

# Multi-criteria Decision Analysis for Customization of Estimation by Analogy Method AQUA<sup>+</sup>

Jingzhou Li

Software Engineering Decision Support  
Laboratory, University of Calgary, Calgary AB,  
Canada, T2N1N4  
jingli@ucalgary.ca

Guenther Ruhe

Software Engineering Decision Support  
Laboratory, University of Calgary, Calgary AB,  
Canada, T2N1N4  
ruhe@ucalgary.ca

## ABSTRACT

The quality of results from a predictor model depends on the proper customization of the parameters of the model. For Estimation by Analogy (EBA), the impact of the parameter "Attribute weighting technique" has been shown by several authors. The decision problem "Which attribute weighting technique is preferable for EBA in which situation?" is considered in this paper from the perspective of multi-criteria decision analysis (MCDA). The empirical results are given for the EBA method AQUA<sup>+</sup>.

More specifically, two MCDA techniques, ELECTRE and Pareto-optimality are applied. Three evaluation criteria MMRE (Mean Magnitude of Relative Error), Pred (Prediction at certain accuracy level), and Strength are considered. We discuss the insights gained from this more in-depth decision analysis for the stated decision problem.

## Categories and Subject Descriptors

D.2.9 [Software Engineering]: Management – *cost estimation*.

## General Terms

Management, Measurement, Experimentation.

## Keywords

Software effort estimation, Estimation by analogy, Technology customization, Software engineering decision support, Multi-criteria decision analysis.

## 1. INTRODUCTION

### 1.1 Motivation

Customization of software engineering technologies is important to fully exploit their inherent potential. Often, different alternative methods and techniques are available, but not so much is known for when and why to use which of them? While this is true in general, we focus on supporting decisions in the process of customization for (effort) estimation by analogy (EBA) [1, 2]. Customization in this context means to adjust the possible

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

PROMISE'08, May 12–13, 2008, Leipzig, Germany.

Copyright 2008 ACM 978-1-60558-036-4/08/05...\$5.00.

parameters (options) within the course of the method to the given context of the problem.

EBA in general and the EBA method AQUA<sup>+</sup> [3] in particular, have proven to be promising software effort estimation methods. This paper continues on our previous research regarding the decision support for customization of EBA in [4]. Different from the single criterion approach used in [4], we look at the problem of customization of EBA from a multi-criteria decision analysis (MCDA) perspective. The reason for that is that the very nature of the problem is multi-criteria. Form applying two of the proven MCDA techniques, we aim at providing more insight into the decision problem, thus supporting (better) decisions towards customization.

### 1.2 Customization of EBA Method AQUA<sup>+</sup>

In order to apply EBA, a sequence of steps must be followed to complete required tasks. The sequence of steps constitutes a process for executing EBA. A decision centric process model of EBA method AQUA<sup>+</sup> is proposed in [4] so that AQUA<sup>+</sup> can be applied in a systematic and repeatable way. Some example tasks of EBA are (i) defining attribute weighting, (ii) dealing with missing values in a data set, (iii) determining similarity measures, and (iv) selecting analogy adaptation strategy. In [4], twelve process steps and their possible options for execution are studied. For all of them, it is unclear upfront which one is preferable in which situation.

This paper does not aim to solve the whole customization question, but to propose a decision-centric method for how it can be done. The method is applied and illustrated for the question of selecting the "right" heuristic for attribute weighing within AQUA<sup>+</sup> [5].

Choosing the "right" heuristic depends on the different criteria of consideration. Without making any upfront assumption about their relative importance, multi-criteria decision analysis aims at finding most promising alternatives and providing the reason why.

### 1.3 The decision problem

As one of the tasks of the EBA method AQUA<sup>+</sup>, the selection of the attribute weighting heuristic is described as a decision problem. It is composed of the following elements:

**Decision alternatives:** Attribute weighting heuristics known from literature. We will study six heuristics in this paper.

**Evaluation method:** The alternative heuristics are evaluated by applying them for different data sets for the learning-based method AQUA<sup>+</sup>. We will consider their performance for six publicly available data sets.

**Evaluation criteria:** There are different evaluation criteria for the quality of effort prediction methods. We select three of the most commonly used ones: *MMRE*, *Pred* [6], and *Strength* [3]. In order to keep the criteria consistent for minimization, we consider *MMRE*, *1-Pred*, and *1-Strength*

**Decision objective:** Determine solution alternatives (heuristics) such that evaluation criteria *MMRE*, *1-Pred*, and *1-Strength* get minimized in a balanced manner.

## 1.4 Objectives of this research

In our former papers [4, 5] we used a single aggregated criterion to judge which attribute weighting heuristic works best for the data sets under consideration. The downside of such an approach is that the performance against the individual criteria is no longer visible and not taken into account. Instead of considering a heuristic favorable because it performs “good enough” for all three stated criteria, it might be preferable to select one that is especially good in terms of *MMRE* (as an example).

The advantage of modeling the decision problem from a multi-criteria perspective is that you allow the decision-maker to bring in his/her preference between the criteria at the latest point in time (instead of upfront). In other words, timeliness of decisions is considered to be a quality factor for selection of the heuristics.

