A Generalized Minimum Disagreement Problem for Learning Applications

Paul Fischer and Norbert Klasner

Abstract

For applications in Machine Learning and forecasting the Minimum Disagreement problem plays a fundamental role. This is the combinatorial problem of selecting a best explanation for observed data from a set of possible ones. It has recently been shown that the Minimum Disagreement strategy can be outperformed by randomized strategies if the observed data is noisy. This new strategy amounts to finding not a single best explanation for the data, but to finding a number of good explanations which satisfy a pairwise independence condition. Here we show a general approach for solving this new combinatorial problem. We also show that for some classes the problem can be solved more efficiently by a problem specific approach.

Key words: Combinatorics, Minimum Disagreement, Optimization, Learning

1 Introduction

In many applications in Machine Learning and forecasting one is looking for a good explanation (or *hypothesis* [1]) of observed data in order to make predictions on further events or to derive an abstraction of the process which generated the observations. In this context the Minimum Disagreement problem plays a fundamental role. Using the Minimum Disagreement strategy means that one selects an empirically best hypothesis (i.e., one which makes the fewest number of mistakes on the observed data) from a fixed set of allowed hypotheses. One then uses this hypothesis for the predictions on new data. Within a number of learning models it can be proved that one achieves optimal prediction or generalization using this strategy, see, e.g., [5,4,3,1,2]. Depending on the set of allowed hypotheses, selecting a best one can sometimes be done efficiently but is often a combinatorially hard problem. The Minimum Disagreement strategy is also appropriate for learning from noisy data. However, for a successful application of the Minimum Disagreement strategy a lot of

[1] We shall use the term *hypothesis* henceforth.

data is required; informally, the reason is that only then the Minimum Disagreement strategy properly filters out the noise.

The Minimum Disagreement strategy is, moreover, not always optimal. For example it was shown that there are more complicated strategies, sometimes randomized ones, which achieve the same prediction performance but require fewer observations, see [1,2]. In [2] learning from very few and, in addition, noisy data is considered and a generic prediction algorithm for this situation is presented.

In [2] a situation is considered where information theoretic reasons prevent an application of the Minimum Disagreement strategy. Moreover, it is assumed that there is so few data such that the empirically best hypothesis might not even be close to the best possible one. The generic algorithm, replacing Minimum Disagreement in this situation, works as follows: Instead of finding a best hypothesis, one tries to find two (or three) "quite good" ones, which satisfy a certain independence condition. This condition requires that empirical the errors of the individual hypotheses occur at different parts of the observed data. Intuitively this means that, with high probability, only one of them makes a wrong prediction. It is proven ([2]) that selecting one of the two at random (or taking majority in the case of three) leads to a prediction rule with good performance. The generic algorithm and its sample size behavior are well analyzed ([2]). However, there are no examples for solving the key problem of finding independent hypotheses. Here we consider the combinatorial aspect of the problem, namely the problem of finding a set of two or three mutually independent hypotheses or certifying that no such pair exists. We first present a straightforward method, which is applicable in many cases. We then show that the problem can be solved more efficiently if the explanations are intervals. This corresponds to the situation where only one feature of the observed data is considered, a situation which is of great practical importance: Observed data usually consists of feature vectors (e.g., results of lab test of patients) and classifications (e.g., ill or healthy). It is well known in practice that restricting to a single feature allows to find a good rule of the thumb quickly, while considering a big number of features might blur the result or be computationally intractable.

