

A BOOLEAN MEASURE OF SIMILARITY

Martin Anthony^a Peter L. Hammer^b

RRR 27-2004, AUGUST 2004

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
<http://rutcor.rutgers.edu/~rrr>

^aDepartment of Mathematics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom, m.anthony@lse.ac.uk

^bRUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway, NJ 08854-8003, USA, hammer@rutcor.rutgers.edu

RUTCOR RESEARCH REPORT

RRR 27-2004, AUGUST 2004

A BOOLEAN MEASURE OF SIMILARITY

Martin Anthony

Peter L. Hammer

Abstract. We propose a way of measuring the similarity of a Boolean vector to a given set of Boolean vectors, motivated in part by certain data mining or machine learning problems. We relate the similarity measure to one based on Hamming distance and we develop from this some ways of quantifying the ‘quality’ of a dataset.

Acknowledgements: Part of this work was carried out while Martin Anthony was visiting RUTCOR, Rutgers University. Martin Anthony’s work is supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. The authors thank Alex Kogan, Ersoy Subasi, Mine Subasi, and Ben Veal for useful discussions.

1 Introduction

In this paper we propose a way of measuring the similarity $s(x, A)$ of a Boolean vector x to a set A of such vectors. The measure we propose is based on the *absence* of certain substrings of x from the set of vectors in A .

In the context of machine learning classification problems, we may think of A as a dataset, a set of observations on which we know the correct classifications. For example, each observation in the data set might arise from a set of medical tests on a patient and may represent, suitably encoded, the absence or presence—or degree of presence—of a number of symptoms the patient may have. We believe that, in this context, the similarity measure provides a natural way of deciding which unseen possible observations it would be plausible to classify with some confidence once a classifier has been found that correctly classifies all (or most of) the observations in the dataset.

Elegant and useful theories of classification error and confidence have been developed, but these usually make probabilistic assumptions about the way in which the observations have been generated. Specifically, the PAC model of learning and its variants (see, for instance [15, 7, 1, 10]) assume that each observation in the data set has been chosen independently of the others, at random, according to a fixed probability distribution on $\{0, 1\}^n$, the set of all conceivable observations. Vovk *et al.* [17, 18, 16] have studied on-line learning in which one wants not only to predict classifications, but to give some indication of how ‘credible’ such predictions are, or not to predict if the predictions are not to be credible; and this is similar to the type of application we have in mind for the similarity measure. But in these papers, it is also assumed that the observations are generated independently according to the same probability distribution. In practice, what can one do without such probabilistic assumptions? It may be hard to prove anything sensible about classification accuracy in this case. Nonetheless, it might be at least useful not only to determine a classifier and to classify unseen observations with it, but also to attach to such predicted classifications the indication $s(x, A)$ of how similar the observation x is to those in the dataset that we trained on. Equally, one may decide not to classify at all those unseen observations that have a low similarity with the dataset. Empirical investigations (to be reported in [4]) appear to indicate that a higher classification accuracy is then achieved on the region of the domain $\{0, 1\}^n$ on which we *do* decide to classify.¹

We also propose a way of using the extremal values of $s(x, A)$ as an indication of the ‘quality’ or ‘representativeness’ of a dataset.

One obvious measure of how similar an observation is to a set of observations is the minimum Hamming distance of the observation from the set. We relate the similarity measure proposed here to this Hamming distance based one.

¹Here, we are discussing similarity of an observation to the dataset—that is, to our set of training observations. However, another approach that is potentially of use in classification problems, is to use, separately, the similarities of an observation to the separate classes of observations in the dataset (that is, to the set of training examples of each class). We are currently investigating the use of classifiers derived from such considerations [6].

2 A similarity measure

2.1 Formal definition and motivation

Suppose $x \in \{0, 1\}^n$, $I \subseteq [n] = \{1, 2, \dots, n\}$, and $|I| = k$. Then the projection of x onto I is the k -vector obtained from x by considering only the coordinates in I . For example, if $n = 5$, $I = \{2, 4\}$ and $x = 01001$ then $x|_I = 10$.