To actually perform the multi-criteria analysis, we consider two proven techniques. The first one is ELECTRE [7] which aims at establishing an outranking relationship among alternatives. The second analysis is to look for Pareto-optimal alternatives, or Pareto-optimality [8]. We compare our previous results from using single criterion (aggregated) evaluation with the results from running MCDA and discuss the additional insight gained.

The rest of the paper is organized as follows. Section 2 gives the necessary background of MCDA. Sections 3 and 4 present the analysis results gained from the application of ELECTRE and Pareto-optimality analyses, respectively. A discussion of the results is provided in Section 5, and the paper concludes with an outlook on future research in section 6.

## 2. INTRODUCTION TO AQUA<sup>+</sup>

In AQUA<sup>+</sup>, the historical data set *DB* is defined as a triple:

$$DB = \langle R, P, V \rangle.$$

Therein, *R* is the set of objects  $R = \{r_1, r_2, \dots, r_n\}$ , and  $P = A \cup \{Effort\}$ ;  $A = \{a_1, a_2, \dots, a_m\}$  is the set of attributes to describe the objects; *Effort* is a specific attribute characterizing the effort for implementing the respective object.  $Effort(r_i)$  represents the effort to develop object  $r_i$ .  $V = \{a_j(r_k)\}$  is the domain of attribute values of all objects in *R*, and  $a_j(r_k)$  represents the value of attribute  $a_j \in P$  for object  $r_k \in R$ . The set  $S = \{s_1, s_2, \dots, s_i\}$  denotes the given objects to be estimated and *S* shares the same attributes *A* with *R*.

The problem of EBA can be stated as follows:

For all  $s_g \in S$ , the effort of  $s_g$ ,  $Effort(s_g)$ , is to be estimated based on the values of *Effort* from a set of most similar objects  $r_i \in R$  to  $s_g$ .

The set of most similar objects are also called analogs of  $s_g$ , which are retrieved from *R* using certain similarity measures. Estimation using effort information from the analogs is also known as analogy adaptation.

In AQUA<sup>+</sup>, the weights of attributes were involved in calculating

the global similarity between two objects. Therefore, there are three phases in AQUA<sup>+</sup>:

- Attribute weighting
- Learning over the historical data set
- Predicting

In order to measure the quality of the estimation results, Jack-knife cross-validation [9] is performed to determine the prediction accuracy distribution of *DB*, denoted by  $AccuDistr(DB)$ . This is done by varying the threshold for the number of analogs (*N*) and the threshold for similarity measure (*T*) of the analogs taken for analogy adaptation. Each row of  $AccuDistr(DB)$  contains a vector of criteria. The relevant criteria to the current research are *MMRE*, *Pred*, and *Strength*, whose brief definitions are given below.

**Definition 1.**  $MRE(r_k)$ —Magnitude of Relative Error [6]

$$MRE(r_k) = \frac{|Effort(r_k) - \tilde{Effort}(r_k)|}{Effort(r_k)} \quad (1)$$

for a given object  $r_k \in R$  under estimation, where  $Effort(r_k)$  is the actual effort and  $\tilde{Effort}(r_k)$  is the estimated effort of object  $r_k$ . ■

**Definition 2.**  $MMRE(N, T)$ —Mean Magnitude of Relative Error [6]

$$MMRE(N, T) = \frac{1}{n} \sum_{r_k \in R} MRE(r_k) \quad (2)$$

for a given pair of values of (*N*, *T*) for all the *n* objects in *R* in a single run of Jack-knife cross-validation. ■

**Definition 3.**  $Pred(\alpha, N, T)$ —prediction at level  $\alpha$  [6]

$$Pred(\alpha, N, T) = \frac{\tau}{\lambda} \quad (3)$$

where  $\lambda$  is the total number of objects that are estimated in a single run of Jack-knife cross-validation with a given pair of values of (*N*, *T*), and  $\tau$  is the number of objects with  $MRE \leq \alpha$ . ■

$\alpha=0.25$  is normally used for actual evaluation in literature and is used in this paper; and  $Pred(0.25)$  is used when *N* and *T* are given or not considered at all.

**Definition 4.**  $Strength(N, T)$

*Support(N, T)* is the number of objects in *R* that can be estimated with a given values of (*N*, *T*).  $Strength(N, T)$  is then defined as the ratio of *Support* to the total number of objects in *R*. ■

## 3. MULTI-CRITERIA DECISION ANALYSIS

Multi-criteria analysis is briefly introduced by explaining the data and criteria used for the analysis, the decision alternatives, and the methods to determine a solution, i.e. decision alternative.

### 3.1 The data used for analysis

There are two parts of the data used for conducting MCDA in this paper. First part is the data sets, e.g. *DB*, that are used for EBA with AQUA<sup>+</sup>. The second part is the data, e.g.  $AccuDistr(DB)$ , that measure the estimation accuracy of AQUA<sup>+</sup>. Because the attribute weighting heuristics affect the estimation accuracy of AQUA<sup>+</sup>, the accuracy data about AQUA<sup>+</sup> are thus used to measure the performance of an attribute heuristic as well. Therefore, the data used for MCDA are actually the  $AccuDistr(DB)$  that are obtained by applying AQUA<sup>+</sup> to the original data sets *DBs*.

Table 1 shows the parameters of the data sets *DBs* with

- “#Objects” - number of objects in the data set;

- "#Attributes" - number of attributes (excluding Effort);
- "%Missing values" - percentage of missing values;
- "%Non-Quantitative attributes" - percentage of non-quantitative attributes.

