_{k \in [K]} g_k(x).$$

It is important to note at this point that a hypothesis h is completely specified in terms of distances between the given input problem instance x and the subsets S_k in the metric space. There is no parameters hence the class of hypotheses that our learning framework uses is non-parametric. It is much richer than the class of nearest-neighbor classifiers since the regions $S_k^{(h)}$ can be any subsets of the metric space.

The event of misclassification of a labeled point (x, l) by h means that there exists some component g_k with $k \neq l$ such that $g_l(x) < g_k(x)$. So the event that h misclassifies a labeled point (x, l) can be expressed as the event that $g_l(x) < \max_{k \neq l} g_k(x)$. Thus for a randomly drawn pair $(X, L) \in \mathcal{X} \times [K]$, we have

$$P\,(h(X) \neq L) = P(g_L(X) < \max_{k \neq L} g_k(X))$$

where $g = g^{(h)}$ is the margin function corresponding to h. We henceforth denote this by the *error* $\text{err}(h)$ of h,

$$\text{err}(h) := P\,(h(X) \neq L)\,.$$

The *empirical error* of h is the average number of misclassifications that h makes on a labeled sample $\chi^{(m)} = \{(x_i, l_i)\}_{i=1}^m$. A more stringent measure is the average number of examples which h does not classify to within some pre-specified minimal width level $\gamma > 0$; that is, the average number of examples (x_j, l_j) for which $g_{l_i}(x_i) - \max_{k \neq l_i} g_k(x_i) < \gamma$. We call this the *empirical margin* error of h (at scale γ) and denote it as

$$\hat{\text{err}}_\gamma(h) := \frac{1}{m} \sum_{i=1}^m \mathbb{I}\left\{ g_{l_i}(x_i) - \max_{k \neq l_i} g_k(x_i) < \gamma \right\}.$$

(Here, $\mathbb{I}$ denotes the indicator function of an event.)

In [18], the general problem of learning multi-category classifiers defined on metric spaces is investigated, and a generalization error bound is presented. In order to describe this, we first need to define what we mean by covering numbers of a metric space.

Suppose, as above, that $(\mathcal{X}, d_{\mathcal{X}})$ is any metric space and that $\alpha > 0$. Then an α-cover of $\mathcal{X}$ (with respect to $d_{\mathcal{X}}$) is a finite subset C of $\mathcal{X}$ such that, for every $x \in \mathcal{X}$, there is some $c \in C$ such that $d_{\mathcal{X}}(x, c) \leq \alpha$. If such a cover exists, then the minimum cardinality of such a cover is the *covering number* $\mathcal{N}(\mathcal{X}, \alpha, d_{\mathcal{X}})$. If the context is clear, we will abbreviate this to $\mathcal{N}_\alpha$.

We will see that the covering numbers (for both $\mathcal{X}$ and $\mathcal{Y}$) play a role in our analysis. So, in practice, it would be useful to know these or to be able to estimate them.

For the moment, let us focus on the case in which we have a finite metric space $\mathcal{X}$ of cardinality N. Then, the problem of finding a minimum γ-cover C_γ for $\mathcal{X}$ can be phrased as a classical *set-cover problem* as follows: find a minimal cardinality collection of sets $C_\gamma := \{B_\gamma(j_l) : j_l \in \mathcal{X}, 1 \leq l \leq \mathcal{N}_\gamma\}$ whose union satisfies $\bigcup_l B_\gamma(j_l) = \mathcal{X}$. It is well known that this problem is NP-complete. However, there is a simple efficient deterministic greedy algorithm (see [19]) which yields a solution — that is, a set cover — of size which is no larger than $(1 + \ln N)$ times the size of the minimal cover. Denote by $\hat{C}_\gamma$ this almost-minimal γ-cover of $\mathcal{X}$ and denote by $\hat{N}_\gamma$ its cardinality. Then $\hat{N}_\gamma$ can be used to approximate N_γ up to a $(1 + \ln N)$ accuracy factor:

$$N_\gamma \leq \hat{N}_\gamma \leq N_\gamma(1 + \ln N).$$

We now present two results from [18]. The first bounds the generalization error in terms of a width parameter γ for which the corresponding empirical margin error is zero. All results henceforth apply to any metric space, including infinite spaces. In these results, it is assumed that the labeled examples (x_i, l_i) in the training sample $\chi^{(m)}$ have each been generated randomly according to some fixed (but unknown) probability distribution P on $\mathcal{Z} = \mathcal{X} \times [K]$. Thus, a sample $\chi^{(m)}$ of length m can be regarded as being drawn randomly according to the product probability distribution P^m.