2 Definitions and Statement of the Problem

To formalize the problem introduced before, let X be some set which we call the *universe*. An element $x \in X$ is called an *observation* or an *example*, e.g., an observation can be a vector of results of lab test. Every example is *labeled* with 0 or 1, e.g., 1 stands for "ill". We allow the labeling to be inconsistent, i.e., the same observation x might occur once with label 1 another time with label 0. This reflects the fact, that of two patients only one might be ill, although their lab results are identical. After having seen a *sample*, i.e., a sequence of labeled observations, one wants to set up a good hypothesis for further predictions. This means that, given a new unlabeled observation x (lab data of a new patient), the hypothesis correctly predicts the label of x with high probability. Usually one will not select any odd hypothesis but one from a small set of plausible ones. This is modeled by fixing a set $\mathcal{H} \subseteq 2^X$ of hypotheses one may choose from. Let $S := (\langle x_i, \ell_i \rangle)_{i=1,\ldots,m}$, $x_i \in X$, $\ell_i \in \{0,1\}$, be a (possibly inconsistent) sample. For a hypothesis H let $\mathrm{E}_S(H) := \{i \mid (x_i \in H \wedge \ell_i = 0) \vee (x_i \notin H \wedge \ell_i = 1)\}$ denote the *error set* of H with respect to sample S, and let $\mathrm{dis}_S(H) := |\mathrm{E}_S(H)|$ be the *empirical disagreement* of H on S. For a pair (H,G) of hypotheses let $\mathrm{E}_S(H,G) := \mathrm{E}_S(H) \cap \mathrm{E}_S(G)$ be the *joint error set* of H and G with respect to sample S, and $\mathrm{dis}_S(H,G) := |\mathrm{E}_S(H,G)|$ be the *empirical joint disagreement* of H and G on S. Let ε and λ be real numbers in $(0,1)$, $\lambda < \varepsilon$. A hypothesis H is (empirically) *ε-good* if $\mathrm{dis}_S(H) \leq \varepsilon m$. A pair (H,G) of hypotheses is called (empirically) *λ-independent* if $\mathrm{dis}_S(H,G) \leq \lambda m$. A set $\{H_1, \ldots, H_r\}$ of hypotheses is called *λ-independent* if every pair (H_i, H_j), $i \neq j$, is λ-independent.

Let $\mathcal{H} \subseteq 2^X$ be the set of allowed hypotheses. Then the problem (called the *3-hypotheses-problem for* $\mathcal{H}$, 3-HP($\mathcal{H}$) for short) can be stated as follows: Given $\varepsilon, \lambda \in (0,1)$, $\lambda < \varepsilon$, and any sample S find a λ-independent set of ε-good hypotheses of cardinality 3 or show that no such set exists. In the latter case find a λ-independent pair of ε-good hypotheses or show that no such pair exists. In the latter case find an ε-good hypothesis or show that no such hypothesis exists.

Note, that an algorithm for 3-HP($\mathcal{H}$) can be used to solve the minimum disagreement problem for $\mathcal{H}$: choose $\lambda = 0$ and $\varepsilon = 1/2$ and run the algorithm. Depending on the outcome set ε to 1/4 or 3/4 etc.

3 A General Solution

There is a general solution to the 3-HP($\mathcal{H}$) which is efficient if some conditions on the class $\mathcal{H}$ of allowed hypotheses hold. First, for any sample S there must not be too many essentially different hypotheses. Two hypotheses are called *essentially different* (with respect to S) if there is an example $\langle x, \ell \rangle \in S$ such that x belongs to exactly one of the two hypotheses. Let $\widehat{\mathcal{H}}_S$ denote this set.

Theorem 1 *If the following conditions are satisfied for every sample S*

(1) The class $\widehat{\mathcal{H}}_S$ is of size polynomial in the sample size $|S|$ and $\widehat{\mathcal{H}}_S$ can be enumerated in polynomial time.

(2) For every hypothesis $H \in \mathcal{H}$ its disagreement $\mathrm{dis}_S(H)$ can be computed in polynomial time.

(3) For every pair of hypotheses H and F from $\mathcal{H}$ the joint disagreement $\mathrm{dis}_S(H, F)$ can be computed in polynomial time.

then 3-HP($\mathcal{H}$) is efficiently solvable.

Proof. Let a sample S be given. One can efficiently enumerate the subset of $\widehat{\mathcal{H}}_S$ which contains the ε-good hypotheses. Let $\mathcal{B}_S$ denote this set. If $\mathcal{B}_S$ is empty we are done, otherwise we construct a graph $\mathcal{G} = (\mathcal{B}_S, E)$ where the vertex set is $\mathcal{B}_S$ and $\{F, H\}$ is an edge if F and H are λ-independent. If E is empty we are done. Otherwise, every triangle in $\mathcal{G}$ corresponds to a triple of ε-good and λ-independent hypotheses. Now compute the square $\mathcal{A}^2$ of the adjacency matrix $\mathcal{A}$ of $\mathcal{G}$ using integer arithmetic. The entry $\mathcal{A}^2(F, H)$ is the number of paths of length 2 from F to H. If $\mathcal{A}^2(F, H) \geq 1$ and, in addition, $\mathcal{A}(F, H) = 1$ then there is a vertex in $\mathcal{G}$ (i.e., a hypothesis $K \in \widehat{\mathcal{H}}_S$) such that K, H, F form a triangle in $\mathcal{G}$ (i.e., $\{K, H, F\}$ is a λ-independent set). As matrix multiplication can be performed in time $n^{2.37}$ where n the dimension of the matrix (here $n = |\mathcal{B}_S|$) we have an efficient algorithm for 3-HP($\mathcal{H}$) if the conditions hold. *Q.E.D.*

