Maximal-margin case-based inference

Martin Anthony
Department of Mathematics
London School of Economics
Houghton Street
London WC2A 2AE, UK
Email: m.anthony@lse.ac.uk

Joel Ratsaby
Electrical and Electronics Engineering Department
Ariel University of Samaria
Ariel 40700, Israel
Email: ratsaby@ariel.ac.il

***Abstract*—The central problem in case-based reasoning (CBR) is to produce a solution for a new problem instance by using a set of existing problem-solution cases. The basic heuristic guiding CBR is the assumption that similar problems have similar solutions. CBR has been often criticized for lacking a sound theoretical basis, and there has only recently been some attempts at developing a theoretical framework, including recent work by Hullermeier, who made a link between CBR and the probably approximately correct (or PAC) probabilistic model of learning in his 'case-based inference' (CBI) formulation. In this paper we present a new framework of CBI which models it as a multi-category classification problem. We use a recently-developed notion of geometric margin of classification to obtain generalization error bounds.**

I. Introduction

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 [16]. The basic heuristic that guides CBR is the hypothesis that similar problems have similar solutions [13]. The area of CBR research has had practical success and has been shown to be widely applicable [10]. 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 [13]. These steps are useful for devising a practical CBR system, but can at best represent CBR as an informal or non-rigorous model of AI. CBR has been often criticized for lacking a sound theoretical basis and there has only recently been some attempts at formalizing CBR in a theoretical framework. An important step in this direction was made by Hullermeier [14] who established a link between CBR and the probably approximately correct (PAC) theoretical model of learning (see [1], for instance). Hullermeier defines case-based reasoning as a prediction process, which allows him to make the connection between CBR and learning based on a training sample. He calls this framework *case-based inference* (CBI) which 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. These last two stages of the R^4 framework use not just the set of promising (or, credible) solutions but also domain-knowledge, user input and further problem-solving strategies [13]. As noted in [17] section 5.4, these steps *adapt* the set of promising solutions into a solution that fits the existing problem.

In this paper we present a new framework of CBI which extends that of [13] and enables us to represent the problem of case-based learning in the mathematical framework of 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 [4] 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.

II. Case-based inference (CBI)

In the framework of case-based inference, we have a problem space, denoted by $\mathcal{X}$, and a solution space, denoted by $\mathcal{Y}$. We define $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. We assume each space has a metric $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ associated with it (which, therefore, in particular, satisfy the triangle inequality). We also assume that each of the two metric spaces has a finite diameter $\text{diam}(\mathcal{X}) := \max_{x,x' \in \mathcal{X}} d_{\mathcal{X}}(x,x') < \infty$, $\text{diam}(\mathcal{Y}) = \max_{y,y' \in \mathcal{Y}}(y,y') < \infty$.

In the Introduction, we mentioned that CBI infers as an output a *set* of 'promising', or credible solutions rather than solving the full CBR problem by predicting a single specific solution. This is at the basis of what Hullermeier [14] calls *approximate reasoning.* We describe a slight modification of his framework. It uses a function h from $[0, \text{diam}(\mathcal{X})]$ to $[0, \text{diam}(\mathcal{Y})]$. Suppose we are 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. Given a problem $x \in \mathcal{X}$ for which we wish to infer a set of good solutions, the function h produces a subset of $\mathcal{Y}$. In this way, it defines a mapping from $\mathcal{X}$ to the power set $2^{\mathcal{Y}}$ of $\mathcal{Y}$ (which consists of all subsets of $\mathcal{Y}$), the output of which is the set of solutions inferred for x. This set is $C_h(x)$, where

$$C_h(x) := \bigcap_{i=1}^{m} \Gamma_h(z_i, x) \subseteq \mathcal{Y} \qquad (1)$$

where $\Gamma_h(z_i, x)$ is a ball centered at y_i with a radius given by the value $h(d_{\mathcal{X}}(x, x_i))$. That is, in the CBI framework of [14], each possible h specifies the radius of every sphere centered at a sample solution y_i, based on the distance between the corresponding problem x_i and the given problem x. Note that, in general, not all of the balls play an important role in shaping the credible solution set $C_h(x)$, but rather only those whose radii are relatively small. Thus, depending on the sample, the *effective* number of spheres (cases) which participate in shaping the credible set may be much smaller than the sample size m. The definition of the set C_h is based solely on the function h and the sample, and Hullermeier [14] discusses how to update the function h.