Theorem 4.1 *Suppose that $\mathcal{X}$ is a metric space of diameter* $\operatorname{diam}(\mathcal{X})$ *and that K is a positive integer. Suppose P is any probability measure on $\mathcal{Z} = \mathcal{X} \times [K]$ and let P^m denote the product probability measure on $\mathcal{Z}^m$. Let $\delta \in (0, 1)$. Then, with P^m-probability at least $1 - \delta$, the following holds for $\chi^{(m)} \in Z^m$: for any function $h : X \to [K]$, and for any $\gamma \in (0, \operatorname{diam}(\mathcal{X})]$, if* $\hat{\operatorname{erf}}_\gamma(h) = 0$, *then*

$$\operatorname{err}(h) \leq \frac{2}{m}\left(K\,\mathcal{N}_{\gamma/12}\log_2\left(\frac{36\operatorname{diam}(\mathcal{X})}{\gamma}\right) + \log_2\left(\frac{8\operatorname{diam}(\mathcal{X})}{\delta\gamma}\right)\right),$$

where $\mathcal{N}_{\gamma/12}$ is the $\gamma/12$-covering number of $\mathcal{X}$.

What this bound shows is that a hypothesis h that has a large width value γ on every point of the sample is likely to obtain a low generalization error. An important point of our work in this paper is the fact that we make this large-width advantage appear in learning CBI; that is, the same conclusion holds in our CBI learning bounds, Theorem 6.1 and 6.2.

Note here that γ is not prescribed in advance, but can be chosen after learning and, in particular, it can be set to be the largest value for which the corresponding empirical margin error is zero.

The following result is more general than the one just presented, because it bounds the error in terms of the empirical margin error (which may be nonzero). It has a better dependence on K (being proportional to $\sqrt{K}$ rather than K). However, in terms of m, it is looser when applied to the case of zero empirical margin error (involving $1/\sqrt{m}$ rather than $1/m$.

Theorem 4.2 *With the notation as above, with P^m-probability at least $1-\delta$, the following holds for $\chi^{(m)} \in Z^m$: for any function $h : X \to [K]$, and for any $\gamma \in (0, \mathrm{diam}(X)]$,*

$$\mathrm{err}(h) \leq \hat{\mathrm{err}}_\gamma(h) + \sqrt{\frac{2}{m}\left(K\,\mathcal{N}_{\gamma/12} \ln\left(\frac{18\,\mathrm{diam}(\mathcal{X})}{\gamma}\right) + \ln\left(\frac{2\,\mathrm{diam}(\mathcal{X})}{\gamma\delta}\right)\right)} + \frac{1}{m},$$

where $\mathcal{N}_{\gamma/12}$ is the $\gamma/12$-covering number of $\mathcal{X}$.

What we have in Theorem 4.2 is a high probability bound that takes the following form: for all h and for all $\gamma \in (0, \mathrm{diam}(X)]$,

$$\mathrm{err}(h) \leq \hat{\mathrm{err}}_\gamma(h) + \epsilon(m, \gamma, \delta),$$

where ϵ tends to 0 as $m \to \infty$ and ϵ decreases as γ increases. The rationale for seeking such a bound is that there is likely to be a trade-off between empirical margin error on the sample and the value of ϵ: taking γ small so that the error term $\hat{\mathrm{err}}_\gamma(h)$ is zero might entail a large value of ϵ; and, conversely, choosing γ large will make ϵ relatively small, but lead to a large empirical error term. So, in principle, since the value γ is free to be chosen, one could optimize the choice of γ on the right-hand side of the bound to minimize it.

5 From CBI to supervised learning

In Section 3, we have framed CBI as a multi-category classification learning problem in which hypotheses are two-dimensional multi-category functions $h = [h_1, h_2]$. In the current fairly technical section, we describe how from the case base we can derive the training samples that are necessary for any supervised learning algorithm to learn h_1 and h_2. These 'auxiliary' samples are defined from the cases by a labeling procedure that we describe. One of these auxiliary samples consists of labeled problems and the other consists of the corresponding labeled solutions. Each of the two components of h are learned separately based on these auxiliary samples. Section 6 describes the learning results that follow.

5.1 Two auxiliary samples

The learner is given a random sample, which is also referred to as a collection of problem-solution cases (or case base),

$$\xi : \;=\; \xi^{(m)} = \{(x_i, y_i)\}_{i=1}^{m} . \tag{10}$$

This sample is drawn i.i.d. according to some product probability measure Q^m on $\mathcal{Z}^m$, where $Q \in \mathcal{Q}_\tau$ for some $\tau > 0$.

Denote by

$$\mathcal{X}_{|\xi} := \{x_i \in \mathcal{X} : \exists i \in \{1, \ldots, m\}, (x_i, y_i) \in \xi\}$$

and

$$\mathcal{Y}_{|\xi} := \{y_i \in \mathcal{Y} : \exists i \in \{1, \ldots, m\}, (x_i, y_i) \in \xi\}$$

the sample projection sets of problems and solutions, respectively. Note that the sample ξ may be 'noisy'; that is, a sample problem $x \in \mathcal{X}_{|\xi}$ may appear multiple times in the sample with different solutions $y \in \mathcal{Y}_{|\xi}$ and even solutions from different modes. In other words, the modes of Q may overlap in problem space $\mathcal{X}$, and hence cases drawn according to Q may pair the same problems with different solutions. Needless to say, a solution $y \in \mathcal{Y}_{|\xi}$ may appear multiple times for different problems $x \in \mathcal{X}_{|\xi}$.