4 An Efficient Solution of the 3-HP for Intervals

Let $\mathcal{I}$ be the class of closed intervals on the real line, i.e., $\mathcal{I} := \{[a, b] \mid a, b \in \mathbb{R} \text{ and } a \leq b\}$. Let $S := (\langle x_i, \ell_i \rangle)_{i=1,\ldots,m}$, $x_i \in \mathbb{R}$, $\ell_i \in \{0, 1\}$, be a (possibly inconsistent) sample. The problem 3-HP($\mathcal{I}$) can be solved in $O(m^6)$ time. First note, that we may restrict ourselves to intervals the endpoints of which are sample points because these intervals form a maximum set of essentially different intervals. Let $\mathcal{I}_S$ denote this set; obviously $|\mathcal{I}_S| = O(m^2)$ such. After some preprocessing of time $O(m^2)$ the disagreement of every interval as well as the joint disagreement of any two such intervals can be computed in constant time. Hence time $O(m^6)$ suffices to check for every triple of intervals whether it forms a λ-independent set and whether the three intervals are all ε-good. Using the general approach of Section 3 this can be improved to $O(m^{4.74})$. We now want to show that 3-HP($\mathcal{I}$) can be solved in $O(m^4)$ by an approach which takes structural properties of the intervals into account.

Theorem 2 *Let $\mathcal{I}$ be the class of closed intervals on the real line. 3-HP($\mathcal{I}$) can be solved in time $O(m^4)$, where m is the sample size.*

Proof. We assume that the sample S is sorted according to first component and that the indexing reflects this ordering. In order to solve the 3-HP($\mathcal{I}$) we do some preprocessing. For convenience we introduce the (unlabeled) points $-\infty$ and $+\infty$ with the obvious meaning. For every pair x_i, x_j, such that $x_i < x_j$, of sample points we compute the numbers of 0-labels and 1-labels in the closed interval $[x_i, x_j]$. We also compute the these numbers for all sample points left of x_j including x_j, i.e., for the half-open interval $(-\infty, x_j]$. The corresponding numbers for open and half-open intervals are defined in the obvious way. E.g., for $r \in \{0, 1\}$ we define

$$\begin{aligned} L_r\left([x_i, x_j]\right) &:= |\{k \mid x_i \le x_k \le x_j \wedge \ell_k = r\}| \\ L_r\left((-\infty, x_j]\right) &:= |\{k \mid x_k \le x_j \wedge \ell_k = r\}| \end{aligned}$$

The computation is done by traversing the sample from left to right and can be performed in time $O(m^2)$. ¿From these values the empirical disagreement of an interval $I := [x_a, x_b]$ and the joint disagreement of every pair $I := [x_a, x_b]$, $J := [x_i, x_j]$ can be computed in constant time. Hence, in time $O(m^2)$, one can compute the set $\mathcal{B}_S \subset \mathcal{I}_S$ of those intervals which are ε-good. If $\mathcal{B}_S$ is empty then the problem is solved. We now solve the 3-HP for $\mathcal{B}_S$. Let $n := |\mathcal{B}_S|$ and let $I_1, \ldots, I_n$ be an enumeration of $\mathcal{B}_S$. We have $n = O(m^2)$. All intervals in the rest of this section are from $\mathcal{B}_S$. We shall first show how to check in time $O(n)$ whether $\mathcal{B}_S$ contains an independent **pair**. This algorithm is later also used as a building block in the solution of the complete problem. The algorithm is based on the following observation: For some interval I let I^* be an interval such that $\mathrm{dis}_S(I, I^*)$ is minimal. Clearly, if $\mathrm{dis}_S(I, I^*) > \lambda m$ then I is not part of any λ-independent pair. We now show that after an $O(n)$ preprocessing, I^* can be found in constant time for every I.

Now fix an interval $I := [x_a, x_b]$. Let $J := [x_i, x_j]$ be some other interval. There are six different ways in which I and J can located relative to each other.

P-1a:	$x_a \le x_b < x_i \le x_j$	P-2b:	$x_i \le x_a \le x_j \le x_b$
P-1b:	$x_i \le x_j < x_a \le x_b$	P-3:	$x_a \le x_i \le x_j \le x_b$
P-2a:	$x_a \le x_i \le x_b \le x_j$	P-4:	$x_i \le x_a \le x_b \le x_j$

In our approach we find for each of the six cases an interval J such that $\mathrm{dis}_S(I, J)$ is minimal. The solutions for the different cases are similar. Below we describe the method for Case P-2a in detail (P-2b is symmetric) and then mention the changes necessary for the other cases.

