

Case-Based Reasoning for Reducing Software Development Effort

Adam Brady¹, Tim Menzies¹, Oussama El-Rawas¹, Ekrem Kocaguneli¹, Jacky W. Keung²

¹Lane Department of CS&EE, West Virginia University, Morgantown, USA; ²Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China.

Email: adam.m.brady@gmail.com, tim@menzies.us, orawas@gmail.com, kocaguneli@gmail.com, jacky.keung@comp.polyu.edu.hk

Received June 24th, 2010; revised July 15th, 2010; accepted July 19th, 2010.

ABSTRACT

How can we best find project changes that most improve project estimates? Prior solutions to this problem required the use of standard software process models that may not be relevant to some new project. Also, those prior solutions suffered from limited verification (the only way to assess the results of those studies was to run the recommendations back through the standard process models). Combining case-based reasoning and contrast set learning, the W system requires no underlying model. Hence, it is widely applicable (since there is no need for data to conform to some software process models). Also, W's results can be verified (using holdout sets). For example, in the experiments reported here, W found changes to projects that greatly reduced estimate median and variance by up to 95% and 83% (respectively).

Keywords: *Software Effort Estimation, Case Based Reasoning, Effort Estimation*

1. Introduction

Existing research in effort estimations focuses mostly on deriving estimates from past project data using (e.g.) parametric models [1] or case-based reasoning (CBR) [2] or genetic algorithms [3]. That research is curiously silent on how to change a project in order to, say, reduce development effort. That is, that research reports what is and not what should be changed.

Previously [4-7], we have tackled this problem using STAR/NOVA, a suite of AI search algorithms that explored the input space of standard software process models to find project options that most reduced the effort estimates. That approach had some drawbacks including 1) the dependency of the data to be in the format of the standard process models, 2) the implementation complexity of the Monte Carlo simulator and the AI search engines, and 3) the lack of an independent verification module (the only way to assess the results of those studies was to run the recommendations back through the standard process models).

This paper describes “W”, a simpler, yet more general solution to the problem of finding project changes that most improves project estimates. Combining case-based reasoning and contrast set learning, W requires no underlying parametric model. Hence, W can be applied to

more data sets since there is no requirement for the data sets to be in the format required by the standard software process models. For example, this paper explores five data sets with W, three of which cannot be processed using STAR/NOVA.

W is simpler to implement and easier to use than STAR/NOVA, requiring hundreds of lines of the scripting language AWK rather than thousands of lines of C++/LISP. For example, we describe below a case study that finds a loophole in Brooks' Law (“adding manpower [sic] to a late project makes it later”). Using W, that study took three days which included the time to build W, from scratch. An analogous study, based on state-of-the-art model-based AI techniques, took two years.

Also, the accuracy of STAR/NOVA is only as good as the underlying model (the USC COCOMO suite [1]). W's results, on the other hand, are verified using hold-out sets on real-world data. In the experiments reported below, we show that W can find changes to projects that greatly reduce estimate median and variance by up to 95% and 83%, respectively.

Finally, the difference between W, which finds what to change in order to improve an estimate, and a standard case-based effort estimator, which only generates estimates, is very small. Based on this experiment, we

advise augmenting standard CBR tools with modules like the planning sub-systems in W.

The rest of this paper is structured as follows. After some background notes on effort estimation and STAR/NOVA, we describe the general framework for case-based reasoning. The W extension to CBR is then described (contrast set learning over the local neighborhood), using a small example. This is followed by fourteen case studies, one with Brooks' Law, and thirteen others. Our conclusion will discuss when W is preferred over STAR/NOVA.

2. Background

2.1. Why Study Effort Estimation?

Generating (and regenerating) project effort estimates is an important and continuous process for project managers. Not only do good estimates allow the better use of resources, but by reviewing and improving their estimation process, a software company can learn and improve from their past experience.

Sadly, we often get estimates wrong. Consider the NASA's Checkout Launch Control System, which was canceled when the initial estimate of \$200 million was overrun by an additional \$200 M [8]. This case is not unique, despite the significant effort put into designing more accurate estimation models. It has been reported that many predictions are wrong by a factor of four or more [9,10]. In order to conduct software effort estimation, it is standard practice to use models to estimate effort. Many software process models have emerged aiming to achieve that task, and there has not emerged a single standardized model that is widely used by the software engineering industry. There several reasons for this including generality, data islands and instability. Software models may not be general so it can be inappropriate to apply a software model learned in one environment one to another. Also, many companies prefer to keep cost related data confidential. This data island effect has also contributed to the fragmentation of the field by companies preferring to build private models rather than using publicly available models. This multiplicity of software effort models has lead to scarcity of publicly available, local data needed for model based effort estimation. Without sufficient data to build, audit, and tune models, the predictions generated by these models may be highly unstable. Baker [11] reports a study that learned values for the (a, b) (linear, exponential) constants in Boehm's COCOMO software process model [9]. The study was repeated 100 times, each time selecting from a random sample of 90% of the project data. The learned values for (a, b) exhibited an alarming variance:

$$(2.2 \leq a \leq 9.18) \wedge (0.88 \leq b \leq 1.09) \quad (1)$$

Such large variations make it hard to understand the effects of changing project options. Suppose some proposed change doubles productivity, but a moves from 9 to 4.5. The improvement resulting from that change would be obscured by the tuning variance.