In addition to the sample ξ we assume that expert advice (or domain-knowledge) is available in the form of knowing the value of τ, the parameter of the family $\mathcal{Q}_\tau$ satisfying the properties in Section 3.1.

We now describe a procedure the learner can use to construct two auxiliary labeled samples $\zeta_\mathcal{X}$ and $\zeta_\mathcal{Y}$ from the given sample ξ and the value τ.

Labeling Procedure: We use τ to partition the sample points of ξ into a finite number of categories as follows. Let D_ξ be the $m \times m$ matrix with entries as follows:

$$D_\xi[i, j] \;=\; d_\mathcal{Y}(y_i, y_j)$$

for all pairs of solution examples y_i, $y_j \in \mathcal{Y}_{|\xi}$. Based on D_ξ, let us define the $m \times m$ $\{0, 1\}$ matrix

$$A_\tau : \;=\; [a(i, j)] \tag{11}$$

as follows:

$$a(i,j) := \begin{cases} 1 & \text{if} \quad D_\xi[i,j] \leq \tau \\ 0 & \text{otherwise.} \end{cases}$$

The j^{th} column $a^{(j)}$ of A_τ represents an incidence (binary) vector of a set, or a ball $B_\tau(j)$ which consists of all the points $i \in \mathcal{Y}_{|\xi}$ that are a distance at most τ from the point $j \in \mathcal{Y}_{|\xi}$.

The matrix A_τ defined in (11) is an adjacency matrix of a graph $G_\tau = (\mathcal{Y}_{|\xi}, E_\tau)$, where E_τ is the set of edges corresponding to all adjacent pairs of vertices according to A_τ; that is, we place an edge between any two vertices i, j such that $D_\xi[i,j] \leq \tau$.

Let $\{H_i\}_{i=1}^{K_\tau}$ be the set of K_τ connected components $H_i \subseteq \mathcal{Y}_{|\xi}$ of the graph G_τ, where by a *connected component* we mean a subset of vertices such that there exists a path (sequence of edges) between every pair of vertices in the component. This set of components can be easily found, for instance, by a hierarchical clustering procedure [20].

Note that $K_\tau := K_\tau(\xi)$ is dependent on the sample ξ through $\mathcal{Y}_{|\xi}$ and is no larger than m since the number of connected components is no larger than the number of vertices of G_τ. Let us partition the sample ξ into the subsets $\xi^{(k)} \subseteq \xi$ based on these components H_k as follows:

$$\xi^{(k)} := \{(x,y) \in \xi : \, y \in H_k\}, \quad 1 \leq k \leq K_\tau.$$

Then, define two auxiliary sets of samples as follows:

$$\begin{aligned} \zeta_{\mathcal{X}} &:= \zeta_{\mathcal{X}}^{(m)} = \left\{(x_i,k) : \, x_i \in \mathcal{X}_{|\xi}, (x_i,\cdot) \in \xi^{(k)}, 1 \leq i \leq m, 1 \leq k \leq K_\tau\right\} \\ \zeta_{\mathcal{Y}} &:= \zeta_{\mathcal{Y}}^{(m)} = \left\{(y_i,k) : \, y_i \in \mathcal{Y}_{|\xi}, (\cdot,y_i) \in \xi^{(k)}, 1 \leq i \leq m, 1 \leq k \leq K_\tau\right\} \end{aligned} \quad (12)$$

We use these samples for the classification learning problems in Section 5.2. Note that both samples have K_τ possible categories for the labels of each of the sample points. Since K_τ enters the learning bounds it is important to understand how large it can be. From spectral graph theory [21, 22] the number of connected components of a graph G is equal to the multiplicity $\mu_0(G)$ of the zero eigenvalue of the Laplacian matrix $\mathcal{L} := \Lambda - A$, where Λ

is a diagonal matrix of the degrees of each vertex and A is the adjacency matrix. It follows that

$$K_\tau = \mu_0(G_\tau)$$

and clearly $K_\tau \leq m$.

We now state two lemmas that together imply that the labels l_i of pairs of examples (x_i, l_i) and (y_i, l_i) in $\zeta_\mathcal{X}$ and $\zeta_\mathcal{Y}$ equal the true unknown mode values of the unknown underlying distribution $Q(Z)$, up to a permutation. That is, under a permutation π of the set $[K]$ a label value $j \in [K]$ is in one-to-one correspondence with a mode value $\pi(j) \in [K]$.