By a *positional substring* of $x \in \{0, 1\}^n$, we mean a pair (z, I) where $z = x|_I$. The key point here is that the coordinates in I are specified: we will want, as part of our later definitions, to indicate that two vectors x and y have the same entries *in exactly the same places*, as specified by some $I \subseteq [n]$. For instance, although both $x = 10101$ and $y = 01010$ have substrings equal to 00 , there is no I such that $x|_I = y|_I = 00$.

We now give the definition of similarity studied in this paper.

Definition 2.1 For $A \subseteq \{0, 1\}^n$ and $x \in \{0, 1\}^n$, the similarity of x to A , $s(x, A)$, is defined to be the largest s such that every positional substring (x, I) of length s appears also as a positional substring (y, I) of some observation $y \in A$. That is,

$$s(x, A) = \max\{s : \forall I \subseteq [n], |I| \leq s, \exists y \in A, y|_I = x|_I\}.$$

Here $x|_I$ denotes the projection of x onto the coordinates indicated by I .

Equivalently, if r is the smallest length of a positional substring possessed by x that does not appear (in the same positions) anywhere in A , then $s(x, A) = r - 1$.

Notice that $s(x, A)$ is a measure of how similar x is to a *set* of vectors. It is not a metric or distance function. It can immediately be seen, indeed, that if A consists solely of one vector y , not equal to x , then $s(x, A) = 0$, since there must be some coordinate on which x and y differ (and hence a positional substring of length 1 of x that is absent from A).

Informally, the similarity of x to A is low if x has a short positional substring absent from A ; and the similarity is high if all positional substrings of x of a fairly large length can be found in the same positions in some $y \in A$. To use the medical analogy discussed earlier, if x has a small combination of symptoms (that is, a simple syndrome) that does not appear in any of the patients in the set A then x has low similarity to A . Conversely, if $x \notin A$ then, certainly, it has some positional substring absent from A (as this is trivially true for the case $I = [n]$), but if the smallest such substring is long, then all simple syndromes indicated in x can be found among the patients of A . In this sense, x is similar to previously observed patients. One might expect that the presence or absence of a medical condition in a patient would be indicated by the patient having certain syndromes, and that short syndromes might carry more weight in such an explanation. For this reason, if a patient has a small syndrome not previously seen, one may want to be cautious in diagnosing the patient; whereas if all short syndromes possessed by the patient appear somewhere in the previously observed patients, one might have more confidence in a diagnosis on that patient.

We remark that the definition of similarity is related to the ideas of *witness set*. Given a set $V \subseteq \{0, 1\}^n$ and $v \in V$, a witness set for v in V is $I \subseteq [n]$ such that for all $v \in V \setminus \{v\}$,

$v|_I \neq x|_I$; that is, the entries in positions I are sufficient to distinguish v from all other vectors in V . Witness sets have been studied in [13] and, in a slightly different context (in which they correspond to *teaching sequences*, *keys*, or *specifying samples*) in computational learning theory [8, 14, 3]. It is easy to see that the similarity $s(x, A)$ is one less than the size of the smallest witness set for x in $A \cup \{x\}$.

Note that $s(x, A)$ is always between 0 and n ; and will equal n if and only if $x \in A$. The highest possible similarity to A that an element outside A can have is $n - 1$, and if x is not in A , then $s(x, A) = n - 1$ if and only if all n neighbors of x are in A . It is clear also that if $A \subseteq A'$ then $s(x, A') \geq s(x, A)$.

We now illustrate the definition with a small example.

Example Suppose the set A consists of the following 10 points of $\{0, 1\}^5$.