In this study, a subset of ISBSG04 [10], called ISBSG04-2, which contains objects whose effort is in the range of 10,000 and 20,000 (hours), is used to keep computational effort reasonable.

**Table 1. Summary of the data sets for analysis**

Name	#Objects	#Attributes	%Missing values	%Non-quantitative attributes	Source
USP05-FT	121	14	2.54	71	[3, 13]
USP05-RQ	76	14	6.8	71	[3, 13]
ISBSG04-2	158	24	27.24	63	[10]
Kem87	15	5	0	40	[11]
Mends03	34	6	0	0	[12]
Desh89	81	10	0.006	20	[2, 13]

Due to the size of the *AccuDistr(DB)* for each data set, it is impossible to present them in the paper. Some example segments of *AccuDistr(DB)* can be found in [5].

### 3.2 The decision alternatives

Our decision alternatives under investigation are different attribute weighting heuristics applicable to EBA. In this research, we focus on weighting heuristics based on Rough Set Analysis (RSA) [14]. Four RSA-based attribute weighting heuristics H1 to H4 were proposed in [5] and the heuristic H0 that uses equal weights for all attributes is regarded as a baseline heuristic. Because the core attributes do not exist in some data sets, H2 is ignored. Only H1, H3, and H4, along with H0 are used in the following analysis.

In addition to the four RSA-based heuristics, two other heuristics CfsSubset and Wrapper provided in Weka [15] are also investigated in this paper. These two heuristics provide attribute selection, i.e. a subset of attributes is selected, other than weighting of attributes.

The six heuristics are:

- H0: Equal weights for all involved attributes,
- H1: Reducts based heuristic,
- H3: Decision rule based heuristic using the number of occurrences,
- H4: Decision rule based heuristic using relative strength;
- CsfSubset: Attribute Evaluator is CfsSubset and Search Method is Greedy Stepwise for attribute selection;
- Wrapper: Attribute Evaluator is Wrapper and Search Method is Greedy Stepwise for attribute selection.

### 3.3 The evaluation criteria

The evaluation criteria for evaluating the attribute weighting heuristics in the multi-criteria analysis are the same as those used for evaluating AQUA<sup>+</sup>: *MMRE*, *Pred*, and *Strength*. Therefore, the *AccuDistr(DB)* contains actually the values of the three criteria after applying an attribute weighting heuristic to AQUA<sup>+</sup>.

### 3.4 ELECTRE method

ELECTRE-IS [7] is a method providing support for the problem of selecting one (or a set of) alternative(s) out of a given finite set of alternatives. ELECTRE-IS is one of a family of methods for multi-criteria decision support. ELECTRE methods in general comprise two main procedures: (i) construction of one or several outranking relation(s) followed by (ii) an exploitation procedure.

The foundation of ELECTRE-IS is the outranking relation  $R$  in which  $R(a, b)$  means alternative  $a$  is at least as good as alternative  $b$ . The relation  $R$  is not required to be complete, e.g., there exist alternatives  $a$  and  $b$ , that are incomparable. Each alternative can be evaluated against a set of individual criteria. This idea is applied in our context for the question of determining a preference relationship between different candidate heuristics as part of AQUA<sup>+</sup>.

Each individual order relation (derived from a criterion) is assigned a weight representing the relative importance of the criterion in the overall comparison. The method then aggregates these relations using their weights to achieve a final outranking relation from which one (or a set of) alternative(s) can be identified as not outranked by any alternative. For an outranking relation  $R(a, b)$  to be validated, a sufficient majority of criteria should be in favor of (concordance with) this assertion. In addition to that, when the concordance condition holds, none of the criteria in the minority should too strongly oppose the assertion  $R(a, b)$ .

The rationale of our choice of a method such as ELECTRE-IS is based on its ability to handle uncertainty in preference building. When asked about what makes a good attribute weighting heuristic, any project manager can come up with at least more than one criteria. However, especially under the assumption of non-compensative criteria, it is difficult to combine these criteria into an objective function. The final outranking relation (expressed by a directed graph) helps to decide which heuristic is preferable to others in which situation.

### 3.5 Pareto-optimality method

We consider a set  $X$  of solution alternatives  $x$  and a vector  $F(x) = (f_1(x), f_2(x), f_3(x))$  of three criteria as formulated in the decision problem in Section 1.3. Each alternative  $x$  is one attribute weighting heuristic under investigation.

The objective is to minimize the vector function  $F(x)$ . For a given vector-valued function  $F$ , a solution  $x^*$  belongs to the set Pareto- $X$  (called Pareto-optimal solutions), if there is no alternative  $x'$  of  $X$  such that

$$F(x') \leq F(x^*) \text{ and } F(x') \neq F(x^*).$$

For a given problem, we will call the set of all Pareto-optimal solutions the Pareto frontier.