In Case P-2a the joint disagreement of I and J can be written as

$$\mathrm{dis}_S(I, J) = L_1\left((-\infty, x_a)\right) + L_1\left((x_j, +\infty)\right) + L_0\left([x_i, x_b]\right) \ . \qquad (1)$$

We want to find a J which minimizes (1). As the first term in (1) is constant (as I is fixed) this amounts to minimizing $L_1((x_j,+\infty)) + L_0([x_i,x_b])$, subject to the condition (P-2a). This expression can be rewritten as follows:

$$\begin{aligned}
&L_1((x_j,+\infty)) + L_0([x_i,x_b]) &(2)\\
&= [L_1((x_b,+\infty)) - L_1((x_b,x_j])] + [L_0([x_a,x_b]) - L_0([x_a,x_i))] \\
&= L_1((x_b,+\infty)) + L_0([x_a,x_b]) - L_1((x_b,x_j]) - L_0([x_a,x_i)) \ . &(3)
\end{aligned}$$

As I is fixed the first two terms in (3) are constant. Hence minimizing (1) is equivalent to maximizing $L_1((x_b,x_j]) + L_0([x_a,x_i))$. Again, because $I = [x_a,x_b]$ is fixed, maximizing this expression is equivalent to maximizing

$$L_1((-\infty,x_j]) + L_0((-\infty,x_i)) \ , \tag{4}$$

of course, subject to conditions (P-2a) and that $[x_i,x_j]$ is ε-good. The latter condition prevents that one can maximize the two terms in (4) independently. Recalling that the indexing of the sample points is compatible with their ordering on the real line we compute for all i,j, such that $1 \le i \le j \le m$

$$U_{i,j} := \begin{cases} L_0((-\infty,x_i)) + L_1((-\infty,x_j]) & \text{if } [x_i,x_j] \in \mathcal{B}_S \\ -\infty & \text{otherwise} \end{cases}$$

Computing the matrix U from the precomputed L_r-values can be done in time $O(m^2)$. Noting that the finite values in U correspond to ε-good intervals we rewrite (4)

$$\max_{\substack{i,j\\ a\le i\le b\le j}} U_{i,j} = \max_{\substack{i\\ a\le i\le b}} \max_{\substack{j\\ b\le j}} U_{i,j} \ . \tag{5}$$

The inner and outer maxima in (5) are computed by a postfix computation, which can be performed in time $O(m^2)$:

For $i = 1,\ldots,m$ do
  $Z_{i,m} := U_{i,m}$
  For $b = m-1,\ldots,1$ do
    $Z_{i,b} := \begin{cases} \max\{Z_{i,b+1}, U_{i,b}\}, & \text{if } i \le b \\ -\infty, & \text{otherwise} \end{cases}$

For $b = 1,\ldots,m$
  $R_{m,b} := Z_{m,b}$
  For $a = m-1,\ldots,1$ do

$$R_{a,b} := \begin{cases} \max\{R_{a+1,b}, Z_{a,b}\}, & \text{if } a \le b \\ -\infty, & \text{otherwise} \end{cases}$$

Then $R_{a,b}$ is a maximum of (5) (and hence of (4)) for fixed a and b. In order to produce an interval $J = [x_i, x_j]$ which maximizes (5) one has to store the indices of the sample points which form the maxima when computing the entries of the matrices Z and R. All together, after computing the matrices Z and R, given $I = [x_a, x_b]$ an interval $J = [x_i, x_j]$ minimizing $\mathrm{dis}_S(I, J)$ can be found in constant time. Hence the total time (including the computation of the L-values and the matrices Z and R) is $O(m^2)$.

For Cases P-1a, P-1b, P-3, and P-4 minimizing the joint disagreement can be shown equivalent to minimizing $L_1((-\infty, x_i)) + L_1((x_j, +\infty))$, $L_1((-\infty, x_i)) + L_1((x_j, +\infty))$, $L_0([x_i, x_j])$, and $L_1((-\infty, x_i)) + L_1((x_j, +\infty))$, respectively. Case P-2b is symmetric to P-2a. In all cases the relevant quantity does only depend on x_i and x_j and one has to obey the corresponding condition on the relative position of x_a, x_b, x_i and x_j. Again by precomputing appropriate data an optimal partner interval for each I can then be found in constant time. Hence the problem of determining whether there is a pair of λ-independent hypotheses can be solved in time $O(m^2)$.

