Using Dissimilarity Metrics to Identify Interesting Designs

Martin S. Feather
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Dr
Pasadena CA 91109-8099
Martin.S.Feather@Jpl.Nasa.Gov

James D. Kiper Selcuk Kalafat
Dept. of Computer Science and Syst. Analysis
Miami University
Oxford, OH 45056
kiperjd@muohio.edu,
kalafaas@muohio.edu@muohio.edu

Abstract

Finding a preferred solution to a complex design problem is challenging. On the one hand the problem space is too large and convoluted for human comprehension, while on the other hand it is infeasible to elicit the entirety of design knowledge required for fully automatic problem solving. We face this challenge repeatedly when planning the development of technologies for spacecraft applications. In this context numerous risk abatement options give rise to a huge space of potential design solutions.

We report on application of a combination of data mining steps guided by user-directives, and appropriate visualizations to present the results of these steps to the users. The core ideas are use of heuristic search to locate near-optimal design solutions, and user-defined "dissimilarity" metrics to characterize interestingly distinct design solutions. The search, data mining and visualization capabilities are implemented features of the risk management tool we use to support a risk-centric design methodology developed and applied at JPL and NASA.

This approach is demonstrated on the selection of risk abatement solutions in the design of advanced technology. The study was performed at JPL to plan the development of technology for future spaceflight missions.

1. Introduction

Risk concerns play a prominent role in planning the design and development of complex systems. Typically there are numerous risk abatement options (e.g., analyses, tests) and decision alternatives (e.g., architecture choices, implementation alternatives) with significant risk implications. Their costs (time, budget, etc.) and benefits (their effectiveness at risk abatement) have to be taken into account when selecting among them.

For complex systems, the selection space of risk abatement solutions can be very large. While the number of individual risk abatement options may be modest, the choice of selections from among them is large (e.g., given 50 binary options, there are $2^{50} \cong 10^{15}$ ways of selecting from among them). We use data mining and visualization to assist experts to make these selections.

The main steps of our approach are sketched in Figure 1. An engineering design model is developed based on inputs from technical experts. Heuristic search is used to reveal the cost/benefit trade space implied by this model. Experts identify the design neighborhood in which they are most interested. The space is culled to just the solutions within that neighborhood. Experts provide "dissimilarity metrics" that indicate what they consider to be important distinctions among neighboring solutions. Dispersal and clustering algorithms are used to locate a modest number of distinct designs from within that neighborhood. Custom visualization presents the located designs to the experts, allowing them to make their choice

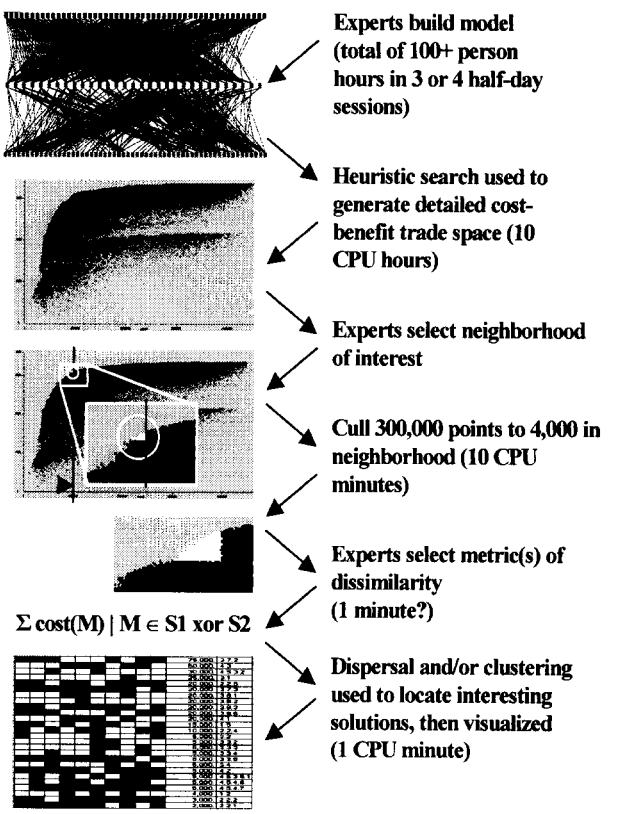

Figure 1. Overall approach

of preferred design.