### 3.6 Clustering

Looking at the Pareto frontier, the question becomes if there are clusters of points (Pareto solutions) which are originated by the same heuristics. This would give us the "hot spots" (range of relative importance between the criteria considered) for the application of this heuristic

In order to explore which set of heuristics are in favor for which combination of criteria, we apply clustering technique to the Pareto frontier. The clustering tool we use is RapidMiner [16]; the clustering algorithm is DBScan with Euclidian Distance for similarity measure. The main reason we chose to use DBScan is that it is a density-based algorithm, in which clusters are regions in the data space densely occupied with data points and are separated from other clusters by regions of low density. Therefore, DBScan meets our goal of clustering the Pareto frontier in terms of the  $n$ -dimensional space of the  $n$  criteria. Detailed discussion about the suitability of using DBScan for our clustering can be found in [17].

One advantage of applying the Pareto-optimal approach is that more data points for each heuristic over a data set can be used for searching the optimal solutions, as opposed to one data point only for each heuristic in ELECTRE. Consequently, it is likely to include more candidate alternatives for consideration.

## 4. DECISION ANALYSIS USING ELECTRE

### 4.1 Process for analysis using ELECTRE-IS

(1) Prepare data set: For each data set, four heuristics H0, H1, H3, and H4 with their corresponding values of three criteria, *MMRE*, *Pred(0.25)*, and *Strength* are used for outranking using the ELECTRE-IS tool [18]. In order for the comparison of weighting heuristics across all the data sets, the first estimate at full strength (*Strength* =1) in the accuracy distribution database [3] is used for each data set. Because *Strength* =1 holds for all the data sets, this criterion is omitted in all the results in this analysis.

(2) Set parameters: There are three major parameters to be defined: thresholds of Level of Confidence (LC), Indifference  $q$  and Preference  $p$ . For simplicity, default values of  $p$  and  $q$  are used.

As LC is allowed to be varied in [0.5, 0.7], for the six data sets, we tested LC using 0.5, 0.6, and 0.7 respectively. The same results were obtained when LC=0.6 or 0.7; indifferent outranking relations were obtained between five out of six heuristics when LC=0.5. Therefore, LC=0.6 is chosen for this study.

The weights of the three criteria *MMRE*, *Pred(0.25)* and *Strength* are 0.4, 0.3, and 0.3, as used in [3].

(3) Derive outranking relations using the ELECTRE-IS tool. The relations are presented in the form of the outranking graphs in Figure 1 to Figure 6 for the six data sets.

(4) Analyze the outranking relation and determine the optimal heuristics (the least out-ranked) or rank the heuristics for each data set.

### 4.2 Analysis results over each data set

The outranking graphs and corresponding analysis data used for ranking the heuristics for each data set are presented in Figure 1 to Figure 6 respectively. In the directed graph,  $1 \rightarrow 2$  means node 1 out-performs node 2;  $1 \leftrightarrow 2$  means node 1 and 2 are equivalent. In the following graphs, "Cfs" stands for CfsSubset and "Wp" for Wrapper.

Because of the equivalent relation, we will present a partial-ordering of the involving heuristics in the sense that the less a heuristic is out-performed the better it is.

Interestingly, although it is well-known method, Wrapper selected no attributes for five out six data sets. Therefore, only Desh89 list the results of using Wrapper.

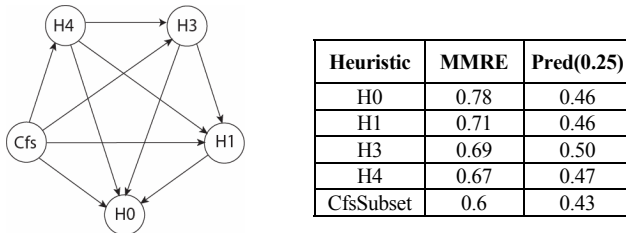


Figure 1. Outranking graph and analysis data for USP05-RQ

In Figure 1, H0 is out-performed by all the other heuristics, while CfsSubset out-performs all other heuristics. The partial-ordering of the heuristics according to the graph is CfsSubset, H4, H3, H1, H0.

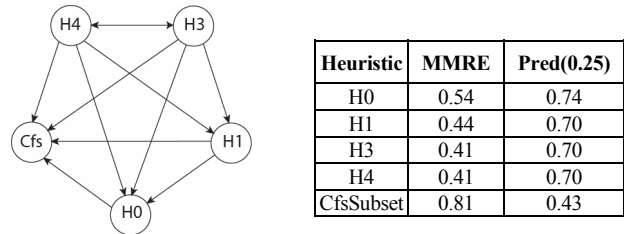


Figure 2. Outranking graph and analysis data for USP05-FT

In Figure 2, H0 is out-performed by the other three RSA-based heuristics. H3 and H4 are equivalent, while CfsSubset performs the worst. The partial-ordering of the heuristics is H4 and H3, H1, H0, CfsSubset.

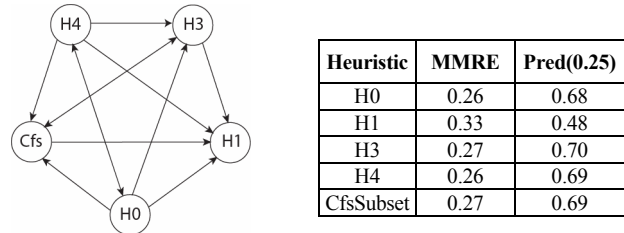


Figure 3. Outranking graph and analysis data for ISBSG04-2

In Figure 3, H1 is out-performed by other three heuristics. H0 and H4 are equivalent, so are H3 and CfsSubset. The partial-ordering of the heuristics is H4 and H0, H3 and CfsSubset, H1.