Lemma 5.1 *Let H be a connected component of G_τ. Then there exists a $k \in [K]$ such that $H \subseteq supp\left(Q_{Y|M}(Y|k)\right)$.*

Proof: Denote by $R_k = \text{supp}(Q_{Y|M}(y|k))$, $k \in [K]$, the mode regions. Suppose there does not exist a j such that $H \subseteq R_j$. Then there is a connected pair $y, y' \in H$ such that $y \in R_k$ and $y' \in R_{k'}$ for some $k' \neq k$. This means that on any path that connects y and y' there exists some edge $e \in E_\tau$ that connects two vertices u, $v \in \mathcal{Y}_{|\xi}$ (which may be y or y') where $u \in R_k$ and $v \in R_{k'}$. But by condition (B) of Section 3 it follows that $d_\mathcal{Y}(u, v) > \tau$ hence by definition of G_τ the pair u, v is not connected. Hence y, y' are disconnected. This is a contradiction and hence the statement of the lemma holds. □

Lemma 5.2 *Let $\alpha \in (0, 1)$ and suppose that the sample size m is at least $m_0^Q(\alpha)$. Let $\{H_j\}_{j=1}^{K_\tau}$ be the connected components of the graph G_τ. Then, with probability at least $1 - \alpha$, the sample is such that, for every $k \in [K]$, there exist at most one single component $H_j \subseteq supp\left(Q_{Y|M}(Y|k)\right)$.*

Proof: Suppose there are two distinct connected components H, H' of the graph contained in a mode-region $R_k = \text{supp}(Q_{Y|M}(y|k))$ for some $k \in [K]$.

Then there exist two points $y \in H$, $y' \in H'$ such that every path $p = \{y, y_1, \ldots, y_n, y'\}$ from y to y' must have at least one pair of consecutive points y_i, y_{i+1} such that $d_{\mathcal{Y}}(y_i, y_{i+1}) > \tau$. But, by condition (C) of Section 3, if $m \geq m_0^Q(\alpha)$, with probability at least $1 - \alpha$, this cannot be. Hence the statement of the lemma holds. □

From these two lemmas, the following observation follows.

Proposition 5.3 *For any $\alpha \in (0, 1)$, with probability at least $1-\alpha$, provided m is large enough ($m \geq m_0^Q(\alpha)$), a connected component H_k of the graph G_τ is always contained in the support of a mode-conditional distribution $Q_{Y|M}$ and there is never more than a single such component in a mode-region.*

This implies that if an example $(x_i, l_i) \in \zeta_{\mathcal{X}}$ corresponds to a case $(x_i, y_i) \in \xi$ with y_i in a connected component H_k of the graph G_τ then l_i equals k where k equals (up to a permutation) the value of the true (unknown) mode from which the case was drawn from. Similarly, if an example $(y_i, l_i) \in \zeta_{\mathcal{Y}}$ is such that y_i falls in a connected component H_k of the graph G_τ then l_i equals k where k equals (up to a permutation) the value of the mode from which the case was drawn from.

Thus the labels l_i of the sample points of $\zeta_{\mathcal{X}}$ and $\zeta_{\mathcal{Y}}$ are representative of the mode numbers and thus these auxiliary samples are proper labeled samples needed for supervised learning of the classifiers h_1 and h_2, respectively.

5.2 Two classification problems

Given the two auxiliary samples $\zeta_{\mathcal{X}}$ and $\zeta_{\mathcal{Y}}$ of (12), we learn two multi-category classification problems, independently, by finding a component hypothesis h_1 and h_2. *Any* supervised learning algorithms could be used to produce h_1 and h_2: we do not propose or limit discussion to any particular ones, but derive performance results (Theorems 6.1 and 6.2) that apply to

all algorithms. Based on h_1 and h_2 we form a hypothesis $h = [h_1, h_2]$ as in (5), where by (9) its error is bounded by the sum of the errors of h_1 and h_2.

As mentioned above, the number of categories $K_\tau(\xi)$ is dependent on the sample ξ, or more specifically on the set $\mathcal{Y}_{|\xi}$. Thus we need to make the bounds of Section 4 apply for any value K and not just for a K which is fixed in advance. To do that we use a 'sieve' method in the error-bound proof.

To be able to use the standard-learning theory bounds we need the auxiliary samples $\zeta_\mathcal{X}$ and $\zeta_\mathcal{Y}$ to be drawn i.i.d.. The next lemmas state that they are effectively drawn in an i.i.d. manner.

Lemma 5.4 *Let $\alpha \in (0,1)$ and $m \geq m_0^Q(\alpha)$. Let ξ be a random sample consisting of i.i.d. pairs of problem-solution cases. Let $\zeta_\mathcal{Y}$ be a sample obtained by the labeling procedure applied on ξ. Then, with probability at least $1-\alpha$, $\zeta_\mathcal{Y}$ consists of m i.i.d. random pairs of solution-mode values each drawn according to $P_\mathcal{Y}$.*

Proof: Let $L^{(m)} = [L_1, \ldots, L_m]$ denote the label vector random variable and $Y^{(m)} = [Y_1, \ldots, Y_m]$ the solution vector random variable, where $\{(Y_i, L_i)\}_{i=1}^m = \zeta_\mathcal{Y}$ is a random sample produced by the labeling procedure of Section 5.1. Denote by $M^{(m)} = [M_1, \ldots, M_m] \in [K]^m$ where M_i is the mode index corresponding to the solution Y_i. For any given sample realization $\zeta_\mathcal{Y} = \{(y_i^*, l_i^*)\}_{i=1}^m$ with $y^{*(m)} = [y_1^*, \ldots, y_m^*] \in \mathcal{Y}^m$ and $l^{*(m)} = [l_1^*, \ldots, l_m^*] \in [K]^m$ we have