From (1), a credible set must take the form of an intersection of spheres in the solution space. While this is sensible from a practical viewpoint, it sets an inductive bias [18] which might result in worse generalization error bounds. In this paper we extend this CBI framework such that no *a priori* inductive bias is placed through a choice of a particular class of mappings from $\mathcal{X}$ to sets of credible solutions.

Our idea is based on learning 'hypotheses' (multi-category classifiers) on each of the metric spaces $\mathcal{X}$ and $\mathcal{Y}$ individually, and by taking account of the sample-width, an idea we introduced in [2, 4] and applied in [3, 6]. This leads to the favoring of 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.

We now describe the new CBI framework.

III. A NEW CBI FRAMEWORK

In this section we extend the CBI model of [14] described in section II. The underlying assumption that similar problems must have similar solutions is represented in this new framework through a preference for smooth hypotheses that map similar problems to similar solutions (discussed in section IV). The advantage of our new framework compared to that of [14] is that we derive rigorous error-bounds that are sample-dependent, which allows hypotheses to be any functions.

A. 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 credible solution) 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:

$$\begin{aligned} Q(Z) &= \sum_{k\in[K]} Q_{Z|M}(Z\,|k)Q_M(k) \qquad (2)\\ &= \sum_{k\in[K]} Q_{Y|M}(Y\,|k)Q_{X|Y,M}(X|Y,k)Q_M(k), \end{aligned}$$

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.*

For any probability distribution P on $\mathcal{Y}$ denote by supp$(P) \subseteq \mathcal{Y}$ the probability-1 support of P. We further 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 $\mathcal{Y}$:

1) For $k \neq k'$, we have supp $\left(Q_{Y|M}\left(Y|k\right)\right) \bigcap$ supp $\left(Q_{Y|M}\left(Y|k'\right)\right) = \emptyset$
2) For any y, $y' \in \mathcal{Y}$ such that $d_\mathcal{Y}(y, y') \leq \tau$, there exists $k \in [K]$ such that $y, y' \in$ supp $\left(Q_{Y|M}(Y|k)\right)$
3) For any $\alpha \in (0, 1)$, there is $m_0^Q(\alpha)$ such that if a sequence of $m \geq m_0^Q$ elements of $\mathcal{Z}$, $\mathrm{s}^{(m)} = \{(x_i, y_i)\}_{i=1}^m$, is drawn according to the product probability measure Q^m, then, with probability at least $1 - \alpha$, for each $k \in [K]$, 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 $d_\mathcal{Y}(y_{j_N}, y_{i_2}) \leq \tau$, for $1 \leq l \leq N - 1$.

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 τ distance apart in the solution space. Condition (C) implies 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. 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 above conditions imply that, for a given x, if its true (unknown) solution y 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 credible solution for x; and if the region is not small, then from condition (C) it still must contain 'typical' points of this mode rather than just 'outliers' that deviate from the expected value of the mode. With these typical points in the credible set, the third and fourth stages of the R^4 model may induce a good candidate solution for x based on the credible set.

Thus, in the new CBI framework, we choose a whole mode region k as the inferred credible-solution set for any problem x whose true (unknown) solution y is predicted to fall in mode region k.

Learning CBI amounts to learning to map an x to a mode that, with high confidence, contains the true solution y, and then predict the corresponding mode region as a credible-solution 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 [9], 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 the framework, let us define two probability functions that we refer to in subsequent sections,

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

B. 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

$$h(z) : \;=\; [h_1(x), h_2(y)] \tag{3}$$

and refer to the vector value function $h : \mathcal{Z} \to [K]^2$ as a *hypothesis* for case-based inference. Note that $[K]$ is the same set in (2) 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 and a problem x, the credible solution *set* $C(x)$ predicted by h is defined as

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

or, equivalently,

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

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 new CBI framework amounts to classifying x into one of a finite number of solution regions.

In section V we discuss how to learn h by learning the two classifiers h_1 and h_2. We learn each individually 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 framework.

In section V 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 the unknown solution y of x) to be the credible solution set for x.

If h is sufficiently accurate then, with a large confidence, the predicted credible set contains the true unknown solution y of 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.

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 $Q_{Y|M}(Y|k)$, $k = 1, \ldots, 3$ 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 $S_k^{(2)}$, $1 \le k \le 3$. 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, $y' \in S_2^{(2)}$. In general, the mode regions of Q need not be circular, and the decision regions of h need not be box-shaped as in this example.