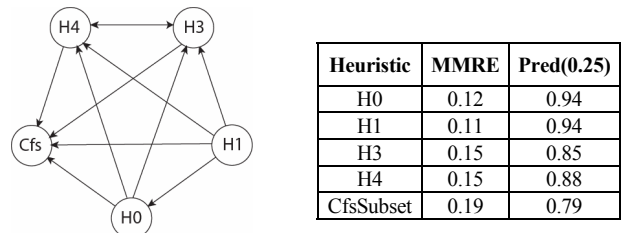


Figure 4. Outranking graph and analysis data for Mends03

In Figure 4, H3 and H4 are equivalent and are out-performed by H0 and H1. CfsSubset is outperformed by all other heuristics. The partial-ordering of the heuristics is H1, H0, H3 and H4, CfsSubset.

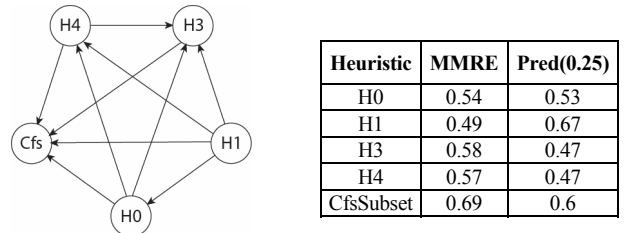
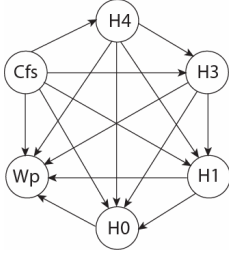


Figure 5. Outranking graph and analysis for Kem87

In Figure 5, CfsSubset is out-performed by all the other heuristics and H1 performs the best. The partial-ordering of the heuristics is H1, H0, H4, H3, CfsSubset.



Heuristic	MMRE	Pred(0.25)
H0	0.62	0.44
H1	0.61	0.44
H3	0.6	0.42
H4	0.59	0.42
CfsSubset	0.52	0.4
Wrapper	0.66	0.43

**Figure 6. Outranking graph and analysis data for Desh89**

In Figure 6, Wrapper is out-performed by all the other heuristics and CfsSubset performs the best. The partial-ordering of the heuristics is CfsSubset, H4, H3, H1, H0, Wrapper.

We obtain the following three major observations from the above analysis using ELECTRE-IS:

- (1) In five out of the six data sets, the baseline heuristic H0 is always out-performed by at least one RSA based heuristic H1, H3, or H4; and H0 is equivalent to H4 in ISBSG04-2. This result is consistent with the observations obtained in [5] that the RSA based heuristics perform better than the baseline heuristic H0 in general.
- (2) Regarding the four RSA based heuristics, the results from the first five data sets from Figure 1 to Figure 5 are: H3 or H4 perform well in USP05-RQ and USP05-FT, H1 performs best in Kem87 and Mends03. CfsSubset performs best in USP05-RQ and Desh89, but almost the worst in other data sets. While Wrapper could not select any attributes for the first five data sets, it performs the worst for Desh89 for which Wrapper produced a subset of attributes.
- (3) Quite different rankings of heuristics are obtained from different data sets. This evidence again supports the point of customization of EBA to suit different data sets.

## 5. PARETO ANALYSIS

### 5.1 Data set and weighting heuristics

Due to space limitation, only data set Desh89 is used for Pareto-optimal analysis in the current paper. All the six heuristics H0, H1, H3, H4, CfsSubset, and Wrapper are investigated. This demonstrates the principal contribution we can expect from this kind of analysis.

### 5.2 Process for analysis

We first present the results of using just the two criteria  $1-Pred(0.25)$  and  $1-Strength$ . Later, the results for three criteria  $MMRE$ ,  $1-Pred(0.25)$  and  $1-Strength$  are given. The process of analysis is consists of three major steps:

- (1) Prepare data points for each heuristic over the given data set, i.e., Desh89 in the current analysis.

The top fifteen data points for each of the six heuristics over the data set Desh89 have been selected, resulting in 90 data points in total. The ranking of the data points is based on a lexicographical ordering of all the data points in the Learned Accuracy Distribution database [5]; the first ranked is best in terms of smaller,  $1-Pred$ , and  $1- Strength$ . More data points may be used, but those low-ranked data points are very unlikely to be at the Pareto frontier, as they are out-performed by high-ranked points.

- (2) Calculating the Pareto frontier

The Pareto frontier is calculated for the cases of two and three criteria through a tool developed by the authors of the current paper.

- (3) Clustering at the Pareto frontier

The clustering algorithm DBScan is applied to obtain clusters of the points in the Pareto frontier in hopes to identify which heuristics are optimal in favor of which criteria. The following parameters are used in the RapidMiner tool:

Minimum Points=2, which means the minimum number of points in a cluster; the default value of the tool is 2.

Maximum Distance=0.1: which means the maximum Euclidian distance among the points in a cluster.

Other values of Maximum Distance were tested, but the resulting clusters contained either too many or too few points, leading to some extreme cases of clustering.