2.2. STAR and NOVA

The SBSE literature inspired us to try simulated annealing to search the what-ifs associated with Equation 1. This lead to the STAR system [4,6]. NOVA was a generalization of STAR that included simulated annealing and other search engines [5,7].

STAR/NOVA handled model variance by finding conclusions that were stable across the space of possible tunings. This analysis assumed that, for a mature effort estimation model, the range of possible tunings was known (this is the case for models like COCOMO). For such models, it is possible for the AI search engines to find conclusions that hold across the space of all tunings.

STAR and NOVA constrain project options P but not the tuning options T. Hence, their recommendations contain subsets of the project options P that most improve the score, despite variations in the tunings T. This approach meant we could reuse COCOMO models requiring using local tuning data. The following is a description that briefly presents the operation of STAR and NOVA:

1) *SAMPLE*: We sample across the ranges of all the attributes in the model n_1 times. Often we sample randomly across the range. Some heuristics allow us to concentrate more on the extremes of the range.

2) *DISCRETIZE*: The data seen in the n_1 samples is then discretized into $D = 10$ bins. Equal frequency bins were used.

3) *CLASSIFY*: The top $n\%$ projects are classified as best or rest.

4) *RANK*: The ranges are then ranked in increasing order using Support-Based Bayesian Ranking.

5) *PRUNE*: STAR runs n_2 experiments with the models where the top ranked ranges $1 \dots X$ ranges are pre-set and the remaining ranges can be selected at random.

6) *REPORT*: STAR returns the $1 \dots X$ settings that optimize the best for the fitness function. These settings are determined by iterating back from the minimum point achieved towards the first point that is statistically similar to the minimum point.

In practice, STAR/NOVA approach was very effective. **Figure 1** shows the large effort reductions found by STAR in three out of four cases presented at ASE'09. It is insightful to reflect about STAR/NOVA's failure to find large reductions in the fourth case study (nasa93 osp2). In this project, management had already fixed most of the project options. STAR/NOVA failed, in this case, since there was very little left to try and change.

Study	NOVA
nasa93 flight	72%
nasa93 ground	73%
nasa93 osp	42%
nasa93 osp2	5%

Figure 1. Effort estimate improvements found by NOVA from [12].

This fourth case study led to one of the lessons learned of STAR/NOVA: apply project option exploration tools as early as possible in the lifecycle of a project. Or, to say that more succinctly: if you fix everything, there is nothing left to fix [13].

While a successful prototype, STAR/NOVA has certain drawbacks:

- **Model dependency:** STAR/NOVA requires a model to calculate (e.g.) estimated effort. In order to do so, we had to use some software process models to generate the estimates. Hence, the conclusions reached by STAR/NOVA are only as good as this model. That is, if a client doubts the relevance of those models, then the conclusions will also be doubted.
- **Data Dependency:** STAR/NOVA's AI algorithms explored an underlying software process model. Hence, it could only process project data in a format compatible with the underlying model. In practice, this limits the scope of the tool.
- **Inflexibility:** It proved to be trickier than we thought to code up the process models, in a manner suitable for Monte Carlo simulation. By our count, STAR/NOVA's models required 22 design decisions to handle certain special cases. Lacking guidance from the literature, we just had to apply "engineering judgment" to make those decisions. While we think we made the right decisions, we cannot rigorously justify them.
- **Performance:** Our stochastic approach conducted several tens of thousands of iterations to explore the search space, with several effort estimates needing calculation with each iteration. This resulted in a performance disadvantage.
- **Size and Maintainability:** Due to all the above factors, our code base proved difficult to maintain. While there was nothing in principle against applying our techniques to other software effort models, we believe that the limiting factor on disseminating our technique is the complexity of our implementation. As partial evidence for this, we note that in the three years since we first reported our technique [6]:
- We have only coded one set of software process models (COCOMO), which inherently limited the scope of our study.

- No other research group has applied these techniques. Therefore, rather than elaborate a complex code base, we now explore a different option, based on Case Based Reasoning (CBR). This new approach had no model restrictions (since it does not use a model) and can accommodate a wide range of data sets (since there are no restrictions of the variables that can be processed).

While there was nothing in principle against applying our techniques to other software effort models, we believe that the limiting factor on disseminating our technique is the complexity of our implementation. As partial evidence for this, we note that in the three years since we first reported our technique [6]:

- We have only coded one set of software process models (COCOMO), which inherently limited the scope of our study.
- No other research group has applied these techniques.