$$\begin{aligned} P\left(\{(Y_i, L_i)\}_{i=1}^m = \{(y_i^*, l_i^*)\}_{i=1}^m\right) &= P\left(L^{(m)} = l^{*(m)} \,\middle|\, Y^{(m)} = y^{*(m)}\right) P\left(Y^{(m)} = y^{*(m)}\right) \\ &= \sum_{l^{(m)} \in [K]^m} P\left(L^{(m)} = l^{*(m)} \,\middle|\, M^{(m)} = l^{(m)}, Y^{(m)} = y^{*(m)}\right) \\ &\quad \cdot P\left(M^{(m)} = l^{(m)} \mid Y^{(m)} = y^{*(m)}\right) P\left(Y^{(m)} = y^{*(m)}\right) \quad (13) \end{aligned}$$

Conditioned on $Y^{(m)} = y^{*(m)}$ being drawn from mode values $M^{(m)} = l^{(m)}$, from Proposition 5.3, if $m \geq m_0(\alpha)$, then with probability at least $1-\alpha$, the labels equal the mode values; that is, $L^{(m)} = l^{(m)}$. (In fact, as noted earlier,

the labels are equal to the mode values up to a permutation, by which we mean there is some fixed permutation π such that $L^{(m)} = \pi(l^{(m)})$. However, without loss of any generality, we can assume that the labels are the same as the mode values because what matters is that the labels on the two auxiliary samples match.) Hence (13) equals

$$\begin{aligned}
&\sum_{l^{(m)} \in [K]^m} \mathbb{I}\left\{l^{(m)} = l^{*(m)}\right\} P\left(M^{(m)} = l^{(m)} \mid Y^{(m)} = y^{*(m)}\right) P\left(Y^{(m)} = y^{*(m)}\right) \\
&= P\left(M^{(m)} = l^{*(m)} \middle| Y^{(m)} = y^{*(m)}\right) P\left(Y^{(m)} = y^{*(m)}\right) \\
&= P\left(Y^{(m)} = y^{*(m)}, M^{(m)} = l^{*(m)}\right) \\
&= \sum_{x^{(m)}} P\left(X^{(m)} = x^{(m)}, Y^{(m)} = y^{*(m)}, M^{(m)} = l^{*(m)}\right) \\
&= P\left(M^{(m)} = l^{*(m)}\right) \sum_{x^{(m)}} P\left(X^{(m)} = x^{(m)}, Y^{(m)} = y^{*(m)} \middle| M^{(m)} = l^{*(m)}\right) \\
&= \sum_{x^{(m)}} \prod_{i=1}^{m} Q_{Z|M}\left(X_i = x_i, Y_i = y_i^* \,|M_i = l_i^*\right) Q_M\left(M_i = l_i^*\right) \qquad (14) \\
&= \prod_{i=1}^{m} Q_M\left(M_i = l_i^*\right) \sum_{x_i \in \mathcal{X}} Q_{Z|M}\left(X_i = x_i, Y_i = y_i^* \,|M_i = l_i^*\right) \\
&= \prod_{i=1}^{m} \sum_{x_i \in \mathcal{X}} Q_{Z|M}\left(X_i = x_i, Y_i = y_i^* \,|M_i = l_i^*\right) Q_M\left(M_i = l_i^*\right) \\
&= \prod_{i=1}^{m} P_{\mathcal{Y}}\left(Y_i = y_i^*, M_i = l_i^*\right) \qquad (15)
\end{aligned}$$

where (14) follows from the fact that the sample ξ is drawn i.i.d. according to $\prod_{i=1}^m Q(Z_i) = \prod_{i=1}^m Q_{Z|M}(Z_i|M_i)Q_M(M_i)$, and (15) follows from (4). Hence it follows that the random sample $\zeta_{\mathcal{Y}}$ consists of m i.i.d. trials according to the distribution $P_{\mathcal{Y}}(Y, M)$. □

The next lemma shows that the sample $\zeta_{\mathcal{X}}$ is also i.i.d..

Lemma 5.5 *Let $\alpha \in (0,1)$ and $m \geq m_0^Q(\alpha)$. Let ξ be a random sample consisting of i.i.d. pairs of problem-solution cases. Let $\zeta_{\mathcal{X}}$ be a sample obtained by the labeling procedure applied on ξ. Then, with probability at least*

$1-\alpha$, $\zeta_{\mathcal{X}}$ *consists of* m *i.i.d. random pairs of problem-mode values each drawn according to* $P_{\mathcal{X}}$.