### 5.3 Analysis results

Table 2 presents the Pareto frontier, which has 8 points and three clusters of the points, for the case of Pareto-optimal analysis using two criteria " $1-Pred(0.25)$ " and " $1-Strength$ ". Figure 7 shows graphically the Pareto frontier in solid symbols and corresponding heuristics, while Figure 8 illustrates the clusters of the heuristics in the Pareto frontier. Here, cluster 0 always contains the points that cannot be clustered into any cluster according to parameters Minimum Points and Maximum Distance.

We can see from Table 2 and Figure 8 that the three heuristics (H0, H1, H4) in cluster 1 are in favor of  $1-Strength=0$ , while the two points (H0) in cluster 2 are relatively balanced between these two criteria. Although points in cluster 0 are not grouped into a proper cluster due to the Maximum Distance, we can still see that the three points (H1) are in favor of  $1-Pred(0.25)$ .

The basic conclusion from the above observations is that H1 favors  $1-Pred(0.25)$ ; and H0 is in some way balanced for both the two criteria. Therefore, the analysis of these three clusters may provide useful decision support for the selection of weighting heuristics with regard to the favor of the two criteria. Likewise, this analysis process can be applied to any other two criteria.

In the case of Pareto-optimal analysis using three criteria, 14 out of 90 points are belonging to the Pareto frontier. Five clusters are obtained using the same parameters as those in the two criteria analysis case. The points and corresponding clusters are presented in Table 3. The clusters are presented in Figure 9. Likewise, cluster 0 contains the points that cannot be clustered into any cluster according to the parameters.

**Table 2. Clusters of Pareto frontier of Desh89 with two criteria**

ID	$1-Pred(0.25)$	$1-Strength$	Heuristic	Cluster
19	0	0.95	H1	0
24	0.14	0.91	H1	0
26	0.21	0.83	H1	0
1	0.56	0	H0	1
28	0.53	0.05	H1	1
59	0.54	0.02	H4	1
12	0.31	0.64	H0	2
13	0.27	0.73	H0	2

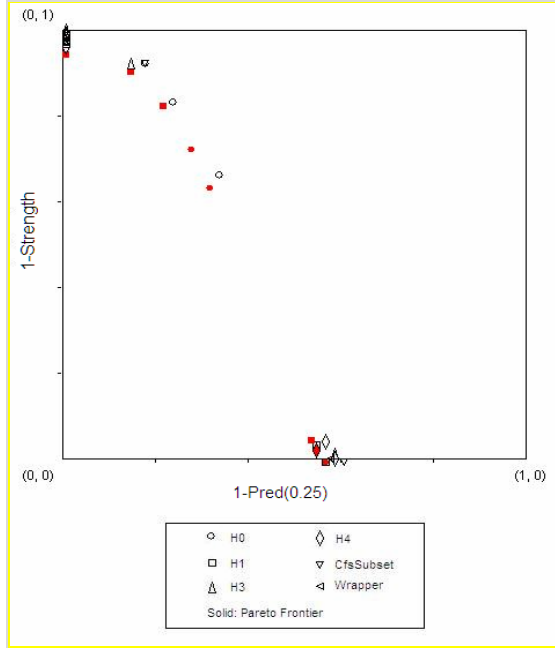


Figure 7. Pareto frontier of Desh89 with criteria 1-Pred(0.25) and 1-Strength

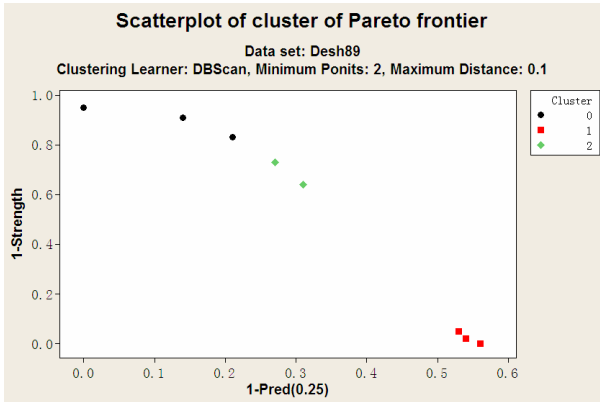


Figure 8. Clusters of Pareto frontier of Desh89 with criteria 1-Pred(0.25) and 1-Strength

Table 3. Clusters of Pareto frontier of Desh89 with three criteria

ID	MMRE	1-Pred(0.25)	1-Strength	Heuristic	Cluster
26	0.23	0.21	0.83	H1	0
8	0.11	0.17	0.93	H0	1
24	0.16	0.14	0.91	H1	1
12	0.27	0.31	0.64	H0	2
13	0.26	0.27	0.73	H0	2
16	0.61	0.56	0	H1	3
28	0.58	0.53	0.05	H1	3
46	0.59	0.58	0	H4	3
58	0.55	0.56	0.04	H4	3
59	0.58	0.54	0.02	H4	3
61	0.52	0.6	0	CfsSubset	3
17	0	0	0.99	H1	4
19	0.08	0	0.95	H1	4
62	0.02	0	0.98	CfsSubset	4