Therefore, rather than elaborate a complex code base, we now explore a different option, based on Case Based Reasoning (CBR). This new approach had no model restrictions (since it does not use a model) and can accommodate a wide range of data sets (since there are no restrictions of the variables that can be processed).

3. Case-Based Reasoning (CBR)

Case based reasoning is a method of machine learning that seeks to emulate human recollection and adaptation of past experiences in order to find solutions to current problems. That is, as humans we tend to base our decisions not on complex reductive analysis, but on an instantaneous survey of past experiences [14]; *i.e.*, we don't think, we remember. CBR is purely based on this direct adaptation of previous cases based on the similarity of those cases with the current situation. Having said that, a CBR based system has no dedicated world model logic, rather that model is expressed through the available past cases in the case cache. This cache is continuously updated and appended with additional cases.

Aamodt & Plaza [15] describe a 4-step general CBR cycle, which consists of:

- 1) *Retrieve*: Find the most similar cases to the target problem.
- 2) *Reuse*: Adapt our actions conducted for the past cases to solve the new problem.
- 3) *Revise*: Revise the proposed solution for the new problem and verify it against the case base.
- 4) *Retain*: Retain the parts of current experience in the case base for future problem solving.

Having verified the results from our chosen adapted action on the new case, the new case is added to the available case base. The last step allows CBR to effectively

learn from new experiences. In this manner, a CBR system is able to automatically maintain itself. As discussed below, W supports retrieve, reuse, and revise (as well as retain if the user collecting data so decides).

This 4-stage cyclical CBR process is sometimes referred to as the R4 model [16]. Shepperd [16] considered the new problem as a case that comprises two parts. There is a description part and a solution part forming the basic data structure of the system. The description part is normally a vector of features that describe the case state at the point at which the problem is posed. The solution part describes the solution for the specific problem (the problem description part).

The similarity between the target case and each case in the case base is determined by a similarity measure. Different methods of measuring similarity have been proposed for different measurement contexts. A similarity measure is measuring the closeness or the distance between two objects in an n -dimensional Euclidean space, the result is usually presented in a distance matrix (similarity matrix) identifying the similarity among all cases in the dataset. Although there are other different distance metrics available for different purposes, the Euclidean distance metric is probably the most commonly used in CBR for its distance measures.

Irrespective of the similarity measure used, the objective is to rank similar cases from case-base to the target case and utilize the known solution of the nearest k -cases. The value of k in this case has been the subject of debate [17,2]: Shepperd [2], Mendes [18] argue for $k = 3$ while Li [3] propose $k = 5$.

Once the actual value of the target case is available it can be reviewed and retained in the case-base for future reference. Stored cases must be maintained over time to prevent information irrelevancy and inconsistency. This is a typical case of incremental learning in an organization utilizing the techniques of CBR.

Observe that these 4 general CBR application steps (retrieve, reuse, revise, retain) do not include any explicit model based calculations; rather we are relying on our past experience, expressed through the case base, to estimate any model calculations based on the similarity to the cases being used. This has two advantages:

1) It allows us to operate independently of the models being used. For example, our prior report to this conference [13] ran over two data sets. This study, based on CBR, uses twice as many data sets.

2) This improves our performance, since data retrieval can be more efficient than calculation, especially given that many thousands of iterations of calculation were needed with our traditional modeling based tool. As evidence of this, despite the use of a slower language, W's AWK code runs faster than the C++/LISP used in

STAR/NOVA. It takes just minutes to conduct 20 trials over 13 data sets with W. A similar trial, conducted with NOVA or STAR, can take hours to run.

4. From CBR to W

A standard CBR algorithm reports the median class value of some local neighborhood. The W algorithm treats the local neighborhood in a slightly different manner:

- The local neighborhood is divided into best and rest.
- A contrast set is learned that most separates the regions (contrast sets contain attribute ranges that are common in one region, but rare in the other).
- W then searches for a subset of the contrast set that best selects for (e.g.) the region with lower effort estimates.

The rest of this section details the above process.

4.1. Finding Contrast Sets

Once a contrast set learner is available, it is a simple matter to add W to CBR. W finds contrast sets using a greedy search, where candidate contrast sets are ranked by the log of the odds ratios. Let some attribute range x appear at frequency N_1 and N_2 in two regions of size R_1 and R_2 . Let the R_1 region be the preferred goal and R_2 be some undesired goal. The log of the odds ratio, or LOR, is:

$$\text{LOR}(x) = \log \left(\frac{N_1/R_1}{N_2/R_2} \right) \quad (2)$$

Note that when $\text{LOR}(x) = 0$, then x occurs with the same probability in each region (such ranges are therefore not useful for selecting on region or another). On the other hand, when $\text{LOR}(x) > 0$, x is more common in the preferred region than otherwise. These LOR-positive ranges are candidate members of the contrast set that selects for the desired outcome.