1	0	1	1	1
0	0	0	1	1
1	1	1	1	1
1	1	1	0	1
1	1	1	0	0
1	0	0	0	0
0	0	1	0	0
1	0	0	1	0
0	0	1	0	1
1	0	1	0	0

Note, first, that no x can have $s(x, A) = 0$, since this could only happen if, on one of the five coordinates, all elements of A had a fixed value, either 0 or 1. Consider any x of the form $x = 01x_3x_4x_5$. Since there is no $y \in A$ with $y|_{\{1,2\}} = x|_{\{1,2\}} = 01$, we have $s(x, A) = 1$. Consider, however, $x = 10101$. For this x , we have $s(x, A) = 3$, because all (positional) substrings of x of length 3 belong to A , but there is no $y \in A$ such that $y|_{\{1,2,4,5\}} = x|_{\{1,2,4,5\}} = 1001$. Suppose now that $x = 00001$. Then, since all (positional) substrings of x of length 2 appear in A , $s(x, A) \geq 2$. However, there are substrings of length 3 missing from A : for example, there is no $y \in A$ with $y|_{\{1,3,4\}} = x|_{\{1,3,4\}} = 000$. So $s(x, A) = 2$.

2.2 A Boolean function formulation

Any Boolean function $f : \{0, 1\}^n \rightarrow \{0, 1\}$ can be expressed by a *disjunctive normal formula* (or DNF), using *literals* $u_1, u_2, \dots, u_n, \bar{u}_1, \dots, \bar{u}_n$, where the \bar{u}_i are known as *negated literals*. A disjunctive normal formula is one of the form

$$T_1 \vee T_2 \vee \dots \vee T_k,$$

where each T_i is a *term* of the form

$$T_i = \left(\bigwedge_{i \in P} u_i \right) \wedge \left(\bigwedge_{j \in N} \bar{u}_j \right),$$

for some disjoint subsets P, N of $\{1, 2, \dots, n\}$. A Boolean function is said to be a k -DNF if it has a disjunctive normal formula in which, for each term, the number of literals ($|P \cup N|$) is at most k . For two Boolean functions f and g , we write $f \leq g$ if $f(x) \leq g(x)$ for all x ; that is, if $f(x) = 1$ implies $g(x) = 1$. Similarly, for two Boolean formulae ϕ, ψ , we shall write $\phi \leq \psi$ if, when f and g are the functions represented by ϕ and ψ , then $f \leq g$. A term T of a DNF is said to *absorb* another term T' if $T' \leq T$. A term T is an *implicant* of f if $T \leq f$; in other words, if T true implies f true. The terms in any DNF representation of a function f are implicants of f . The most important type of implicants are the *prime implicants*. These are implicants with the additional property that there is no other implicant of f absorbing T . Thus, a term is a prime implicant of f if it is an implicant, and if the deletion of any literal from T results in a non-implicant T' of f (meaning that there is some x such that $T'(x) = 1$ but $f(x) = 0$). If we form the disjunction of all prime implicants of f , we have a DNF representation of f .

Given A , we now define a sequence of $n + 1$ Boolean functions g_0, g_1, \dots, g_n , as follows. The function g_0 is taken to be the identically-0 function and, for $1 \leq k \leq n$, g_k is the most ‘general’ k -DNF function that is 0 on every member of A , in the sense that if f is a k -DNF function and $f(x) = 0$ for all $x \in A$ then $f \leq g_k$. It can be seen that g_k is the disjunction of all terms corresponding to positional substrings of length at most k that are not present in any element of A . For example, if the positional substring $(10, \{2, 4\})$ is not in A (that is, there is no $y \in A$ with $y_{\{2,4\}} = 10$) then, for $k \geq 2$, g_k will have as a term $u_2\bar{u}_4$.

For a subset B of $\{0, 1\}^n$ we denote by \mathbb{I}_B the characteristic, or indicator function, of B , satisfying $\mathbb{I}_B(x) = 1 \iff x \in B$. Then we have the following observation.

Proposition 2.2 *If \bar{A} denotes the complement $\{0, 1\}^n \setminus A$ of A , then we have*

$$0 \equiv g_0 \leq g_1 \leq g_2 \leq \dots \leq g_{n-1} \leq g_n = \mathbb{I}_{\bar{A}}.$$

Furthermore, $s(x, A) \geq r$ if and only if $g_r(x) = 0$.