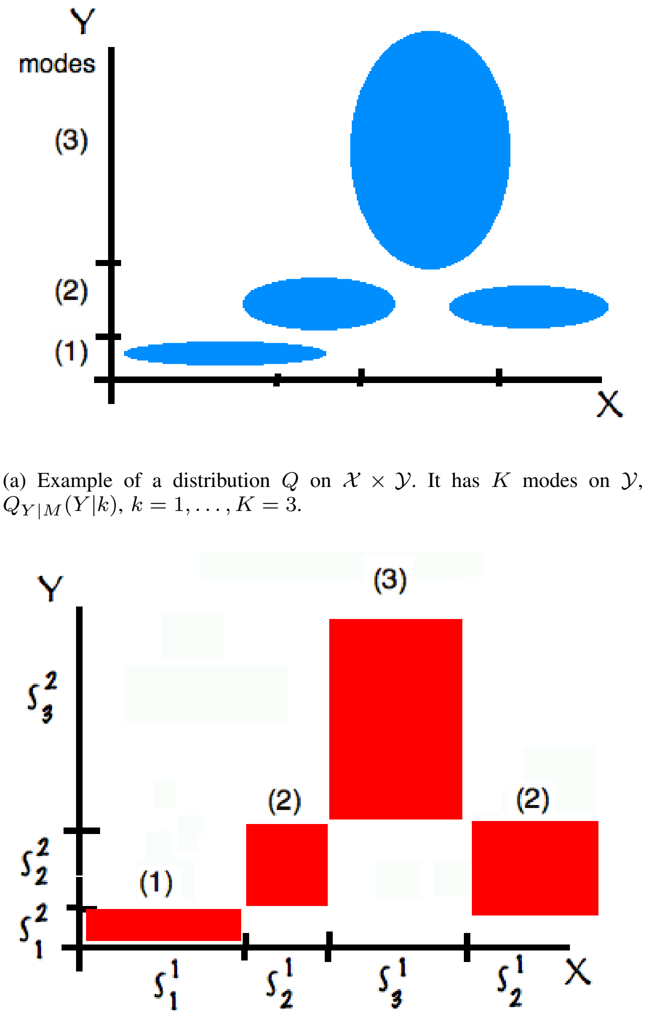
In the learning framework that we introduce in section V the number of modes K is not assumed to be known. The value of K is estimated based on a training sample of problem-solution pairs and on knowing the value of τ (which is given as domain knowledge). The estimate of K may be as large as the sample size m (as shown in (10)).

C. 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 mispredicts 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

$$\mathrm{err}\,(h) := P\left(Y \notin C_h(X)\right). \tag{4}$$

Since the two components of h are classifiers, then the event that h mispredicts (X, Y) implies that the two component classifiers disagree on the category of the credible-solution. We can represent this as follows: denote by M the 'true' unknown solution category of the random solution Y where $M \in [K]$.



(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 , $K = 3$.

Fig. 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 $K = 3$.

Then the probability of mispredicting is

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

which, it can be shown (see [5]), is bounded from above by

$$P_{\mathcal{X}}\left(h_1(X) \neq M\right) + P_{\mathcal{Y}}\left(h_2(Y) \neq M\right). \qquad (5)$$

The first and second term in (5) 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 $\text{err}(h_1)$ and $\text{err}(h_2)$ and therefore we have

$$\text{err}(h) \leq \text{err}(h_1) + \text{err}(h_2). \qquad (6)$$

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 (6) are implicitly dependent on the set $[K]$. In section V, 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 V-A.

We now proceed to present new results on learning multi-category classification on metric spaces which we subsequently use for the CBI learning framework in section V.

IV. 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 III, that is, on each of the two terms on the right side of (6). 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 III.)

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 mutually exclusive, $S_k \bigcap S_{k'} = \emptyset$ 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:

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

As in [4] we define the notion of *width* of a classifier h at a point x as follows,

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

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* [12] which we denote by $f_h : \mathcal{X} \times [K] \to \mathbb{R}$ and which is defined as follows:

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

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:

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

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

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

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

$$\text{er}_P(h) := P\left(h(X) \neq L\right).$$

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 as the *empirical margin* error of h and denote it as

$$\begin{aligned}\hat{\text{err}}_\gamma\,(h) \quad &:= \quad \hat{P}_m\left(g_L(X) - \max_{k \neq L} g_k(X) < \gamma\right) \\ &= \quad \frac{1}{m}\sum_{i=1}^m \mathbb{I}\left\{g_{l_i}(x_i) - \max_{k \neq l_i}(x_i) < \gamma\right\}.\end{aligned}$$

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

In [7], 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 α-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 [11]) 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 [7]. 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.