It turns out that, for many data sets, the LOR values for all the ranges contain a small number of very large values (strong contrasts) and a large number of very small values (weak contrasts). The reasons for this distribution do not concern us here (and if the reader is interested in this master-variable effect, they are referred to [19,20]). What is relevant is that the LOR can be used to rank candidate members of a contrast set. W computes the LORs for all ranges, and then conducts experiments applying the top i -th ranked ranges.

For more on LOR, and their use for multi-dimensional data, see [21].

4.2. The Algorithm

CBR systems input a query q and a set of cases. They return the subset of cases C that is relevant to the query. In the case of W:

- Each case C_i is an historical record of one software

projects, plus the development effort required for that project. Within the case, the project is described by a set of attributes which we assume have been discretized into a small number of discrete values (e.g. analyst capability $\in \{1, 2, 3, 4, 5\}$ denoting very low, low, nominal, high, very high respectively).

- Each query q is a set of constraints describing the particulars of a project. For example, if we were interested in a schedule over-run for a complex, high reliability projects that have only minimal access to tools, then those constraints can be expressed in the syntax of **Figure 2**.

W seeks q' (a change to the original query) that finds another set of cases C' such that the median effort values in C' are less than that of C (the cases found by q). W finds q' by first dividing the data into two-thirds training and one-third testing. Retrieve and reuse are applied to the training set. Revising is then applied to the test set.

1) *Retrieve*: The initial query q is used to find the N training cases nearest to q using a Euclidean distance measure where all the attribute values are normalized from 0 to 1.

2) *Reuse* (adapt): The N cases are sorted by effort and divided into the K_1 best cases (with lowest efforts) and K_2 rest cases. For this study, we used $K_1 = 5, K_2 = 15$. Then we seek the contrast sets that select for the K_1 best cases with lowest estimates. All the attribute ranges that the user has marked as “controllable” are scored and sorted by LOR. This sorted order S defines a set of candidate q' queries that use the first i -th entries in S :

$$q_i' = q \cup S_1 \cup S_2 \dots \cup S_i \tag{3}$$

Formally, the goal of W is find the smallest i value such q_i' selects cases with the least median estimates.

According to **Figure 3**, after retrieving and reusing comes revising (this is the “verify” step). When revising q' , W prunes away irrelevant ranges as follows:

- 1) Set $i = 0$ and $q_i' = q$.
- 2) Let $Found_i$ be the test cases consistent with q_i' (i.e., that do not contradict any of the attribute ranges in q_i').
- 3) Let $Effort_i$ be the median efforts seen in $Found_i$.

```
@project example
@attribute ?rely 3 4 5
@attribute tool 2
@attribute cplx 4 5 6
@attribute ?time 4 5 6
```

Figure 2. W 's syntax for describing the input query q . Here, all the values run 1 to 6. $4 \leq cplx \leq 6$ denotes projects with above average complexity. Question marks denote what can be controlled in this case, *rely*, *time* (required reliability and development time).

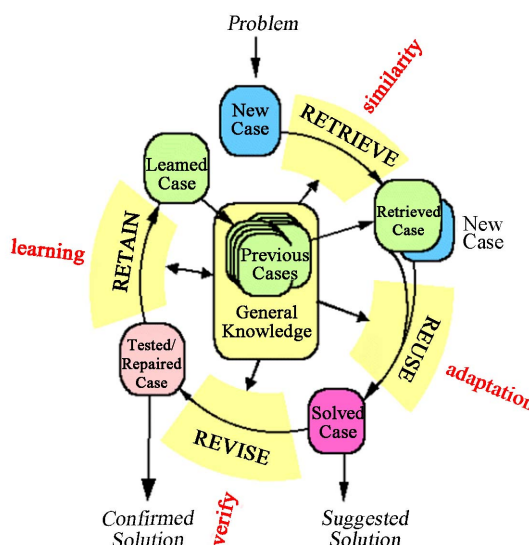


Figure 3. A diagram describing the steps of CBR (source: <http://www.peerscience.com/Assets/cbrcycle1.gif>).

4) If $Found$ is too small then terminate (due to over-fitting). After Shepperd [2], we terminated for $|Found| < 3$.

5) If $i > 1$ and $Effort_i < Effort_{i-1}$, then terminate (due to no improvement).

6) Print q_i' and $Effort_i$.

7) Set $i = i + 1$ and $q_i' = q_{i-1} \cup S_i$

8) Go to step 2.

On termination, W recommends changing a project according to the set $q' - q$. For example, in **Figure 2**, if $q' - q$ is $rely = 3$ then this treatment recommends that the best way to reduce the effort for this project is to reject $rely = 4$ or 5 .

One useful feature of the above loop is that it is not a black box that offers a single “all-or-nothing” solution. Rather it generates enough information for a user to make their cost-benefit tradeoffs. In practice, users may not accept all the treatments found by this loop. Rather, for pragmatic reasons, they may only adopt the first few S_i changes seen in the first few rounds of this loop. Users might adopt this strategy if (e.g.) they have limited management control of a project (in which case, they may decide to apply just the most influential S_i decisions).