2.3 Computing similarity

One approach to computing the similarity is to compute the functions g_k and use the fact that, for a given x , $s(x, A) \geq k$ precisely if $g_k(x) = 0$. For fixed k , a k -DNF formula for g_k can be computed in time $O(|A|n^k)$ by using what is essentially Valiant’s k -DNF learning algorithm [15, 2]. This proceeds as follows. Start with all terms of degree at most k and run through each observation in A in turn, deleting from the current set of terms those that are true on the current observation. Then, the disjunction of the remaining terms is g_k . Given any x , one can now determine whether $s(x, A) \geq k$ by establishing whether $g_k(x) = 0$. Of course, this algorithm is only efficient for (small) fixed k , not depending on n .

The problem of determining similarity can also be posed as a set covering problem. Note first that if we can determine the shortest positional substring possessed by x and absent from A , then $s(x, A)$ is one less than the length of this string. Now, fix $x \in \{0, 1\}^n$, and suppose $x \notin A$ (it being easy to check quickly whether $x \in A$). For $i = 1, 2, \dots, n$, let

$S_i = \{y \in A : y_i \neq x_i\}$. Then the smallest I such that for all $y \in A$, $y|_I \neq x|_I$ is exactly the smallest number of sets S_i needed to cover A . The standard greedy set-covering heuristic will therefore provide an efficient way of determining a number $s'(x, A)$ such that $s'(x, A) \leq s(x, A) \ln |A|$, enabling us at least to lower-bound the similarity.

3 Quantifying the representativeness of a dataset

We now propose ways in which the similarity measure can be used to indicate how representative a dataset is of the whole of $\{0, 1\}^n$. Assume henceforth that A is a proper nonempty subset of $\{0, 1\}^n$ ($\emptyset \neq A \neq \{0, 1\}^n$).

Definition 3.1 *The extent of $A \subseteq \{0, 1\}^n$, $e(A)$, is defined to be the maximum similarity of $x \notin A$ to A ; that is,*

$$e(A) = \max\{s(x, A) : x \in \bar{A}\},$$

where $\bar{A} = \{0, 1\}^n \setminus A$ is the complement of A .

Definition 3.2 *The pervasiveness, $p(A)$, of $A \subseteq \{0, 1\}^n$ is defined to be the minimum similarity of $x \in \{0, 1\}^n$ to A ; that is,*

$$p(A) = \min\{s(x, A) : x \in \{0, 1\}^n\}.$$

Note that, for $A \neq \{0, 1\}^n$, $0 \leq p(A) \leq e(A) \leq n - 1$.

Definition 3.3 *The porosity, $\pi(A)$, of A is defined to be the pervasiveness of \bar{A} , where $\bar{A} = \{0, 1\}^n \setminus A$ is the complement of A . That is,*

$$\pi(A) = p(\bar{A}) = \min\{s(x, \bar{A}) : x \in \{0, 1\}^n\}.$$

For $A \neq \emptyset$, $0 \leq \pi(A) \leq n - 1$.

All three of these quantities can be thought of in alternative ways, both geometrically and in terms of Boolean functions, as we now explore.

By a *cube* we mean a subcube of $\{0, 1\}^n$. Explicitly, a cube is a subset of $\{0, 1\}^n$ of the form $C_a = \{x \in \{0, 1\}^n : x|_I = a\}$, for some $I \subseteq [n]$ of cardinality k (where $0 \leq k \leq n$) and some $a \in \{0, 1\}^k$. The dimension of such a cube is $n - k$ (and its co-dimension is k). We refer to C_a as an $(n - k)$ -cube.