Throughout this paper we use data from one of the technology assessment and infusion planning efforts performed at JPL. The details are proprietary, so we avoid revealing specifics. Nevertheless, all the quantities we report (e.g., the number of distinct mitigations) are actual figures from the assessment effort, and charts shown have all been generated from this data. Briefly, the assessment concerned an electronics packaging technique that has seen wide use on Earth, and on some space missions but only inside a temperature controlled housing. The focus of the assessment was its novel application to settings where the electronics would be exposed to the harsh conditions (e.g., extreme cold temperature) of planetary environments. The end goal was to identify and select appropriate design, fabrication, assembly and testing methods for the packaging technique so that it could be incorporated reliably into future spacecraft.

2. A Risk-Informed Design Methodology

A risk-informed design methodology underpins our work. Motivation for this stems from a vision of using risk as a *resource*, one that can be traded against other resources such as schedule, cost and performance [Greenfield, 1998]. The methodology we have developed and applied at JPL and NASA supports this vision. It combines insights and skills of spacecraft experts, a model for representing their knowledge, a process for building and exploring the model, and custom software to support this process.

At its heart, it relies on users to identify: *objectives* to be achieved (and their relative priorities), the various *risks* to achieving those objectives, and options for risk *mitigation* (prevention, detection ahead of time, and alleviation). The connectivity among these pieces of information is as follows: risks are connected to the objectives that they would impact (should those risks occur), and mitigations are connected to the risks they reduce (should those mitigations be applied). Note that a risk may impact multiple objectives, an objective may be impacted by multiple risks, etc. Note also that different risk impacts and mitigation effects may have different strengths, for example, one risk may detract from one objective more than it does from another.

Models of actual technologies and systems are typically voluminous and convoluted, as illustrated by the data in Figure 2 extracted from the technology study used throughout this paper. This comprises 50 objectives, 31 risks, 58 mitigations, and some 800 links among them, numbers typical of the order of magnitude of data gathered in these assessment efforts.

This data was gathered from experts in a series of facilitator-led sessions, following the elicitation process we have established for our risk-centric models.

The risk-centric design model offers a selection of risk mitigations. For any given selection, the model can be evaluated to yield two measures:

- *cost*, calculated as the sum of the costs of the selected mitigations (e.g., the cost to perform a test), and of the repairs of risks they detect (e.g., the cost of fixing the bugs revealed by testing), and
- *benefit*, calculated as the sum total attainment of objectives taking risks into account. Risks detract from objectives' attainment, however risks themselves are reduced (in likelihood and/or impact) by the selected mitigations. Risk reduction leads to increased attainment of objectives.

For further details of our model, and how it is applied, see [Feather&Cornford, 2003].

3. Optimal Designs and Heuristic Search

An optimal design is one that attains its objectives at minimal expense. Generally speaking mitigations increase expected objective attainment (by reducing risks), but incur costs. In most instances the total cost of all possible mitigations far exceeds the resources available. The primary purpose of our methodology is to help identify and select the set of mitigations to apply to achieve an optimal design within some cost bound.

In typical designs there will be many mitigations (dozens, possibly hundreds). The combinatorial choices from among these imply a space containing huge numbers of possible candidate solutions. In our application, 58 mitigations represented design and development choices whose costs range from the low thousands of dollars to, in a few cases, hundreds of thousands of dollars. Since there are 58 mitigations, there are in principle 2^{58} (approximately 10^{17}) different selections from among them.