Implementing W is simpler than the STAR/NOVA approach:

- Both NOVA and STAR contain a set process models for predicting effort, defects, and project threats as well as Monte Carlo routines to randomly select values from known ranges. STAR and NOVA implement simulated annealing while NOVA also implements other search algorithms such as A^* , LDS,

MAXWALKSAT, beam search, etc. STAR/NOVA are 3000 and 5000 lines of C++ and LISP, respectively.

- W, on the other hand, is a 300 line AWK script.

Our pre-experimental suspicion was that W was too simple and would need extensive enhancement. However, the results shown below suggest that, at least for this task, simplicity can suffice (but see the future work section for planned extensions).

Note that W verification results are more rigorous than those of STAR/NOVA. W reports results on data that is external to its deliberation process (*i.e.*, on the test set). STAR/NOVA, on the other hand, only reported the changes to model output once certain new constraints were added to the model input space.

5. Data

Recall that a CBR system takes input a query q and cases C . W has been tested using multiple queries on the data sets of **Figure 4**. These queries and data sets are described below.

5.1. Data Sets

As shown in **Figure 4**, our data includes:

- The standard public domain COCOMO data set (Cocoma81);
- Data from NASA;
- Data from the International Software Benchmarking Standards Group (ISBSG);
- The Desharnais and Maxwell data sets.

Except for ISBSG, all the data used in this study is available at <http://promisedata.org/data> or from the authors.

Note the skew of this data (min to median much smaller than median to max). Such asymmetric distributions complicate model-based methods that use Gaussian approximations to variables.

There is also much divergence in the attributes used in our data:

- While our data include effort values (measured in terms of months or hours), no other feature is shared by all data sets.
- The COCOMO and NASA data sets all use the attributes defined by Boehm [9]; e.g. analyst capability,

required software reliability, memory constraints, and use of software tools.

- The other data sets use a variety of attributes such as the number of data model entities, the number of basic logical transactions, and number of distinct business units serviced.

This attribute divergence is a significant problem for model-based methods like STAR/NOVA since those systems can only accept data that conforms to the space of attributes supported by their model. For example, this study uses the five data sets listed in **Figure 2**. STAR/NOVA can only process two of them (coc81 and nasa93). CBR tools like W, on the other hand, avoid these two problems:

- W makes no assumption about the distributions of the variables.
- W can be easily applied to any attribute space (caut: as long but there are some dependent variables).

5.2. Queries

Figure 2 showed an example of the W query language:

- The idiom “@attribute name range” defines the range of interest for some attribute “name”.
- If “range” contains multiple values, then this represents a disjunction of possibilities.
- If “range” contains one value, then this represents a fixed decision that cannot be changed.
- The idiom “?x” denotes a controllable attribute (and W only generates contrast sets from these controllables). Note that the “range”s defined for “?x” must contain more than one value, otherwise there is no point to making this controllable.

6. Case Studies

6.1. Case Study #1: Brooks’ Law

This section applies W to Brooks’ Law. Writing in the 1970s [22], Brooks noted that software production is a very human-centric activity and managers need to be aware of the human factors that increase/decrease productivity. For example, a common practice at that time at IBM was to solve deadline problems by allocating more resources. In the case of programming, this meant

Dataset	Attributes	Number of cases	Content	Units	Min	Median	Mean	Max	Skewness
coc81	17	63	NASA projects	months	6	98	683	11400	4.4
nasa93	17	93	NASA projects	months	8	252	624	8211	4.2
desharnais	12	81	Canadian software projects	hours	546	3647	5046	23940	2.0
maxwell	26	62	Finnish banking software	months	6	5189	8223	63694	3.3
isbsg	14	29	Banking projects of ISBSG	minutes	662	2355	5357	36046	2.6

Figure 4. The 328 projects used in this study come from 5 data sets.

```
@project brooks
@attribute apex 1
@attribute plex 1
@attribute ltex 1
@attribute kloc 0.9 2.2 3 [snip] 339 350 352 423 980
```

Figure 5. The brookslaw query.

Treatment	median	spread
AsIs = Nasa93	225	670
ToBe1 = nasa93 \cup q	380	680
ToBe2 = nasa93 \cup q'	220	290

Figure 6. Effort estimates seen in different treatments for the brooks' law experiment.

adding more programmers to the team. Brooks argued that this was an inappropriate response since, according to Brooks' law "adding manpower [sic] to a late software project makes it later". The reason for this slowdown is two-fold:

- The more people involved the greater the communication overhead. While this is certainly an issue if all parts of the software system are accessible to all other parts, with an intelligent module design, this first issue can be mitigated.
- The second issue is more fundamental. Software construction is a complex activity. Newcomers to a project suffer from inexperience in the tools, the platform, the problem domain, etc.