Proof: Let $L^{(m)} = [L_1, \ldots, L_m]$ denote the label vector random variable and $X^{(m)} = [X_1, \ldots, X_m]$ the problem vector random variable. Denote by $M^{(m)} = [M_1, \ldots, M_m] \in [K]^m$ where M_i is the mode index corresponding to the problem X_i. For any sample $\zeta_{\mathcal{X}} = \{(x_i^*, l_i^*)\}_{i=1}^m$ with $x^{*(m)} = [x_1^*, \ldots, x_m^*] \in \mathcal{X}^m$ and $l^{*(m)} = [l_1^*, \ldots, l_m^*] \in [K]^m$ we have

$$
\begin{aligned}
&P\left(\{(X_i, L_i)\}_{i=1}^m = \{(x_i^*, l_i^*)\}_{i=1}^m\right) = P\left(X^{(m)} = x^{*(m)}, L^{(m)} = l^{*(m)}\right) \\
&= \sum_{l^{(m)} \in [K]^m} P\left(X^{(m)} = x^{*(m)}, L = l^{*(m)} \,\middle|\, M^{(m)} = l^{(m)}\right) P\left(M^{(m)} = l^{(m)}\right) \\
&= \sum_{l^{(m)} \in [K]^m} \sum_{y^{(m)}} P\left(X^{(m)} = x^{*(m)}, L^{(m)} = l^{*(m)}, Y^{(m)} = y^{(m)} \,\middle|\, M^{(m)} = l^{(m)}\right) P\left(M^{(m)} = l^{(m)}\right) \\
&= \sum_{l^{(m)} \in [K]^m} \sum_{y^{(m)}} P\left(L^{(m)} = l^{*(m)} \mid X^{(m)} = x^{*(m)}, Y^{(m)} = y^{(m)}, M^{(m)} = l^{(m)}\right) \\
&\quad \cdot P\left(X^{(m)} = x^{*(m)} \,\middle|\, Y^{(m)} = y^{(m)}, M^{(m)} = l^{(m)}\right) P(Y^{(m)} = y^{(m)} \,\middle|\, M^{(m)} = l^{(m)}) P\left(M^{(m)} = l^{(m)}\right).
\end{aligned} \tag{16}
$$

Conditioned on $X^{(m)} = x^{*(m)}$ being drawn from mode values $M^{(m)} = l^{(m)}$, from Proposition 5.3, if $m \geq m_0(\alpha)$, then with probability at least $1-\alpha$, we can assume as before the labels are equal to the modes, that is, $L^{(m)} = l^{(m)}$.

Hence (16) equals

$$\begin{aligned}
&\sum_{l^{(m)}\in[K]^m}\sum_{y^{(m)}}\mathbb{I}\left\{l^{(m)}=l^{*(m)}\right\}P\left(X^{(m)}=x^{*(m)}\left|Y^{(m)}=y^{(m)},M^{(m)}=l^{(m)}\right.\right)\\
&\qquad\cdot P(Y^{(m)}=y^{(m)}\mid M^{(m)}=l^{(m)})P\left(M^{(m)}=l^{(m)}\right)\\
=\;& P\left(M^{(m)}=l^{*(m)}\right)\sum_{y^{(m)}}P\left(X^{(m)}=x^{*(m)},Y^{(m)}=y^{(m)}\left|M^{(m)}=l^{*(m)}\right.\right)\\
=\;&\sum_{y^{(m)}}\prod_{i=1}^{m}Q_{Z|M}\left(X_i=x_i^*,Y_i=y_i\,|M_i=l_i^*\right)Q_M\left(M_i=l_i^*\right) \qquad (17)\\
=\;&\prod_{i=1}^{m}Q_M\left(M_i=l_i^*\right)\sum_{y_i\in\mathcal{Y}}Q_{Z|M}\left(X_i=x_i^*,Y_i=y_i\,|M_i=l_i^*\right)\\
=\;&\prod_{i=1}^{m}\sum_{y_i\in\mathcal{Y}}Q_{Z|M}\left(X_i=x_i^*,Y_i=y_i\,|M_i=l_i^*\right)Q_M\left(M_i=l_i^*\right)\\
=\;&\prod_{i=1}^{m}P_{\mathcal{X}}\left(X_i=x_i^*,M_i=l_i^*\right) \qquad (18)
\end{aligned}$$

where (17) follows from the fact that the sample ξ is drawn i.i.d. according to $\prod_{i=1}^{m}Q(Z_i)=\prod_{i=1}^{m}Q_{Z|M}(Z_i|M_i)Q_M(M_i)$, and (18) follows from (4). Hence it follows that the random sample $\zeta_{\mathcal{X}}$ is drawn as m i.i.d. trials according to the distribution $P_{\mathcal{X}}(X,M)$. □

6 Error bounds for learning CBI

In this section, we give bounds on the credible set prediction error of any hypothesis $h=[h_1,h_2]$. In particular, if h_1,h_2 happen to have a large width value γ on every point of the auxiliary samples, then the bounds indicate that the prediction will likely be a good one. Having these bounds can serve as a guiding criterion for supervised learning algorithms to learn to predict credible sets more accurately by producing hypotheses which make these bounds small.