We use simulated annealing (a form of heuristic search), included as part of our software, to locate near-optimal solutions. We have also explored genetic algorithms, and machine learning [Cornford et al, 2003]. On our electronics technology dataset, a detailed search was performed, organized as a series of individual

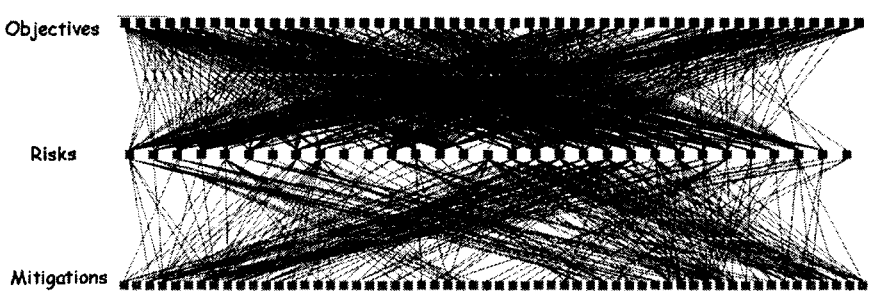

Figure 2. Topology of data in a completed risk model

cost-bounded optimal searches at successive cost levels. The resulting cost-benefit trade space is the background portion of Figure 3. The sum total cost of all mitigations (approximately $4,750,000) determines the rightmost value of the x-axis, and the sum total value of all objectives (approximately 3,600) determines the topmost value of the y-axis. Its generation took on the order of 10 hours running on a 1.8 GHz PC.

Each of the approximately 300,000 individual points in the black "cloud" corresponds to a design solution (i.e., selection of mitigations). For a given solution, the software uses the quantitative risk-centric model to calculate cost and benefit. A small black point corresponding to the solution is then drawn on the plot – solution cost determines horizontal position, solution benefit vertical position. The upper-left frontier of the cloud is thus the "optimal" boundary, also referred to as the "Pareto front" [Sen&Yang, 1998]. Note that we plot a point for *every* solution investigated by the search, not just the "near-optimal" solution points on the boundary. This is important data, since the steps that follow investigate points close to, but not necessarily on, that optimal frontier. The simulated annealing search is designed to concentrate towards this optimal boundary.

4. Experts Select Neighborhood of Interest

The experts know of funding availability, and level of benefit (attainment of objectives) desired. They use this knowledge to identify their neighborhood(s) of interest within the cost/benefit trade space revealed by the previous search.

The reasons we identify a *neighborhood* of interest, rather than simply picking specific near-optimal solutions on the frontier are twofold. First, the data over which the search is performed was produced through expert judgment. We do not assume all of those expert judgments to be perfectly accurate. That is, solutions within a small percentage of the near optimal solution may in truth be no more costly and/or attain no less benefit than ones calculated as the "near-optimal" in that neighborhood. Thus such a neighborhood encompasses different design solutions that, from the standpoint of cost and benefit, we judge to be equally acceptable within the accuracy of our data. Second, there may be factors other than overall cost and benefit that would lead us to prefer one solution within this neighborhood over another. For example, one design solution in the neighborhood may make use of a hard-to-schedule test facility, while another does not, in which case we might prefer the latter.

While it should in principle be possible to encode such preferences within the utility function that guides the optimal search, we believe this to be infeasible in practice. In the first place, much of this may be "tacit" knowledge – not evident until the experts see concrete examples of design solutions. Even if the experts could foresee all these preferences in advance, our suspicion is that it would be a waste of time to ask them to try to articulate them all. Better, have them list just the ones that they recognize as