The query of **Figure 5** models this second issue. In this query, all the experience attributes have been set to their lowest value (*apex*, *plex*, *ltex* are analyst experience, programmer language experience, and language and tool experience, respectively). The remaining attributes are all controllable and are allowed to move over their full range. For a (dataset, query) of (nasa93, $q = \text{brookslaw}$), W returns

$$q' = q \cup (\text{data} = 2) \quad (4)$$

That is, W is recommended setting the database size to its lowest value. Databases are used to store program and data elements. In effect, W is recommending reigning in the scope of the project. The recommendation can be paraphrased as follows:

If the project is late, and you add more staff with less experience, you can still finish on time if you decrease the scope of the project.

Figure 6 shows the effects of this recommendation. The AsIs row shows the median and spread of the effort values in nasa93. The ToBe1 row shows the effect of Brooks' Law. In the subset of the data consistent with $aexp = plex = ltex = 1$, the median effort has nearly doubled. The ToBe2 row shows the impact of W 's recommendation: the project will now finish in nearly the time

as the AsIs row, and the spread is greatly reduced.

One of the reasons we are exploring W is the simplicity of the implementation. In this regard, it is useful to compare our results on Brooks' Law to other researchers. Brooks' Law is a well researched effect and other researchers have found special cases where the general law does not hold. For example, using sophisticated qualitative reasoning techniques, Zhang *et al.* [23] found their own loopholes in Brooks' Law. One of us (Keung) worked on site with the Zhang team and reports that the Brooks analysis was the main result of a two year masters graduate thesis. In contrast, writing W took three days and the specific analysis of Brooks' Law took twenty minutes from first posing the question, to graphing the output.

6.2. Thirteen More Case Studies

Apart from the Brooks' Law experiment, we have tested W on thirteen other case studies:

- For the ISBSG data set, we used our recent experience to describe the constraints suitable for a stand-alone or client server system (denoted *stdalone* and *clientsrv*).
- For the Desharnais data set, we posed queries representing:
 - (s; m; l) denotes (small, medium, large) projects;
 - (mngnr; team) denotes (manager, team) experience being low.
- For the Cocomo and NASA data sets, we used our contacts at the Jet Propulsion Laboratory to write queries describing (a) osp (see above); (b) the second version of that system called osp2; as well as (c) generic flight and (d) ground systems.
- Lacking direct experience with the Finnish financial system, we could not pose specific queries to the Maxwell dataset. Instead, we made half the attributes controllable and used that for the Maxwell query.

Figure 7 shows the improvements seen in our 13 queries, running on the data sets of **Figure 4**. As shown by the last line of **Figure 7**, the usual improvements where (36, 68) for (median, spread). Note that, unlike STAR/NOVA, these are results on real-world data sets not used during training.

Figure 8 displays the **Figure 7** results graphically. The dashed line indicates the median of the improvements for each axis. One data set had consistently worse results than any other. The gray cells of **Figure 7** indicate when W failed (*i.e.*, where the treatments increased median development effort or spread, or both). Note that the gray cells are found only in the Desharnais results.

On investigation, the root cause of the problem was

dataset	query q	Improvement	
		median	spread
coc81	ground	95%	83%
coc81	osp	82%	68%
nasa93	ground	77%	15%
ISBSG	stdalone	69%	100%
nasa93	flight	61%	73%
nasa93	osp	49%	48%
maxwell		44%	76%
ISBSG	clientserv	42%	88%
desharnais	mngr-l	36%	-45%
coc81	osp2	31%	71%
nasa93	osp2	27%	42%
desharnais	mngr-s	23%	85%
desharnais	mngr-m	23%	85%
coc81	flight	0%	18%
desharnais	team-m	0%	60%
desharnais	team-s	-15%	-206%
desharnais	team-l	-15%	-93%
	median	36%	68%

Figure 7. Improvements (100 * (initial - final)/initial) for 13 queries, sorted by median improvement. Gray cells show negative improvement.

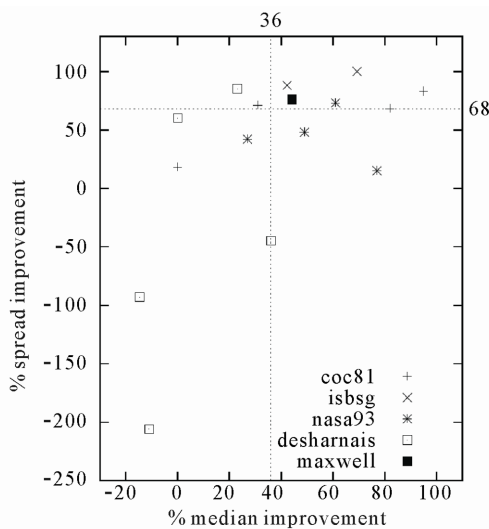


Figure 8. Median and spread improvements for effort estimation. dashed lines mark median values.

the granularity of the data. Whereas (e.g.) coc81 assigns only one of six values to each attribute, Desharnais' attributes had a very wide range.