Recall that what we want to do is obtain a high-probability bound on the error $\mathrm{err}(h)$ of a hypothesis h, which is the probability that for a randomly drawn problem-solution pair $Z = (X, Y) \in \mathcal{Z}$, h mispredicts Z; that is, h predicts a bad credible solution set $C_h(X)$ for X. Now, by (8), this error is bounded by the sum

$$P_{\mathcal{X}}\left(h_1(X) \neq M\right) + P_{\mathcal{Y}}\left(h_2(Y) \neq M\right) = \mathrm{err}(h_1) + \mathrm{err}(h_2).$$

We may use Theorem 4.1 and Theorem 4.2 to bound each of the two probabilities here. This results in the following error bounds. To be clear, in these bounds, $\hat{\mathrm{err}}_{\gamma_1}(h_1)$ means the empirical margin error of h_1 at scale γ_1 on sample $\zeta_{\mathcal{X}}$, and $\hat{\mathrm{err}}_{\gamma_2}(h_2)$ means the empirical margin error of h_2 at scale γ_2 on sample $\zeta_{\mathcal{Y}}$. The fact that the theorems below hold for all K means that the number of modes of Q does not have to be known in order to apply the theorems.

Theorem 6.1 *With the notation as above, with probability at least $1-\delta$, the following holds for all integers $m \geq m_0^Q(\delta/2)$. For all positive integers K for all $\gamma_1 \in (0, \mathrm{diam}(\mathcal{X})]$ and $\gamma_2 \in (0, \mathrm{diam}(\mathcal{Y})]$, and for all $h = [h_1, h_2]$ mapping $\mathcal{X} \times \mathcal{Y}$ into $[K]^2$: if $\hat{\mathrm{err}}_{\gamma_1}(h_1) = 0$ and $\hat{\mathrm{err}}_{\gamma_2}(h_2) = 0$, then the error of h is at most*

$$\frac{2}{m}\left(K(A + B + 2) + C + 10\right),$$

where

$$A = \mathcal{N}\left(\mathcal{X}, \gamma/12, d_{\mathcal{X}}\right) \log_2\left(\frac{36\,\mathrm{diam}(\mathcal{X})}{\gamma_1}\right),$$

$$B = \mathcal{N}\left(\mathcal{Y}, \gamma/12, d_{\mathcal{Y}}\right) \log_2\left(\frac{36\,\mathrm{diam}(\mathcal{Y})}{\gamma_2}\right),$$

$$C = \log_2\left(\frac{\mathrm{diam}(\mathcal{X})\mathrm{diam}(\mathcal{Y})}{\delta^2\gamma_1\gamma_2}\right).$$

Proof: Fix K. We apply Theorem 4.1 simultaneously to both auxiliary samples. It is the case, since $m \geq m_0^Q(\delta/2)$, that with probability at least $1 - \delta/2$, each auxiliary sample will be i.i.d., by Lemma 5.4 and Lemma 5.5. Call this event the 'independence event'. Assuming the independence event

holds, Theorem 4.1 then shows that, with probability at least $1 - \delta/2^{K+1}$, the sample will be such that we have both

$$\mathrm{err}(h_1) \leq \frac{2}{m} AK + \frac{2}{m} \log_2 \left(\frac{32\, 2^K \mathrm{diam}(\mathcal{X})}{\gamma_1 \delta} \right)$$

and

$$\mathrm{err}(h_2) \leq \frac{2}{m} BK + \frac{2}{m} \log_2 \left(\frac{32\, 2^K \mathrm{diam}(\mathcal{Y})}{\gamma_2 \delta} \right).$$

This is for fixed K. It follows that, if the independence event holds, then with probability at least $1 - \sum_{K=1}^{\infty} \delta/2^{K+1} = 1 - \delta/2$, the error of h is at most

$$\frac{2}{m} AK + \frac{2}{m} \log_2 \left(\frac{32\, 2^K \mathrm{diam}(\mathcal{X})}{\gamma_1 \delta} \right) + \frac{2}{m} BK + \frac{2}{m} \log_2 \left(\frac{32\, 2^K \mathrm{diam}(\mathcal{Y})}{\gamma_2 \delta} \right).$$

So, the probability that either the independence event does not hold, or it does but the stated error bound fails, is at most $\delta/2 + \delta/2$. The result follows. □

Theorem 6.2 *With the notation as above, with probability at least $1 - \delta$, the following holds for all integers $m \geq m_0^Q(\delta/2)$. For all positive integers K for all $\gamma_1 \in (0, \mathrm{diam}(\mathcal{X})]$ and $\gamma_2 \in (0, \mathrm{diam}(\mathcal{Y})]$, and for all $h = [h_1, h_2]$ mapping $\mathcal{X} \times \mathcal{Y}$ into $[K]^2$:*

$$\mathrm{err}(h) \leq \hat{\mathrm{err}}_{\gamma_1}(h_1) + \hat{\mathrm{err}}_{\gamma_2}(h_2) + \frac{2}{m} + (A + B) \sqrt{\frac{2}{m}},$$

where

$$A = \sqrt{K\, \mathcal{N}(\mathcal{X}, \gamma_1/6, d_{\mathcal{X}}) \ln \left(\frac{18\, \mathrm{diam}(\mathcal{X})}{\gamma_1} \right) + \ln \left(\frac{8\, \mathrm{diam}(\mathcal{X})}{\gamma_1 \delta} \right) + K}$$

and

$$B = \sqrt{K\, \mathcal{N}(\mathcal{Y}, \gamma_2/6, d_{\mathcal{Y}}) \ln \left(\frac{18\, \mathrm{diam}(\mathcal{Y})}{\gamma_2} \right) + \ln \left(\frac{8\, \mathrm{diam}(\mathcal{Y})}{\gamma_2 \delta} \right) + K}$$