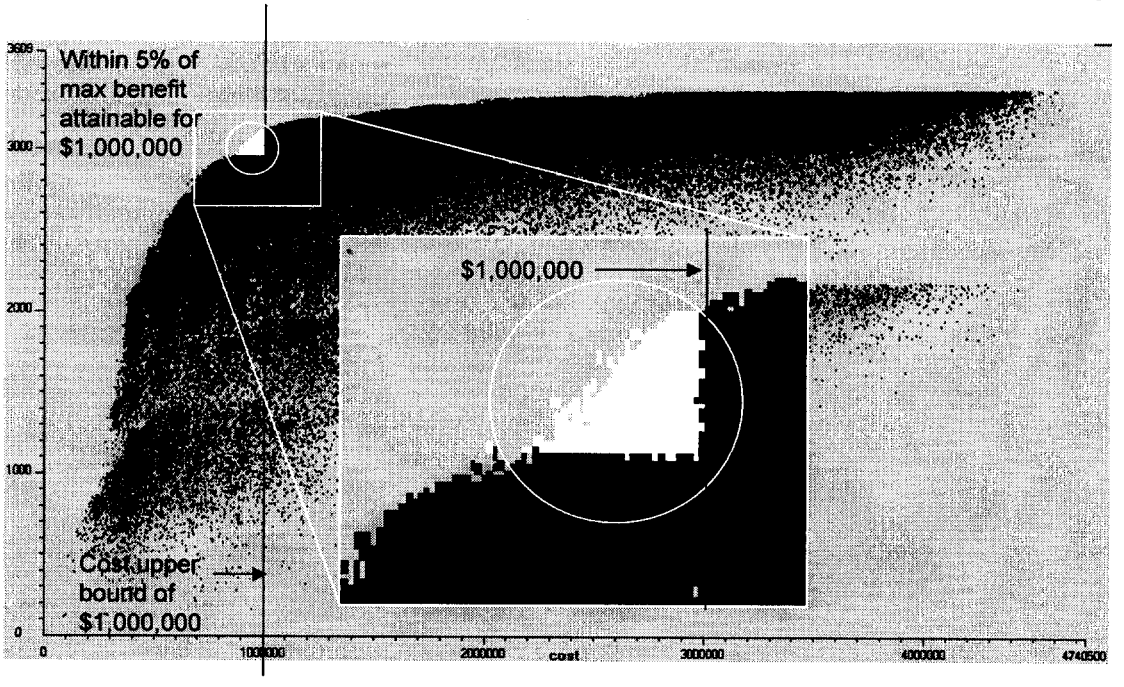

Figure 3. Selection of a neighborhood of interest

the major impediments, perform the search, identify the neighborhood, scrutinize the results, and select the preferred solution.

An expert-identified neighborhood of interest defined as solutions costing ≤ $1,000,000 and attaining ≥ 95% of the objectives attainment of the best solution found at or below that cost limit is shown superimposed and magnified in Figure 3. For the dataset, 3,391 solutions (i.e., distinct selections of mitigations) fall within the neighborhood.

5. Dissimilarity Metrics

Manual scrutiny of each of 3,391 solutions is tedious at best. Since many will be very similar (differing by a small number of the selected mitigations), we use data mining to more effectively explore this neighborhood. We define a ***metric of dissimilarity*** – two design solutions (each a set of selected mitigations) will have a larger dissimilarity value according to this the more they differ from one another. Using this metric, solutions that lie within the neighborhood of interest, but which are distinct from one another with respect to this dissimilarity metric, are located and presented to the users.

In our risk-centric model, a design solution is described by a set of mitigations, each of which is either "on" or "off". A dissimilarity metric takes as input a pair of design solutions, and returns a numerical measure of dissimilarity.

A simple metric is to count the number of differences between the two design solutions' mitigations, that is, count how many mitigations are "on" in one but not in the other of those solutions. If bit strings represent solutions, each bit corresponding to a mitigation, with value 1 if that mitigation is "on", 0 otherwise, then this metric is the *Hamming distance* between solutions' representations.

Another metric is to sum the costs of all the mitigations that are "on" in one but not both of two solutions. This metric therefore ranks as more dissimilar solutions that differ by higher cost mitigations.

Yet another metric makes use of relevant groupings of mitigations. For example, in our technology assessment studies, it is common to classify mitigations according to the phase in which they would apply – design, fabrication, assembly, test. We can define a metric of dissimilarity based on the difference in costs *between* these phases.

The experts using our software select which metrics to use to explore the neighborhood of design solutions.

6. Data-Mining for Interesting Design Solutions

Given a neighborhood of solutions, and a dissimilarity metric chosen by the experts, we use two data mining techniques to locate solutions distinct with respect to the chosen metric:

- *Dispersal:* a set of maximally dispersed (using the metric of "dissimilar") solutions is extracted from the neighborhood of interest.
- *Clustering:* similar solutions, i.e., solutions that are close in distance according to the metric of "dissimilar", are aggregated into clusters.

