Spectrum-Based Sequential Diagnosis
(Extended Abstract)

Alberto Gonzalez-Sanchez[1], **Rui Abreu**[2], **Hans-Gerhard Gross**[1], **and Arjan J.C. van Gemund**[1]

[1] *Department of Software Technology, Delft University of Technology, The Netherlands*
{a.gonzalezsanchez, h.g.gross, a.j.c.vangemund}@tudelft.nl
[2] *Department of Informatics Engineering, University of Porto, Portugal*
rui@computer.org

In fault diagnosis an important performance metric is the accuracy of the multiple-fault diagnosis D returned by a diagnostic algorithm (DA). In many cases diagnostic accuracy is insufficient (D comprises too many candidates). Diagnostic uncertainty can be reduced by supplying more modeling information, and more observation information. Observation information can be in spatial sense (more probing points), and in temporal sense (more tests), the latter of which is the focus of this paper. Ignoring the computational cost of the DA itself, the diagnostic process may be viewed as (1) expending costs to generate observation input to the DA (testing cost C_t), and (2) expending costs to find the actual diagnosis from D returned by the DA by the diagnostician (diagnostic cost C_d). The latter is inverse to diagnostic accuracy (utility) as false positives and negatives in D result in wasted effort by the diagnostician finding the actually faulted components.

In *sequential diagnosis* (SD) one computes the optimal sequence of tests (observations) that, on average, minimizes C_t and C_d. SD approaches are typically found in the (hardware) systems testing domain where system knowledge is encoded in terms of predetermined test matrices, or in the model-based diagnosis domain where system models are available. In the software domain, the domain considered in this paper, the problem of diagnosis has received considerable attention. Due to complexity problems most model-based approaches are limited to small programs and/or must take a single-fault assumption. Typically, statistical approaches are used, that abstract from the program code and compute the correlation between component involvement in tests on the one hand (expressed in terms of a so-called *spectra*), and the test outcomes (pass/fail) on the other, deriving a ranking of suspect components. While these approaches essentially involve multiple observations (e.g., an entire regression test suite) they do not qualify as SD since the choice of tests is random (a *passive* diagnosis approach), rather than generating/selecting the next best test based on the diagnosis obtained thus far (an *active* diagnosis approach). Hence, the decrease of C_d per test is far from optimal.

In this paper we present a spectrum-based SD approach and an associated algorithm dubbed SEQUOIA (SEQUencing fOr dIAgnosis) that greedily computes a sequence out of a large set of given tests that delivers near-optimal diagnostic performance in terms of the decay of C_d as function of the number of tests. Unlike statistical techniques, which have insufficient diagnostic accuracy to select the next best test, we use an approximate, Bayesian reasoning approach which provides the required accuracy. Similar to approaches based on a test coverage matrix, the tests are selected from a fixed test set. Unlike multiple-fault, matrix-based approaches, however, our approximate, Baysian approach can handle large system sizes and test set sizes. Furthermore, unlike MBD approaches, no system information is required, other than component test coverage, which is typically known from test execution profiling.

To achieve polynomial complexity SEQUOIA uses an approximate information gain computation approach. Synthetic data shows, that the dynamic selection of the next best test based on the observations made so far, allows SEQUOIA to achieve much better decay of diagnostic uncertainty compared to random test sequencing. Real programs, taken from the Siemens set, also show that SEQUOIA has better performance, except when the diagnosis involves ambiguity sets that are too large for the entropy estimation to handle. Future work therefore includes solving this problem by a better expansion technique, expanding the MHS in order of posterior probability (which is currently ignored in our prototype approach).