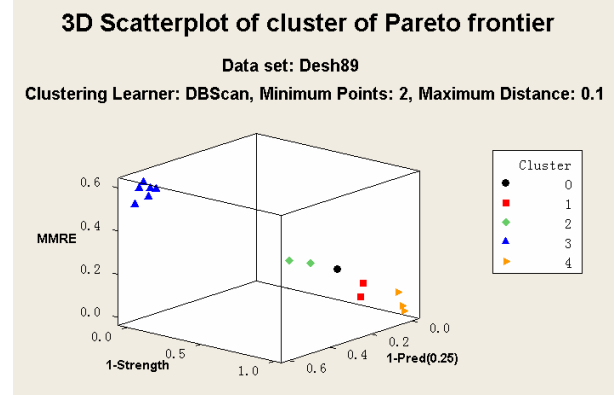


Figure 9. Clusters of Pareto frontier of Desh89 with criteria MMRE, 1-Pred(0.25) and 1-Strength

Cluster 3 and 4 in Figure 9 contain points with extreme small values in one or more of the criteria. Heuristics (H1, H4, CfsSubset) in cluster 3 are in favor of 1-Strength, while heuristics (H1, CfsSubset) in cluster 4 are in favor of both MMRE and 1-Pred(0.25). Heuristics (H0, H1) in cluster 1 and 2 are relatively balanced between the three criteria and favor 1-Pred(0.25) and 1-Strength.

It is clear that H0 is more balanced between the three criteria than the other heuristics. H4 favors 1-Strength while keeping other two criteria balanced. H1 and CfsSubset do not behave in a stable manner.

## 6. DISCUSSION OF RESULTS

### 6.1 Application of decision support techniques for choosing attribute weighing heuristics

The overall goal of the research is to provide a framework to qualify decision-making for selection of the "best" options of attribute weighing heuristic in the context of effort estimation with AQUA<sup>+</sup>. We present two application scenarios:

#### Scenario 1: Analysis of a single data set

Given a single data set for EBA, one can apply ELECTRE to determine an outranking relation between the heuristics. This outranking relation can provide support to the decision on which heuristic is preferable to the given data set. This information can be used for any new object under estimation using the data set.

Pareto-optimality provides the answer to the question of which heuristic is recommended in dependence of the relative importance of the criteria. We illustrate this point by the Pareto frontier shown in Figure 8. If strong emphasis is on the criterion "1-Pred(0.25)", then heuristic H1 of cluster 0 is the recommended choice. If "1-Strength" gets important as well, H0 of cluster 2 is recommended. If "1-Strength" becomes the most important criterion, all heuristics other than H2 and H3 seem to be appropriate in this case.

#### Scenario 2: Analysis from multiple data sets

After a certain number of data sets are investigated, empirical knowledge regarding which heuristic is suitable for which types of data sets can be acquired and reused. Figure 9 is used to demonstrate the idea. In Figure 9, we provide a classification of data sets, represented by bubbles, in terms of the percentage of

missing values, the percentage of non-quantitative attributes, and the number of objects (i.e. the size of the bubble). Three classes are created according to their spatial distance: Class I, Class II, and Class III. Based on our knowledge from previous empirical studies, corresponding heuristics suitable for the data sets in each class are indicated in the call-outs.

Given a new data set, it can be easily positioned in the two-dimensional space according to %Missing values and %Non-quantitative attributes. If it falls in or "close enough" to any of the existing classes, the heuristics that are suitable to this class may be recommended to the new data set; otherwise, empirical studies using the techniques in scenarios (1) should be conducted to investigate the suitable heuristics of the data set, which may be a new class to the existing ones. The knowledge base is thus updated with this knowledge regarding the new data set.

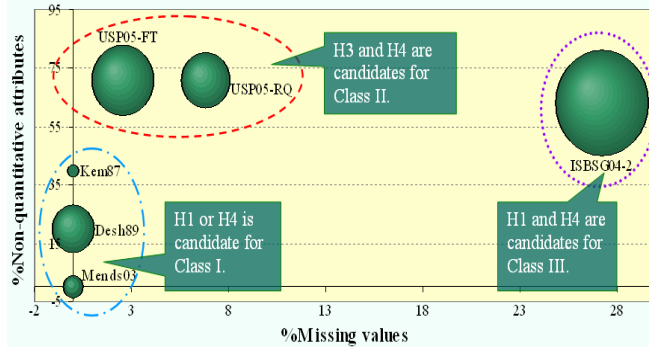


Figure 10. Classification of data sets and candidate heuristics

## 6.2 Applicability of ELECTRE and Pareto-optimality

ELECTRE works with only one data point regarding the involved criteria for each alternative. An outranking relation can be produced and presented in an outranking graph, based on which a partial-ordering of the heuristics can be obtained. Therefore, a heuristic can be easily selected according to the partial-ordering.

Through the above Pareto analysis process, in combination with clustering, decision support for the selection of weighting heuristics can be made with regard to the favor of corresponding criteria. Instead of a specific heuristic or a ranking of heuristics, this type of analysis normally recommends a group of heuristics. This is helpful in the sense that at least some heuristics are eliminated while revealing which heuristics are in favor of which criteria. With the reduced scope of alternatives, other means such as ELECTRE may be then applied to choose the optimal one.

## 6.3 Comparison with previous results

In our previous studies, only the baseline H0 and the RSA based heuristics were investigated. Therefore, only the four heuristics H0, H1, H3, and H4 will be compared here.