Currently, we are exploring discretization policies to [24,25] reduce attributes with large cardinality to a smaller set. Tentatively, we can say that discretization solves the problem with Desharnais but we are still

studying this aspect of our system. Even counting the negative Desharnais results, in the majority of cases W found treatments that improved both the median and spread of the effort estimates. Sometimes, the improvements were modest: in the case of (coc81, flight), the median did not improve (but did not get worse) while the spread was only reduced by 18%. But sometimes the improvements are quite dramatic.

- In the case of (ISBSG, stdalone), a 100% improvement in spread was seen when q' selected a group of projects that were codeveloped and, hence, all had the same development time.
- In the case of (coc81, ground), a 95% effort improvement was seen when q' found that ground systems divide into two groups (the very expensive, and the very simple). In this case W found the factor that drove a ground system into the very simple case.

7. Discussion

7.1. Comparisons to NOVA

Figure 9 shows that estimation improvements found by W (in this report) to the improvements reported previously (in Figure 1). This table is much shorter than Figure 7 since, due to the modeling restrictions imposed by the software process models, NOVA cannot be applied to all the data sets that can be processed by W. The numbers in Figure 9 cannot be directly compared due to the different the different goals of the two systems: W tries to minimize effort while NOVA tries to minimize effort and development time and delivered defects (we are currently extending W to handle such multiple-goal optimization tasks). Nevertheless, it is encouraging to note that the results are similar and that the W improvements are not always less than those found by STAR/NOVA.

Regardless of the results in Figure 9, even though we prefer W (due to the simplicity of the analysis) there are clear indicators of when we would still use STAR/NOVA. W is a case-based method. If historical cases are not available, then STAR/NOVA is the preferred method. On the other hand, STAR/NOVA is based on the USC COCOMO suite of models. If the local business users do not endorse that mode, then W is the preferred method.

Treatment	median	spread
AsIs = Nasa93	225	670
ToBe1 = nasa93 \cup q	380	680
ToBe2 = nasa93 \cup q'	220	290

Figure 9. Comparing improvements found by NOVA (from Figure 1) and W (Figure 7).

7.2. Threats to Validity

External validity is the ability to generalize results outside the specifications of that study [26]. To ensure the generalizability of our results, we studied a large number of projects. Our datasets contain a wide diversity of projects in terms of their sources, their domains and the time period they were developed in. Our reading of the literature is that this study uses more project data, from more sources, than numerous other papers. Table 4 of [27] lists the total number of projects in all data sets used by other studies. The median value of that sample is 186, which is less much less than the 328 projects used in our study.

Internal validity questions to what extent the cause-effect relationship between dependent and independent variables hold [28]. For example, the above results showed reductions in the effort estimates of up to 95%; *i.e.*, by a factor of 20. Are such massive reductions possible?

These reductions are theoretically possible. Making maximal changes to the first factor (personnel/team capability) can affect the development effort by up to a 3.53. Making maximal changes just to the first four factors could have a net effect of up to:

$$3.53 * 2.38 * 1.62 * 1.54 \approx 21 > 20 \quad (5)$$

By the same reasoning, making maximal changes to all factors could have a net effect of up to eleven thousand. Hence, an improvement of 95% (or even more) is theoretically possible. As to what is pragmatically possible, that is a matter for human decision making. No automatic tool such as STAR/NOVA/W has access to all the personnel factors and organizational constraints known to a human manager. Also, some projects are inherently expensive (e.g. the flight guidance system of a manned spacecraft) and cutting costs by, say, reducing the required reliability of the code is clearly not appropriate. Tools like W are useful for uncovering options that a human manager might have missed, yet ultimately the actual project changes must be a human decision.

8. Conclusions

If a manager is given an estimate for developing some software, they may ask “how do I change that?” The model variance problem makes it difficult to answer this question. Our own prior solution to this problem required an underlying process model that limited the data sets that can be analyzed.

This paper has introduced W, a case-based reasoning approach (augmented with a simple linear-time greedy search for contrast sets). W provides a mechanism that allows projects to improve over time based on historical events, greatly assist in the project planning and resource allocation. W has proven to be simpler to implement and use than our prior solutions. Further, since this new me-

thod has no model-based assumptions, it can be applied to more data sets. When tested on 13 real-world case studies, this approach found changes to projects that could greatly reduce the median and variance of the effort estimate.