For the case of three hypotheses construct an $n \times n$ matrix M. The entry $M(i, j)$ is 1 if $I_i, I_j \in \mathcal{B}_S$ are λ-independent and 0 otherwise. The matrix M is constructible in time $O(n^2) = O(m^4)$. Now fix $I_i \in \mathcal{B}_S$ and let $\mathcal{K}_i := \{I_j | \mathrm{dis}_S(I_i, I_j) \le \lambda m\}$. Each $\mathcal{K}_i$ is constructible in time $O(n) = O(m^2)$. For each $I_j \in \mathcal{K}_i$ find an $I_k^* \in \mathcal{K}_i$ such that $\mathrm{dis}_S(I_j, I_k^*)$ is minimal. This requires total time $O(m^2)$ using the method for the two interval case (for fixed i). Now if I_i, I_j, I_k^* form an independent set (i.e., if $\mathrm{dis}_S(I_j, I_k^*) \le \lambda m$) we are done. The condition $\mathrm{dis}_S(I_i, I_k^*) \le \lambda m$ is automatically satisfied, because $I_k^* \in \mathcal{K}_i$). On the other hand, if no independent set of size 3 is found by this procedure, then no such set exists. *Q.E.D.*

Remark. With the general approach based on matrix multiplication the problem can be solved in $O(n^{2.37})$. The last result for intervals does not imply that (boolean) matrix multiplication is possible in n^2, because not every matrix can be constructed as a "disagreement matrix" $\mathcal{A}$.

5 Discussion

One might ask whether there are always special solutions to 3-HP which are more efficient than exhaustive search. For example it might be the case that there always exists a specific hypothesis H^* such that the following holds: If there exists a pair (or triple) of independent hypotheses then there is also one such pair or triple containing H^*. A natural choice for H^* is an hypothesis H with minimum disagreement $\mathrm{dis}_S(H)$. The following example shows that this is not correct. We show, that it may happen that the concept which minimizes the disagreement is not a member of any solution for 3-HP.

Consider the class $\mathcal{F}$ of functions $F : \{0,1\}^2 \to \{0,1\}$ for which $F(x_1, x_2) = 1$ for at most one input, i.e. $\mathcal{F} = \{F_0, F_1, F_2, F_3, F_4\}$, where $F_0(x_1, x_2) = 0$, $F_1(x_1, x_2) = \overline{x_1} \wedge \overline{x_2}$, $F_2(x_1, x_2) = x_1 \wedge \overline{x_2}$, $F_3(x_1, x_2) = \overline{x_1} \wedge x_2$, and $F_4(x_1, x_2) = x_1 \wedge x_2$. Let S be the sample of size 12 which consists of the following examples. Each of the possible four instances $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ is contained three times in the sample, once labeled 1 and twice labeled 0. Then $\mathrm{dis}_S(F_0) = 4$ and $\mathrm{dis}_S(F_i) = 5$, $i = 1,2,3,4$. Now, let the maximal tolerated disagreement for a hypothesis be 5 (i.e, $\varepsilon = 5/12$) and the maximal tolerated joint disagreement of two hypothesis be 2 (i.e, $\lambda = 2/12$). Therefore, the disagreement of all hypotheses is tolerated. F_0 minimizes the disagreement, but the joint disagreement of F_0 and F_i for each $i = 1,2,3,4$ is 3. On the other hand, for all $1 \leq i, j \leq 4$ for which $i \neq j$ the joint disagreement is 2.

The above counterexample remains valid even if one allows *all* boolean functions on two variables as hypotheses. The reason is that any function computing a 1 on at least two instances has disagreement at least 6.

References

[1] N. Cesa-Bianchi, E. Dichterman, P. Fischer, and H.U. Simon. Noise-tolerant learning near the information-theoretic bound. In *Proc. 28th Annual ACM Symposium on Theory of Computing, (STOC'96)*. ACM Press, 1996.

[2] N. Cesa-Bianchi, P. Fischer, E. Shamir, and H. U. Simon. Randomized hypotheses and minimum disagreement hypotheses for learning with noise. In *Proc. 3rd EuroCOLT*, 1997.

[3] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. Comput.*, 100(1):78–150, September 1992.

[4] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. In *Proc. 5th Annu. Workshop on Comput. Learning Theory*, pages 341–352. ACM Press, New York, NY, 1992.

[5] Ph. Laird. *Learning from Good and Bad Data.* Kluwer, 1988.