Proposition 3.4 *The extent $e = e(A)$ can be characterized in the following ways.*

1. *e is the largest number such that for all $x \notin A$, there is some $I \subseteq [n]$ with $|I| = e + 1$ and $x|_I \notin A|_I = \{y|_I : y \in A\}$; that is, the fact that $x \notin A$ can be demonstrated by exhibiting an $(e + 1)$ -length (positional) substring of x that cannot be found in A .*
2. *e is the smallest number such that every $x \notin A$ is contained in some $(n - e - 1)$ -cube that is disjoint from A .*

3. e is the smallest number such that the complement \bar{A} of A is a union of cubes of dimension $(n - e - 1)$.
4. The longest prime implicant of $\mathbb{I}_{\bar{A}}$ (the indicator function of \bar{A}) is of size $e + 1$.
5. e is the largest value of k such that g_k is not equal to the indicator function $\mathbb{I}_{\bar{A}}$.

Proposition 3.5 *The pervasiveness $p = p(A)$ can be characterized in the following ways.*

1. p is the largest number such that for all $I \subseteq [n]$ with $|I| = p$, $A|_I = \{0, 1\}^p$.
2. p is the largest number such that A intersects every $(n - p)$ -cube.
3. p is the smallest number such that there is some $(n - p - 1)$ -cube disjoint from A .
4. The shortest prime implicant of $\mathbb{I}_{\bar{A}}$ (the indicator function of \bar{A}) is of size $p + 1$.
5. $p + 1$ is the smallest value of k such that g_k is not the identically-0 function.

Pervasiveness is related to some other notions that have been studied in extremal combinatorics and learning theory. The first characterization given in Proposition 3.5 is equivalent to saying that I is an (n, p) -universal set (see [9]). Equivalently, suitably interpreting the vectors in A as functions $[n] \rightarrow \{0, 1\}$ in the obvious way, the pervasiveness is the *testing dimension* of A [11, 12]. Quite clearly, we must have $|A| \geq 2^{p(A)}$, so that $p(A) \leq \log_2 |A|$. A probabilistic argument [9] shows that, for each k , there is some set $|A|$ of cardinality at most $k2^k \log_2 n$ with $p(A) \geq k$. The set E of all even-weight vectors in $\{0, 1\}^n$ demonstrates that (for $k = n - 1$) the lower bound is tight: for, $|E| = 2^{n-1}$ and $p(E) = n - 1 = \log_2 |E|$ (since all the neighbors of any odd weight vector x are of even weight and hence $s(x, E) = n - 1$).

Proposition 3.6 *The porosity $\pi(A)$ can be characterized in the following ways.*

1. π is the largest number such that every $(n - \pi)$ -cube contains elements not in A .
2. The largest dimension of a cube contained entirely within A is $n - 1 - \pi$.
3. The shortest prime implicant of \mathbb{I}_A (the indicator function of A) is of size $\pi + 1$.

4 Hierarchies based on similarity and relationship with Hamming distance

The similarity measure provides a way of filtering, or grading, $\{0, 1\}^n$ according to similarity to a given set A . For $0 \leq k \leq n$, let

$$A_k = \{x \in \{0, 1\}^n : s(x, A) \geq k\}$$

be the set of Boolean vectors which have similarity at least k to A . Then we have the following *hierarchy*:

$$\{0, 1\}^n = A_0 \supseteq A_1 \supseteq \cdots \supseteq A_{n-1} \supseteq A_n = A.$$

So, for large k , A_k is the set of vectors highly similar to A . Such a hierarchy might be useful in machine learning, where we might decide, for instance, to form a classifier on the basis of the dataset A , but not to predict outside A_k for a particular choice of k . The rationale for this would be that vectors in $\{0, 1\}^n \setminus A_k$ are judged to be too dissimilar to those in A . Experimental evidence seems to indicate this is a good strategy [4].