REFERENCES

- [1] B. Boehm, E. Horowitz, R. Madachy, D. Reifer, B. K. Clark, B. Steece, A. W. Brown, S. Chulani and C. Abts, “Software Cost Estimation with Cocomo II,” Prentice Hall, New Jersey, 2000.
- [2] M. Shepperd and C. Schofield, “Estimating Software Project Effort Using Analogies,” *IEEE Transactions on Software Engineering*, Vol. 23, No. 11, November 1997, pp. 736-743.
- [3] Y. Li, M. Xie and T. Goh, “A Study of Project Selection and Feature Weighting for Analogy Based Software Cost Estimation,” *Journal of Systems and Software*, Vol. 82, No. 2, February 2009, pp. 241-252.
- [4] O. El-Rawas, “Software Process Control without Calibration,” Master’s Thesis, Morgantown, 2008.
- [5] T. Menzies, O. El-Rawas, J. Hihn and B. Boehm, “Can We Build Software Faster and Better and Cheaper?” *Proceedings of the 5th International Conference on Predictor Models in Software Engineering (PROMISE’09)*, 2009.
- [6] T. Menzies, O. Elrawas, J. Hihn, M. Feathear, B. Boehm and R. Madachy, “The Business Case for Automated Software Engineering,” *Proceedings of the Twenty-second IEEE/ACM International Conference on Automated Software Engineering (ASE’07)*, 2007, pp. 303-312.
- [7] T. Menzies, S. Williams, O. El-rawas, B. Boehm and J. Hihn, “How to Avoid Drastic Software Process Change (Using Stochastic Stability),” *Proceedings of the 31st International Conference on Software Engineering (ICSE’09)*, 2009.
- [8] K. Cowing, “Nasa to Shut down Checkout & Launch Control System,” August 26, 2002. <http://www.spaceref.com/news/viewnews.html>
- [9] B. Boehm, “Software Engineering Economics,” Prentice Hall, New Jersey, 1981.
- [10] C. Kemerer, “An Empirical Validation of Software Cost Estimation Models,” *Communications of the ACM*, Vol. 30, No. 5, May 1987, pp. 416-429.
- [11] D. Baker, “A Hybrid Approach to Expert and Model-Based Effort Estimation,” Master’s Thesis, Morgantown, 2007.
- [12] P. Green, T. Menzies, S. Williams and O. El-waras, “Understanding the Value of Software Engineering Technologies,” *Proceedings of the 2009 IEEE/ACM International Conference on Automated Software Engineering (ASE’09)*, 2009.
- [13] T. Menzies, O. Elrawas, D. Baker, J. Hihn and K. Lum, “On the Value of Stochastic Abduction (If You Fix Everything, You Lose Fixes for Everything Else),” *International Workshop on Living with Uncertainty (An ASE’07*

- Co-Located Event*), 2007.
- [14] R. C. Schank, "Dynamic Memory: A Theory of Reminding and Learning in Computers and People," Cambridge University Press, New York, 1983.
 - [15] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches," *Artificial Intelligence Communications*, Vol. 7, No. 1, 1994, pp. 39-59.
 - [16] M. J. Shepperd, "Case-Based Reasoning and Software Engineering," Technical Report TR02-08, Bournemouth University, UK, 2002.
 - [17] G. Kadoda, M. Cartwright, L. Chen and M. Shepperd, "Experiences Using Casebased Reasoning To Predict Software Project Effort," Keele University, Staffordshire, 2000.
 - [18] E. Mendes, I. D. Watson, C. Triggs, N. Mosley and S. Counsell, "A Comparative Study of Cost Estimation Models for Web Hypermedia Applications," *Empirical Software Engineering*, Vol. 8, No. 2, June 2003, pp. 163-196.
 - [19] T. Menzies, D. Owen and J. Richardson, "The Strangest Thing about Software," *IEEE Computer*, Vol. 40, No. 1, January 2007, pp. 54-60.
 - [20] T. Menzies and H. Singh, "Many Maybes Mean (Mostly) the Same Thing," In: M. Madravio, Ed., *Soft Computing in Software Engineering*, Springer-Verlag, Berlin Heidelberg, 2003.
 - [21] M. Možina, J. Demšar, M. Kattan and B. Zupan, "Nomograms for Visualization of Naive Bayesian Classifier," *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'04)*, Vol. 3202, 2004, pp. 337-348.
 - [22] F. P. Brooks, "The Mythical Man-Month," Anniversary Edition, Addison-Wesley, Massachusetts, 1975.
 - [23] H. Zhang, M. Huo, B. Kitchenham and R. Jeffery, "Qualitative Simulation Model for Software Engineering Process," *Proceedings of the Australian Software Engineering Conference (ASWEC'06)*, 2006, pp. 10-400.
 - [24] J. Dougherty, R. Kohavi and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features," *International Conference on Machine Learning*, 1995, pp. 194-202.
 - [25] Y. Yang and G. I. Webb, "A Comparative Study of Discretization Methods for Naive-Bayes Classifiers," *Proceedings of PKAW 2002: The 2002 Pacific Rim Knowledge Acquisition Workshop*, 2002, pp. 159-173.
 - [26] D. Milicic and C. Wohlin, "Distribution Patterns of Effort Estimations," *30th EUROMICRO Conference (EUROMICRO'04)*, 2004.
 - [27] B. Kitchenham, E. Mendes and G. H. Travassos, "Cross Versus Within-Company Cost Estimation Studies: A Systematic Review," *IEEE Transactions on Software Engineering*, Vol. 33, No. 5, May 2007, pp. 316-329.
 - [28] E. Alpaydin, "Introduction to Machine Learning," MIT Press, Cambridge, 2004.