Table 4 presents the results obtained in [19] for the comparison among the RSA based four attribute weighting heuristics and the baseline heuristic H0, but using just a single aggregated criterion (*AccuH*). The numbers in the table represent the degrees of preference when comparing between the heuristics. The highlighted numbers in bold represent the "best" ones in the corresponding data sets. For example, H3 is most recommended for data sets USP05-RQ and USP05-FT, but H4 performs the second for these two data sets.

Table 4. Comparison of the four RSA-based heuristics over six data sets using AQUA<sup>+</sup>

( <i>AccuH</i> [ <i>i</i> ])	H0	H1	H2	H3	H4
USP05-RQ	0.22	0.42	-1.53	<b>0.52</b>	0.37
USP05-FT	-0.79	0.03	-	<b>0.62</b>	0.15
ISBSG04-2	0.16	<b>1.81</b>	-2.62	0.30	0.35
Mends03	-0.09	<b>0.15</b>	-	-0.05	-0.05
Kem87	-0.48	<b>1.42</b>	1.42	-0.47	-0.47
Desh89	-0.09	-0.05	-	0.03	<b>0.11</b>

By comparing Table 4 with the observations obtained at the end of section 4.1, it can be seen that the results for data sets Mends03, Kem87, and Desh89 are exactly the same: H1 performs the best. For USP05-RQ and USP05-FT, H3 and H4 are always the best two in both cases. The situation for ISBSG04-2 seems contradictory, i.e. H1 is best in Table 4, but the worst in Figure 3.

Therefore, we can say that the ELECTRE outranking relation is very close to the results obtained by the ranking of heuristics using a single aggregated criterion. On the other side, ELECTRE provides a more detailed view into the inherent preference structure between the alternative heuristics.

## 6.4 Limitations

The major limitation of the above two types of multi-criteria decision analysis is that appropriate parameters must be chosen to apply the ELECTRE, Pareto frontier, and clustering tools. Different combinations of the values of the parameters may lead to quite different results. Although the values of the parameters in the tools are not difficult to determine, there is no generally applicable method to decide about an optimal combination of the parameters. Choosing the appropriate values of the parameters of the tools mainly depends on personal judgment and experience.

Further limitations are caused by the small size in the number of data sets that has been considered. So there is no claim for external validity of the conclusions. The work is explorative in the sense that new methods have been demonstrated to provide more insight into the complex world of decision-making in the context of customization.

## 7. CONCLUSIONS AND FUTURE WORK

Making the right decisions for customization of (software) technologies is essential to exploit their full potential. While this is true in general, this paper looks at the predictor model technology AQUA<sup>+</sup> for estimation by analogy. The decision analysis for its better customization is done for the aspect of selection of the proper attribute weighting heuristic.

The contribution of the paper is preliminary and needs further investigation and empirical validation. The emphasis of the paper is not primarily on the accuracy of the results. Instead, we see the value of the paper in opening the door into this area of research, showing an approach how to do it, and provide initial (explorative) empirical validation.

ELECTRE provides an outranking relation among the alternatives regarding a set of criteria in the form of directed graphs. This relational model of decision-making is less rigorous than the traditional functional approach [7]. It supports strong involvement of the human experts and is less ambitious in the underlying assumptions. Search for the Pareto frontier is completely different. The respective solutions (in conjunction with the clustering) give some indication under which combination of importance level for

the different criteria a solution heuristic is more preferable than another.

With ELECTRE, partial-ordering of heuristics, even the optimal heuristic, can be determined easily based on the identified outranking relation. From the study in this paper over six data sets, the results about the optimal heuristic for a specific data set are very close to that of using a single aggregated criterion in our former studies. Therefore, ELECTRE is strongly recommended for decision analysis when there are only a small number of data points available for each alternative.

On the other hand, Pareto-optimality may be applied if a large set of data points are available for each alternative with regards to each of the multiple criteria, and there are also a great number of alternatives. With the Pareto frontier, only the non-dominated points will be considered, thus reducing the scope of alternatives. Furthermore, clustering techniques help to produce clusters of alternatives in regard to their favor of different criteria. This provides decision makers an opportunity to utilize their own preferences regarding the final decision.

The proposed Workshop contribution is a piece seen in a larger effort to look at decision-making under multiple objectives. Future research is intended to broaden the scope from EBA method AQUA<sup>+</sup> and its weighting attributes heuristics to other classes of decision and prediction problems. In addition, our future work will include the study of more weighting heuristics over additional available data sets. In the same way, other aspects of customization (such as the ones discussed in [4]) are planned to be investigated.

## Acknowledgements

The authors would like to thank the Alberta Informatics Circle of Research Excellence (iCORE) for its financial support of this research. Thanks are also given to Jim McElroy for his comments to improve the readability of the paper. Special thanks are given to the anonymous reviewers for their in-depth comments.

## REFERENCES

- [1] Mukhopadhyay, T., Vicinanza, S., and Prietula, M.J., "Examining the Feasibility of a Case-based Reasoning Model for Software Effort Estimation", *MIS Quarterly*, Vol. 16, No. 2, 1992, pp 155-171.
- [2] Shepperd, M., Schofield, C., "Estimating Software Project Effort Using Analogies", *IEEE Transactions on Software Engineering*, Vol. 23, No. 12, 1997, pp 736-743.
- [3] Li, J.Z., Ruhe, G., Al-Emran, A., and Richter, M.M., "A Flexible Method for Effort Estimation by Analogy", *Empirical Software Engineering*, