MAXIMAL POSTERIOR PROBABILITY ERROR LOCALIZATION

Gunar E. Liepins, Energy Division, Oak Ridge National Laboratory
David J. Pack, Computer Sciences Division, Union Carbide Corporation

ABSTRACT

Given a data base and a system of inter-variable constraints that acceptable data records are required to satisfy, it is of interest to determine the specific combination of incorrect components in each failing record. The determination of the components is error localization and it is shown that in theory the statistically optimal means of error localization is maximal posterior probability error location (MPPEL). Properties of MPPEL are stated, and connections to minimal fields to impute (MFI), minimal weighted fields to impute (MWFI), and maximal prior probability error localization are made. Simulation results are presented. The tentative conclusion is suggested that for practical application, MPPEL is not significantly superior to MFI.

1. Introduction

Error localization is the process of inferring which components of a multivariate datum record are responsible for the record's failing a given system of constraints. Constraints may be specified in many ways. A system of linear constraints is one such, and could be written for data records x and constraint matrix M as

$$Mx \leq b \tag{1}$$

$$x \geq 0 \tag{2}$$

(This paper will deal exclusively with discrete data so the additional constraint "x_i integer" must be appended. However, many of the results of this paper have straight-forward extensions to the continuous case.) Generally, the violation of constraints does not unambiguously specify which of the components is (are) wrong. For example, given two constraints c^1 and c^2

$$c^1: \; x_1 + x_2 + x_3 \leq 2 \tag{3}$$

$$c^2: \; x_2 + 2x_4 \leq 3 \tag{4}$$

and the datum record $y^0 = (1, 2, 0, 1)$, the record fails both constraints, but it is unclear which components are wrong.

A plausible heuristic for error localization is to specify as wrong, the set of components with smallest cardinality, such that changes in exactly those components will render the record acceptable. This is called the minimum fields to impute (MFI) error localization. Again, with the constraints and datum record of the previous example, a change of y^0_2 from 2 to 0, renders the record acceptable and is the only single component change that will do so.

A generalization of the MFI error localization is the minimum weighted fields to impute (MWFI) error localization: given constants c_i (usually non-negative) and an unacceptable record y^0, find the index set s to minimize

$$\sum c_i \, \delta(\varepsilon_i) \tag{5}$$

$$\text{subject to } y^0 + \varepsilon \text{ is acceptable} \tag{6}$$

$$\delta(\varepsilon_i) = \begin{cases} 0 \text{ if } \varepsilon_i = 0 \\ 1 \text{ if } \varepsilon_i \neq 0 \end{cases} \tag{7}$$

$$\varepsilon_i = 0 \iff i \notin s \tag{8}$$

2. MPPEL

Statistical error localization requires a data model, that is, the modelling of the observed datum record as the true value plus an error term:

$$y = x + \varepsilon \tag{9}$$

Given this model of the data process, error localization could be implemented by specification as wrong, a set of components which has maximal posterior probability of error (MPPEL). Formally, let $y^0 = (y^0_1, \ldots, y^0_n)$ be an unacceptable datum record, S the collection of all 2^n subsets of $\{1, \ldots, n\}$. Then MPPEL can be specified as

$$\underset{s \in S}{\text{maximize}} \sum_{w(s, y^0)} p(x \mid y^0) \tag{10}$$

with $w(s, y^0)$ defined as

$$w(s, y^0) = \{ x \mid x \text{ acceptable and } x_i = y^0_i \text{ for } i \notin s; \; x_i \neq y^0_i \text{ for } i \in s \}. \tag{11}$$

In theory, MPPEL is the optimal error localization procedure in the sense specified by thorem 1.

Theorem 1. For the i^{th} observation, let y^0 be an unacceptable record, $t(y^0, i)$ the true error localization, and L the family of error localization procedures. Let S be the collection of all 2^n subsets of $\{1, \ldots, n\}$ and s and t, elements of S. Let $\delta(s, t) = 1$ if only if $s = t$. Then the expectation

$$\underset{\ell \in L}{E} \{\delta[\ell(y^0), t(y^0, i)]\} \tag{12}$$

is maximized by MPPEL.

Proof. The proof is almost a tautology and proceeds in two steps.

Step 1. MPPEL dominates any constant localization.

Step 2. Any localization is a (stochastic) convex combination of constant localizations.

Theorem 1 states that MPPEL is the optimal localization procedure, but gives no indication of how significantly it dominates other procedures, what its performance is, and how its performance is a function of the data model, and more desirably, how its performance is a function of observables of the data process. Little

is known about these issues, though a few insights seem to be supported by limited simultion results and theoretical investigation.

For further development it is useful to assume independence of errors in components, that is

$$p\{\varepsilon_i = 0 \mid \varepsilon_j \neq 0\} \neq p\{\varepsilon_i \neq 0\} = p_i \text{ for } i \neq j \tag{13}$$

If the additional assumption is made that all the prior probabilities p_i are equal and have value p, then for a given data set and erroneous datum record y^0, the probability of MPPEL correctly localizing the error becomes solely a function of p. The conjecture is that the function is (essentially) convex with local maxima at p = 0 and p = 1. Another conjecture is that MPPEL deteriorates as the number of components increases.

Any attempt to evaluate MPPEL seems to require Bayes' theorum

$$p(x \mid y^0) = \frac{p(y^0 \mid x)\ p(x)}{p(y^0)} \tag{14}$$

Hence, for a given failing record y^0, an equivalent formulation of MPPEL is

$$\text{maximize}_{s \in S} \sum_{w(s,\, y^0)} p(y^0 \mid x)p(x) \tag{15}$$

Without loss of generality, it can be assumed that the data space A is a hypercube, that is A is a cartesian product:

$$A = A_1 \times \ldots \times A_n \tag{16}$$

If in addition, it is assumed that ε_i has uniform distribution over its possible values, then MPPEL can be evaluated up to a constant.

Set $s(y, x) = \{ i: x_i \neq y_i \}$. Then

$$p(y^0 \mid x) = \prod_{s(y^0,x)} \frac{p_i}{(|A_i| - 1)} \prod_{\sim s(y^0,x)} (1-p_i) \tag{17}$$

where $|A_i|$ is the cardinality of the allowable entries in the i^{th} component.
Let $|w(s, y^0)| = \sum p(x)$. [If all the probabilities p (x) are equal, then $|w(s, y^0)|$ is just the cardinality of $w(s, y^0)$.] Then, with the assumptions (13) of independence of errors in components and uniform distribution of component errors, MPPEL can be written as (18)

$$\text{maximize}_{s \in S} \sum_{w(s,y^0)} p(x \mid y^0) = \text{maximize}_{s \in S} \frac{|w(s,y^0)|}{\prod_s (|A_i| - 1)} \prod_s p_i \prod_{\sim s} (1-p_i) \tag{18}$$

3. Maximal Prior Probability Error Localization

Given only the assumption $p(\varepsilon_i \neq 0 \mid \varepsilon_j \neq 0) = p(\varepsilon_i \neq 0) = p_i$, the prior probability that exactly the components indexed by the set s are in error is given by

$$J_s = \prod_s p_i \prod_{\sim s} (1-p_i) \tag{19}$$

whenever s is a feasible localization, and zero otherwise. This can be rewritten as

$$J_s = \prod_{i=1}^{n} (1-p_i) \prod_s p_i / \prod_s (1-p_i) \tag{20}$$

The negative log transform yields a constant plus

$$\sum_s \log (1-p_i) - \sum_s \log p_i \tag{21}$$

Set $c_i = \log (1 - p_i) - \log p_i$. It follows that maximization of J_s, $s \in S$ is equivalent to minimization of

$$\sum_s c_i\, \delta(\varepsilon_i),\ s \in S$$

Thus with appropriate choice of coefficients, MWFI error localization can be interpreted to be maximal prior probability error localization. Moreover, it follows immediately that if

$$\frac{|w(s,y^0)|}{\prod_s (|A_i| - 1)}$$

is constant for feasible localizations {s}, then maximal prior probability error localization is MPPEL. Similarly, if for any index sets t and s related by {t} = {s, k} - that is, t has one more index than s - it holds that

$$\frac{|w(s,y)|}{|w(t,y)|} (|A_k| - 1) \frac{(1-p_k)}{p_k} > 1 \tag{22}$$

whenever s is a feasible error localization, then MPPEL will be a MFI error localization.

4. Evaluation

It must conceded at the outset that MPPEL has proven to be difficult both to analyze theoretically and to evaluate by simulation (a good experimental design is lacking.) Nonetheless, a limited attempt has been initiated.

To gain computational experience with the performance of MPPEL, it was applied to two relatively small simulted examples. In the following, direct performance comparison to minimal fields to impute (MFI) is provided.

The first simulated example is a 6 component example with the following possible variable entries:

$$A_1 = \{0,1\} \qquad A_4 = \{0,1,2,3\} \tag{22}$$

$$A_2 = \{0,1,2\} \qquad A_5 = \{0,1,2\}$$

$$A_3 = \{0,1\} \qquad A_6 = \{0,1,2,3\}$$

The locus of points not acceptable is defined by the explicit edits

$$e_1 = A_1 \times \{0,1\} \times \{0\} \times A_4 \times \{0,1\} \times A_6$$

$$e_2 = \{1\} \times A_2 \times \{1\} \times \{0,1\} \times A_5 \times \{2,3\} \quad (23)$$

$$e_3 = \{0\} \times \{1,2\} \times A_3 \times \{1,2,3\} \times A_5 \times A_6$$

$$e_4 = A_1 \times \{0,2\} \times A_3 \times A_4 \times A_5 \times \{0,1\}$$

$$e_5 = \{1\} \times A_2 \times A_3 \times \{0\} \times \{1,2\} \times A_6$$

These edits imply 422 of the 576 points in $A_1xA_2x...xA_6$ are unacceptable.

The second simulated example is a 4 component example with the following possible variable entries:

$$A_1 = \{0,1,2,...,27,28\} \qquad A_3 = \{0,1\} \quad (24)$$

$$A_2 = \{0,1,2\} \qquad A_4 = \{0,1,2,3,4,5\}$$

This structure was chosen to compare with (22) to examine the effect of having one component (here A_1) with much larger cardinality than the other components. The explicit edits in this case were

$$e_1 = \{1,2,3,...,28\} \times \{0\} \times A_3 \times \{0,1,2\}$$

$$e_2 = \{16,17,...,28\} \times A_2 \times \{1\} \times A_4$$

$$e_3 = \{0,1,2,...,10\} \times \{0\} \times A_3 \times A_4 \quad (25)$$

$$e_4 = \{1,2,3,...,28\} \times \{1,2\} \times A_3 \times \{0\}$$

$$e_5 = A_1 \times A_2 \times \{0\} \times \{3,4,5\}$$

The edits imply 749 of the 1044 points in $A_1xA_2xA_3xA_4$ are unacceptable. The edits were structured so that the proportion 749/1044 would be about the same as the proportion 422/576 in the example of (22) and (23).

The simulation process in a given case produces 10,000 unacceptable records by the following means:

1. Generate a record within the given acceptance region employing a uniform distribution over this region.
2. Perturb the above record with errors generated with specified prior probabilities for each field, independently between components and uniformly within components.
3. Does the perturbed record fail one or more explicit edits? If so, proceed. If not, go to step 1.
4. Localize error according to MWFI and MPPEL.

In the cases reported, specified prior probabilities of error for individual components were all equal. For each example, the probabilities were varied over the values .05, .10, .20, .40, .60, .80, .90, .95. When the prior probability of error is high ("high" depends on the dimension, but generally .60 or more) it is indicative of the actual proportion of error introduced in most components of the 10,000 records. When it is low, however, the actual proportion of error in most components will be greater than this probability (because at least one error must be introduced to each record to put it outside the given acceptance region).

5. Simulation Results

The simulation results for Examples 1 and 2 are given in Tables 1 and 2, respectively. Method 1 is the MFI method (with equal c_i coefficients) and method 2 is the MPPEL. The summary statistics in these tables have the following definitions:

True proportion of error: The proportion of the variables in the 10,000 records that were actually in error. This statistic is clearly related to performance, and it may be quite different from the prior probability of error in each component.