For a particular A , the hierarchy will typically look as follows:

$$\{0, 1\}^n = A_0 = \cdots = A_p \supset A_{p+1} \supseteq \cdots \supseteq A_e \supset A_{e+1} = \cdots = A_{n-1} = A_n = A,$$

where ‘ \supset ’ denotes strict containment. (This is modified in the obvious way if $p = e$.) Here, $p = p(A)$ is the pervasiveness of A and $e = e(A)$ is the extent of A (so we see another characterization of these quantities). In terms of the Boolean functions g_k , we can see that A_k has indicator function \bar{g}_k , the complement of g_k . The set A_k can also be thought of geometrically: if B_k is the union of all $(n - k)$ -dimensional cubes that are contained entirely in the complement of A , then $A_k = \bar{B}_k$ is the complement of B_k . That is, A_k is obtained by deleting from $\{0, 1\}^n$ all cubes of co-dimension k that lie entirely outside A .

Another natural way to form a hierarchy of subsets of $\{0, 1\}^n$ is to use Hamming distance. Recall that the Hamming distance $d(x, y)$ between x, y in $\{0, 1\}^n$ is the number of entries on which they differ; and that, for $A \subseteq \{0, 1\}^n$, the Hamming distance of x to the set A is defined by $d(x, A) = \min\{d(x, y) : y \in A\}$. Then, for $0 \leq k \leq n$, let $D_k = \{x \in \{0, 1\}^n : d(x, A) \leq n - k\}$. We have the hierarchy

$$\{0, 1\}^n = D_0 \supseteq D_1 \supseteq \cdots \supseteq D_{n-1} \supseteq D_n = A.$$

The following result relates the similarity measure and Hamming distance, and hence the two hierarchies.

Theorem 4.1 *With the notation as above, suppose that $0 \leq k \leq n$ and that $s(x, A) \geq k$. Then $d(x, A) \leq n - k$. Thus, for all k , $A_k \subseteq D_k$.*

Proof: Suppose that $s(x, A) \leq k$ and that $d(x, A) > n - k$. Then, let y be on a shortest path from x to A , and such that $d(x, y) = n - k$. Then $d(y, A) \geq 1$. In particular, $y \notin A$. Let Q be the $(n - k)$ -dimensional subcube of $\{0, 1\}^n$ with x and y as diagonally opposite vertices. Since x and y are not in A , no point of Q can be (for if $z \in Q \cap A$, then $d(x, A) \leq d(x, z) + d(z, A) = d(x, z) \leq n - k$, which is a contradiction). Therefore the complement of A contains Q and hence x has a positional substring of length at least k that is absent from A . But this contradicts $s(x, A) \geq k$.

The following result indicates that the two hierarchies are, in general, quite different. It shows that the similarity-based hierarchy is more ‘discerning’ or discriminating than that arising from Hamming distance.

Theorem 4.2 *With the above notation, suppose that $1 \leq k \leq n$ and that $A_k \neq \{0, 1\}^n$. Then $\{0, 1\}^n \setminus A_k$ contains an element of $\{0, 1\}^n$ that is at Hamming distance only 1 from A .*

Proof: Since $A_k \neq \{0, 1\}^n$, there is x such that $s(x, A) < k$. Suppose that $s(x, A) = r < k$. We can suppose, without any loss of generality, that x is of the form $x = 0^{r+1}x'$ (where 0^{r+1} denotes a string comprising $r + 1$ contiguous 0s) and that for no $y \in A$ do we have $y|_{\{1, 2, \dots, r+1\}} = 0^{r+1}$. Now, since $s(x, A) = r$, for all I such that $|I| = r$, there is $y \in A$ with $y|_I = x|_I$. In particular, there is some $z \in A$ of the form $0^r 1 z'$. Now, consider $w = 0^{r+1} z'$. Clearly, $d(w, A) = 1$ since $w \notin A$ and $d(w, z) = 1$. Since $w|_{\{1, 2, \dots, r+1\}} = 0^{r+1}$, w has a positional substring of length $r + 1$ that does not appear in A , and so we have $s(w, A) \leq r < k$ and hence $w \in \{0, 1\}^n \setminus A_k$.

5 Future applications-oriented work

We are currently pursuing several lines of research of an empirical nature. Using standard machine learning databases, we are investigating to what extent the generalization accuracy of a classifier can be improved by restricting predictions to some A_k (and not offering predictions at all outside that part of the domain) [4].