Proof: The result follows from Theorem 4.2 in a similar way as the previous theorem followed from Theorem 4.1, making the observation that

$$\ln\left(\frac{8\,2^K\,\mathrm{diam}(\mathcal{X})}{\gamma\delta}\right) \leq K + \ln\left(\frac{8\,\mathrm{diam}(\mathcal{X})}{\gamma\delta}\right).$$

□

7 Conclusions

Hüllermeier introduced a framework of CBI whose goal is to predict a credible set of solutions for a given input problem instance. In his framework, he has a specific learning algorithm that produces a hypothesis function, from which one constructs a credible set of solutions. Taking the goal of his framework, we introduce a new approach to learning CBI which uses a different and much richer class of hypotheses to predict a credible set of solutions. In our approach, a hypothesis is a pair of multi-category classifiers. We model learning CBI as two multi-category learning problems. We provide and mathematically justify an automatic procedure that transforms any given case-base into two sets of samples that can be used by any supervised learning algorithm to learn CBI. We then perform an analysis of the error of a hypothesis that any algorithm may provide as output when training on these samples. We provide bounds on this error that can serve as a guiding criterion for the design of successful algorithms.

One main contribution has been to show how learning CBI over the wide spectrum of complex and unstructured CBR domains can now be tackled by standard off-the-shelf supervised-learning algorithms. Another contribution is in showing how the large-width advantage (related to the branch of learning theory known as large margin-learning) can also be realised for learning CBI.

Acknowledgements

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886 and by a research grant from the Suntory and Toyota International Centres for Economics and Related Disciplines at the London School of Economics. We are grateful to the referees for their careful reading of the paper and their many helpful comments.

References

[1] J. Kolodner. An introduction to case-based reasoning. *Artificial Intelligence Review*, 6:pp. 3–34, 1992.

[2] E. Hüllermeier. *Case-Based Approximate Reasoning*, volume 44 of *Theory and Decision Library*. Springer, 2007.

[3] W. Cheetham and I. D. Watson. Fielded applications of case-based reasoning. *Knowledge Eng. Review*, 20(3):321–323, 2005.

[4] E. Hüllermeier. Credible case-based inference using similarity profiles. *IEEE Trans. Knowl. Data Eng.*, 19(6):847–858, 2007.

[5] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[6] D. B. Leake. CBR in Context: The Present and Future. In D. B. Leake, editor, *Case-Based Reasoning: Experiences, Lessons, and Future Directions*. AAAI Press/MIT Press, 1996.

[7] M. Anthony and J. Ratsaby. Learning on finite metric spaces. Technical Report RRR 19-2012, RUTCOR, Rutgers University, 2012.

[8] D. Dubois, E. Hullermeier, and H. Prade. Fuzzy set-based methods in instance-based reasoning. *Fuzzy Systems, IEEE Transactions on*, 10(3):322–332, Jun 2002.

[9] D. Dubois, H. Prade, F. Esteva, P. Garcia, L. Godo, and R. López de Màntaras. Fuzzy set modelling in case-based reasoning. *International Journal of Intelligent Systems*, 13(4):345–373, 1998.

[10] S. Ontañón and E. Plaza. On knowledge transfer in case-based inference. In B. D. Agudo and I. Watson, editors, *Case-Based Reasoning Research and Development*, volume 7466 of *Lecture Notes in Computer Science*, pages 312–326. Springer Berlin Heidelberg, 2012.

[11] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[12] D. W. Aha. Lazy learning. chapter Lazy Learning, pages 7–10. Kluwer Academic Publishers, Norwell, MA, USA, 1997.

[13] M. Anthony and J. Ratsaby. Maximal width learning of binary functions. *Theoretical Computer Science*, 411:138–147, 2010.

[14] M. Anthony and J. Ratsaby. Analysis of a multi-category classifier. *Discrete Applied Mathematics*, 160(16-17):2329–2338, 2012.

[15] M. Anthony and J. Ratsaby. *The performance of a new hybrid classifier based on boxes and nearest neighbors*. Presented at International Symposium on Artificial Intelligence and Mathematics (ISAIM'12), Fort Lauderdale, FL, USA, January 9–11, 2012.

[16] B. Bollobas and O. Riordan. *Percolation*. Cambridge University Press, 2006.

[17] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

[18] M. Anthony and J. Ratsaby. Sample width for multi-category classifiers. Technical Report RRR 29-2012, RUTCOR Research Report, November 2012.

[19] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):pp. 233–235, 1979.

[20] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[21] B. Mohar. Laplace eigenvalues of graphs - a survey. *Discrete Mathematics*, 109(1-3):171–183, 1992.

[22] B. Mohar. Some applications of Laplace eigenvalues of graphs. *Graph Symmetry: Algebraic Methods and Applications*, 497:227–275, 1997.