6.1. Data-Mining by Dispersal

We have implemented a dispersal method that is a fast approximation of an idealized dispersal algorithm. It works as follows:

Input C (> 1) the number of dispersed solutions that the method is to find. Let N be the set of design solutions in the neighborhood of interest.:

Initialize S to be the singleton set holding the optimal design solution in N;

While S's cardinality < C, do:

- Find a design solution *ds* in (N – S) such that *ds*'s minimum distance from all the design solutions in S is as great as possible.
- Add *ds* to S.

This approximate dispersal algorithm was applied to find 10 dispersed solutions with respect to each of the three metrics discussed above from the expert-identified neighborhood. It took under 2 minutes to find and plot the visualization of the 30 dispersed solutions – 10 dispersed solutions for each of the three metrics. The table below shows the dispersal distances between solution point s_1 and its shortest metric distance to the points $s_2 \ldots s_n$ where n = card(N)

Table 1: Dispersal Distances

	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
Hamming distance	23	20	18	18	18	17	17	16	16
Cost of different mitigations ($1,000's)	65	57	50	45	43	40	37	36	36
Cost difference of phases ($100,000's)	32	23	21	17	15	15	13	12	11

6.2. Data-Mining by Clustering

The intuition behind this method is that we partition similar solutions into groups (clusters). Resulting from this will be clusters such that the solutions within a cluster will all be similar to one another (hence if we examine one, there is little point to examining another solution drawn from that same cluster), while the clusters themselves will tend to represent distinct design solutions.

The clustering method works by first picking a prototype that will serve as a seed for the first cluster. (We have arbitrarily picked the near-optimal solution found by the search to be this initial seed.) Then, each design solution in the neighborhood of interest is

compared to this seed. Design solutions are compared to the seed solution by means of the dissimilarity metric of our choosing (e.g., one of those discussed in the preceding section). If the measure it returns is sufficiently small (below some threshold value), this solution set is included in the cluster.

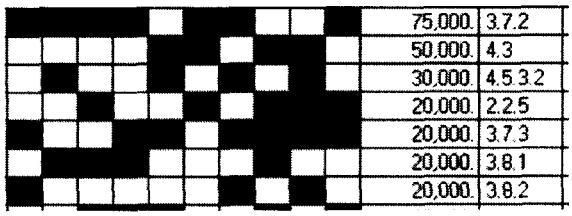

Figure 4. Fragment of visualization of dispersal

7. Visualization of the Data-Mining Results

Visualization is used to present results of dispersal and clustering to the experts, allowing them to see what distinct solutions are available, and understand the ways in which those solutions differ (and the ways in which they do *not* differ).

7.1. Visualization of Dispersal Results

We use a simple tabular display as the means to visualize dispersal results. Figure 4 shows a fragment of such a display, and Figure 5 gives a key to its elements. In practice the visualization would list the title of each of the migitigations, but in order to protect the proprietary nature of this information, such titles have been omitted here.

The display shows dispersal results using the metric of the sum of the costs of all the mitigations that are "on" in one but not both of two solutions. Since the focus is on how solutions differ, mitigations that are either always on in all of the dispersal solutions, or always off, are omitted, hence every row in the displayed results involves at least one black square (mitigation included) and one white square (mitigation not included).

The mitigations are arranged in order of decreasing cost. From the visualization we can readily discern alternative solutions some of which make use of the relatively expensive mitigation 3.7.2, and some of which do not. This kind of

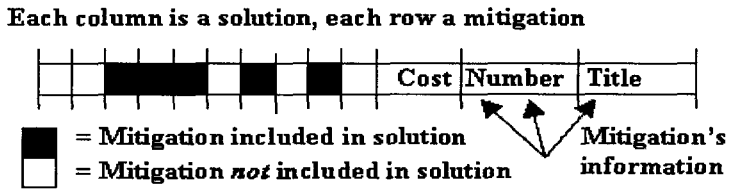

Figure 5. Key to visualization of dispersal results

visualization aids the experts in making their selection of the preferred solution.

7.2. Visualization of Clustering Results

Figure 6 shows a fragment of the visualization of the clustering results from our data, and Figure 7 gives a key to its elements. As before, the names of mitigations would normally be listed alongside their numbers, but are omitted in this paper to conceal proprietary information.