We are also investigating the use of similarity in imputing missing entries in data: for example, given one missing binary attribute value, we might impute the value that gives the observation greater similarity to the known observations [5].

As indicated earlier, since a dataset on which one trains a classifier will consist of observations with different classifications, there may be some merit in considering the similarities of an unclassified observation to the disjoint sets of differently classified observations in the dataset. Explicitly, suppose the classification problem is a two-class classification problem, and that the observations in the dataset A are partitioned into A^+ (observations of class 1) and A^- (observations of class 0). Then the similarities $s(x, A^+)$ and $s(x, A^-)$ indicate how similar x is to the two classes. Precisely when and how to classify an observation on this basis requires careful investigation. For one thing, if the classes A^+ and A^- are not of approximately equal size, one might *a priori* expect, given the properties of the similarity measure, a larger similarity of an unseen observation with the larger of the two classes. Furthermore, one might sensibly demand, in order to have a reliable classification, that the similarity of the unseen observation to one of the classes be not just large, but also significantly larger than its similarity to the other class. We are currently experimenting with classification methods of this type [6].

References

- [1] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, UK, 1999.

- [2] Martin Anthony and Norman Biggs. *Computational Learning Theory: An Introduction*. Cambridge University Press, Cambridge, UK, 1992.
- [3] Martin Anthony, Graham Brightwell and John Shawe-Taylor. On specifying Boolean functions by labelled examples. *Discrete Applied Mathematics*, 61 (1995): 1–25.
- [4] Martin Anthony, Peter L. Hammer, Ersoy Subasi and Mine Subasi. Restricting the regions of confident prediction using a Boolean measure of similarity. In preparation.
- [5] Martin Anthony, Peter L. Hammer, Ersoy Subasi and Mine Subasi. Imputing missing values in data by means of a Boolean similarity measure. In preparation.
- [6] Martin Anthony, Peter L. Hammer, Ersoy Subasi and Mine Subasi. Pattern classification using a Boolean similarity measure. In preparation.
- [7] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler and Manfred Warmuth, Learnability and the Vapnik-Chervonenkis dimension, *Journal of the ACM*, 36(4), 1989: 929–965.
- [8] Sally A. Goldman and Michael J. Kearns. On the Complexity of Teaching. *Journal of Computer and Systems Sciences*, 50(1), 1995: 20–31.
- [9] Stasys Jukna. *Extremal Combinatorics With Applications in Computer Science*, Springer-Verlag, 2001.
- [10] Michael J. Kearns and Umesh Vazirani, *Introduction to Computational Learning Theory*, MIT Press, Cambridge, MA, 1995.
- [11] Kathleen Romanik, Approximate testing and learnability, In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Association for Computing Machinery Press, New York, 1992.
- [12] Kathleen Romanik and Carl Smith, Testing geometric objects, Technical Report UMIACS-TR-90-69, CS-TR-2437, University of Maryland, Maryland, 1990.
- [13] E. Kushilevitz, N. Linial, Y. Rabinovich and M. Saks. Witness sets for families of binary vectors. *Journal of Combinatorial Theory (A)* 73(2), 1996: 376–380.
- [14] Ayumi Shinohara and Satoru Miyano, Teachability in computational learning, *New Generation Computing*, 8, 1991: 337–347.
- [15] Leslie G. Valiant, A theory of the learnable. *Communications of the ACM*, 27 (11), 1984: 1134–1142.
- [16] C. Saunders, A. Gammerman and V. Vovk. Transduction with confidence and credibility. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1999: 722–726.

- [17] V. Vovk. On-line confidence machines are well-calibrated. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, 2002, Los Alamitos, CA, IEEE Computer Society: 187–196.
- [18] V. Vovk. Asymptotic optimality of Transductive Confidence Machine. In *Proceedings of the 13th International Conference on Algorithmic Learning Theory* (ed by N Cesa-Bianchi, M Numao and R Reischuk), Lecture Notes in Artificial Intelligence, vol 2533, 2002: 336–350.