Success index: The proportion of components correctly dealt with over all solutions (both methods of localization can produce multiple solutions for a given record). Components suggested to be in error that were, or components suggested not to be in error that were not are "correctly dealt with".

Matches/solution: The average number of exact matches of localized components to components actually in error per solution over the 10,000 records simulated. (For example, if three alternate solutions are suggested by error localization and one is the true error pattern, then one exact match is added to exact matches and three solutions to solutions.)

The later two statistics are measures of performance, while the first statistic establishes a potential for performance.

Table 1. Simulation Results for Example 1

Summary Statistics	Method	Probability of Error Each Component = p .05	.10	.20	.40	.60	.80	.90	.95
True Prop. Error		.198	.230	.295	.448	.617	.799	.899	.948
Success Index	1	.798	.771	.712	.588	.462	.334	.267	.237
	2	.788	.761	.706	.586	.512	.496	.493	.494
Matches/ Solution	1	.358	.296	.193	.066	.014	.001	.000	.000
	2	.385	.315	.211	.074	.022	.006	.001	.000

Table 2. Simulation Results for Example 2

Summary Statistics	Method	Probability of Error Each Component = p .05	.10	.20	.40	.60	.80	.90	.95
True Prop. Error		.385	.412	.465	.583	.709	.838	.908	.941
Success Index	1	.786	.759	.701	.593	.489	.397	.353	.332
	2	.570	.566	.556	.529	.529	.595	.659	.683
Matches/ Solution	1	.333	.292	.222	.117	.050	.013	.004	.002
	2	.110	.105	.095	.072	.073	.080	.087	.089

Let us first examine the results purely in terms of MPPEL, i.e. method 2. It is interesting to note that this method's performance varies greatly with the general error level in Example 1, but is relatively constant in Example 2. Example 2 may be pathological in that even when p is as low as .05, more than 80% of the simulated records have an error in component 1, the component with relatively large cardinality.

Is performance of MPPEL a convex function of p as conjectured? In Example 1, degraded performance as p increases is reversed only between .90 and .95 in the success index, and is not reversed in the matches/solution. In Example 2, degraded performance is reversed between .40 and .60 in both measures of performance. Thus, performance is suggested to be a convex function of p, but the specific functional form seems to depend significantly on the problem characteristics.

What is the effect of the number of components, which was 6 in Example 1 and 4 in Example 2? One must take care to compare at points where the general error level in the simulation was similar, i.e. "true proportion of error" was similar. This means levels of .385 or more, since .385 was the minimum observed in Example 2. At the lower available levels, one sees little difference. At levels of .8 and more, performance is better in Example 2 with fewer components.

How does the performance of method 1, i.e. MFI, compare to method 2, i.e. MPPEL? Both performance statistics tell about the same story here. In Example 1, MPPEL appears significantly better for $p > .5$, but the two methods look similar for $p \leq .5$. In Example 2, MPPEL appears significantly better for $p > .5$, but significantly worse for $p \leq .5$. Overall, since the realistic range of $\bar{p}$ values in a real problem is surely $p \leq .5$, the simulation results seem to suggest that MFI does not appear to perform significantly worse than MPPEL and should receive serious consideration as a means of error localization.

In examining individual records in some of the simulations, it was noted that MPPEL often localized error to one component whereas there were frequently two or more components in error. This pattern is one that might be expeted if the prior probabilities p_i are generally lower than the true probabilities of error. In effect, that is what happens. Records outside the given acceptance region contain more error than is implied by the p level in Tables 1 and 2, at least for the lower p levels. Multiple errors also occur more frequently than the low p levels might suggest.

Conclusions. With appropriate assumptions MPPEL is equivalent to maximal prior probability error localization, and with slightly weaker assumptions, equivalent to MFI. In theory, MPPEL is statistically the optimal error localization procedure, but the examples simulated suggest that for practical problems, MPPEL does not perform significantly better than MFI. A good experimental design is required but not presently available to thoroughly test the various localization techniques.

Reference

I. P. Fellegi and D. Holt (1976), "A Systematic Approach to Automatic Edit and Imputation," Journal of the American Statistical Association, 71, 17-35.

Research sponsored by the Energy Information Administration, U.S. Department of Energy, under contract W-7405-eng-26 with Union Carbide Corporation.