Each cluster is represented by one column of boxes. The shading of the boxes indicates the proportion of solution sets of that cluster that include that particular mitigation. The darker the shading, the greater the percentage of solutions that include the mitigation.

Using this display, the experts can quickly determine which clusters are preferred. When the best cluster is determined, clicking on that cluster produces another graphic display of all of the solutions in that cluster so that the manager may fine-tune his/her choices.

8. Conclusions

This paper has described application of data mining and visualization to aid experts in making design

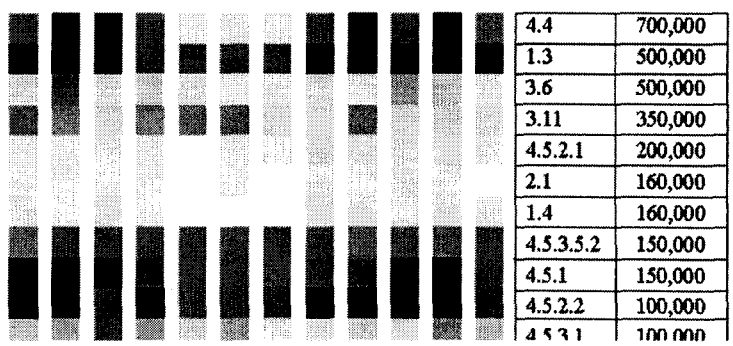

Figure 6. Fragment of visualization of clustering results

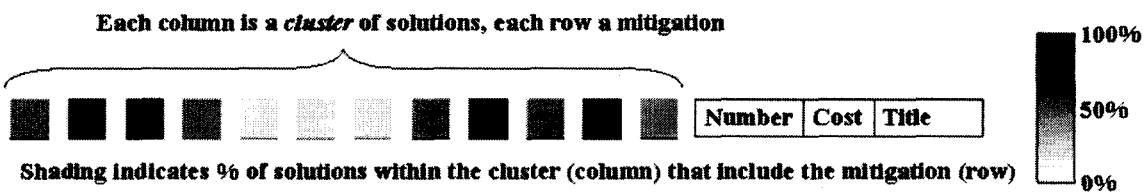

Figure 7. Key to visualization of clustering results

decisions. The approach is demonstrated on the selection of risk abatement solutions in the design of advanced technology. The study was performed at JPL to plan the development of technology for future spacecraft missions.

Our use of dissimilarity metrics bears a resemblance to the "unexpectedness" measures of [Padmanabhan & Tuzhilin, 1999], but is perhaps simpler because we seek only to locate unexpected data, not patterns in that data.

In future work we plan to investigate a closer connection between the dispersal and clustering techniques. In particular, we plan to use the former to rapidly get an overall feel for how solutions are dispersed, and use that information to guide slower but more revealing clustering algorithms. We also would like to take into account knowledge of uncertainty distributions in the input data to help better identify neighborhoods of interest.

More details of our risk-informed methodology, and the tool support we have built to support it, can be found at http://ddptool.jpl.nasa.gov

9. Acknowledgements

The research described in this paper was done at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology. We thank Chet Borden for insightful comments on this presentation.

10. References

[Cornford et al, 2003] S.L. Cornford, M.S. Feather, J.R. Dunphy, J. Salcedo & T. Menzies. "Optimizing Spacecraft Design – Optimization Engine Development: Progress and Plans", 2003 IEEE Aerospace Conference, Big Sky, Montana, March 2003.

[Feather & Cornford, 2003] M.S. Feather, & S.L. Cornford. "Quantitative risk-based requirements reasoning", to appear in *Requirements Engineering* (Springer), published online 25 February 2003, DOI 10.1007/s00766-002-0160-y. Available from: http://eis.jpl.nasa.gov/~mfeather/AvailablePublications/

[Padmanabnah & Tuzhlin, 1999] B. Padmanabhan & A. Tuzhilin. "Unexpectedness as a measure of interestingness in knowledge discovery." Decision Support Systems Journal, 27, 3, 303-318.

[Sen&Yang, 1998] P. Sen & J-B. Yang. *Multiple Criteria Decision Support in Engineering Design*, Springer-Verlag, 1998.