Theorem 4.1: Suppose that $\mathcal{X}$ is a metric space of diameter $\text{diam}(\mathcal{X})$ 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 $\gamma \in (0, \text{diam}(\mathcal{X})]$, if $\hat{\text{err}}_\gamma(h) = 0$, then

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

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

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 (which is more general than the one just presented, but which is looser in the special case of zero empirical margin error) bounds the error in terms of the empirical margin error (which may be nonzero).

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 $\gamma \in (0, \text{diam}(X)]$,

$$\text{er}_P(h) \leq \text{err}_\gamma(h) \\ + \sqrt{\frac{2}{m}\left(K\,\mathcal{N}_{\gamma/6}\ln\left(\frac{18\,d(\mathcal{X})}{\gamma}\right) + \ln\left(\frac{2\,d(\mathcal{X})}{\gamma\delta}\right)\right)} + \frac{1}{m},$$

where $\mathcal{N}_{\gamma/6}$ is the $\gamma/12$-covering number of $\mathcal{X}$ and $d(\mathcal{X}) = \mathrm{diam}(\mathcal{X})$.

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

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

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 $\mathrm{err}_\gamma(h)$ 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.

V. A probabilistic framework for learning CBI

The learning model considered in [14] is based on the candidate-elimination algorithm, specifically, algorithm Find-S of [18], and is applied to concept-learning. As mentioned in section II, Hullermeier [14] uses functions h that map distance values in $\mathcal{X}$ to distance values in $\mathcal{Y}$. His class of possible h consists of piecewise-constant mappings from $\mathbb{R}_+$ to $\mathbb{R}_+$. He uses the generalization heuristic of the Find-S algorithm in order to find a good h which is consistent with the sample.

In this paper we introduce a new learning model for CBI which is based on learning multi-category classification. We use the new framework of CBI (section III) where hypotheses are two-dimensional multi-category functions $h = [h_1, h_2]$. Each of the two components of h are learned individually based on two auxiliary samples, one which consists of labeled problems and the other consists of the corresponding labeled solutions. We now proceed to describe how these samples are defined from the cases.

A. 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),

$$\mathbf{s} : \;=\; \mathbf{s}^{(m)} = \{(x_i, y_i)\}_{i=1}^m . \tag{7}$$

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

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

and

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

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

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

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 $\mathbf{s}$ and the value τ.

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

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

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

$$A_\tau : \;=\; [a(i,j)] \tag{8}$$

as follows:

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

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 $i \in \mathcal{Y}_{|\mathbf{s}}$ that are a distance at most τ from the point $j \in \mathcal{Y}_{|\mathbf{s}}$.

The matrix A_τ defined in (8) is an adjacency matrix of a graph $G_\tau = (\mathcal{Y}_{|\mathbf{s}}, E_\tau)$, 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 $D_\mathbf{s}[i,j] \leq \tau$.

Let $\{H_i\}_{i=1}^{K_\tau}$ be the set of K_τ connected components $H_i \subseteq \mathcal{Y}_{|\mathbf{s}}$ 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 [15].

Note that $K_\tau := K_\tau(\mathbf{s})$ is dependent on the sample $\mathbf{s}$ through $\mathcal{Y}_{|\mathbf{s}}$ 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 $\mathbf{s}$ into the subsets $\mathbf{s}^{(k)} \subseteq \mathbf{s}$ based on these components H_k as follows:

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

Then, define two auxiliary sets of samples as follows:

$$\zeta_\mathcal{X} := \zeta_\mathcal{X}^{(m)}$$

where

$$\zeta_\mathcal{X}^{(m)} = \left\{(x_i, k) : x_i \in \mathcal{X}_{|\mathbf{s}}, (x_i, \cdot) \in \mathbf{s}^{(k)}, 1 \leq i \leq m, 1 \leq k \leq K_\tau\right\},$$

and

$$\zeta_\mathcal{Y} := \zeta_\mathcal{Y}^{(m)}$$

where

$$\zeta_\mathcal{Y}^{(m)} = \left\{(y_i, k) : y_i \in \mathcal{Y}_{|\mathbf{s}}, (\cdot, y_i) \in \mathbf{s}^{(k)}, 1 \leq i \leq m, 1 \leq k \leq K_\tau\right\}. \tag{9}$$

We use these samples for the classification learning problems in section V-B. 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 [19, 20] 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

$$K_\tau \quad = \quad \min\left\{m, \mu_0(G_\tau)\right\}. \tag{10}$$

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 σ of the set $[K]$ a label value $j \in [K]$ is in one-to-one correspondence with a mode value $\sigma(j) \in [K]$.

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

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, $v \in \mathcal{Y}_{|\mathbf{s}}$ (which may be y or y') where $u \in R_k$ and $v \in R_{k'}$. But by condition (B) of section III 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 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 $H_j \subseteq \text{supp}\left(Q_{Y|M}(Y|k)\right)$.

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 III, 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 probability-1 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 an example $(x_i, y_i) \in \mathbf{s}$ with y_i in a connected component H_k of the graph G_τ then l_i equals k where k is the value of the true (unknown) mode (up to a permutation). 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 is the value of the true mode (up to a permutation).

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

B. Two classification problems

Given the two auxiliary samples $\zeta_{\mathcal{X}}$ and $\zeta_{\mathcal{Y}}$ of (9) we learn two multi-category classification problems, independently, by finding a component hypothesis h_1 and h_2 which classify $\zeta_{\mathcal{X}}$ and $\zeta_{\mathcal{Y}}$, with a large sample-width, respectively. Based on h_1 and h_2 we form a hypothesis $h = [h_1, h_2]$ as in (3), where by (6) its error is bounded by the sum of the errors of h_1 and h_2.

As mentioned above, the number of categories $K_\tau(\mathbf{s})$ is dependent on the sample $\mathbf{s}$, or more specifically on the set $\mathcal{Y}_{|\mathbf{s}}$. Thus we need to make the bounds of section IV 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 $\mathbf{s}$ 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 $\mathbf{s}$. 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 $P_{\mathcal{Y}}$.

Lemma 5.5: Let $\alpha \in (0, 1)$ and $m \geq m_0^Q(\alpha)$. Let $\mathbf{s}$ 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 $\mathbf{s}$. 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 $P_{\mathcal{X}}$.

The proofs of these lemmas can be found in [5].

VI. Learning bounds

Recall that what we want to do is obtain a high-probability bound on the error $\text{err}_P(h)$ 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 (5), this error is bounded by the sum

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

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 (the proofs of which may be found in [5]).

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 $\gamma_2 \in (0, \mathrm{diam}(\mathcal{Y})]$ (where $\mathrm{d}(\mathcal{X}) = \mathrm{diam}(\mathcal{X})$ and $\mathrm{d}(\mathcal{Y}) = \mathrm{diam}(\mathcal{Y})$), and for all $h = [h_1, h_2]$: 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

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

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

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

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 $\gamma_2 \in (0, \mathrm{diam}(\mathcal{Y})]$ (where $d(\mathcal{X}) = \mathrm{diam}(\mathcal{X})$ and $d(\mathcal{Y}) = \mathrm{diam}(\mathcal{Y})$), and for all $h = [h_1, h_2]$:

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

where

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

and

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

VII. Conclusions

We have discussed a new way of modeling probabilistically the process of learning for case-based inference. We have done so through framing it as two related multi-category classification problems, and using recently-developed bounds for generalization by any type of classifier.

Acknowledgements

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

References

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

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

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

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

[5] M. Anthony and J. Ratsaby. Large margin case-based reasoning. *RUTCOR Research Report RRR 2-2013, Rutgers University*, 2013.

[6] M. Anthony and J. Ratsaby. The performance of a new hybrid classifier based on boxes and nearest neighbors. In *Proceedings ISAIM 2012*.

[7] M. Anthony and J. Ratsaby. Sample width for multi-category classifiers. *RUTCOR Research Report RRR 29-2012, Rutgers University*, 2012.

[8] M.J.B. Appel and R.P. Russo. The connectivity of a graph on uniform points on $[0,1]^d$. *Statistics and Probability Letters* 60: 351–357.

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

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

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

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

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

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

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

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

[17] 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.

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

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

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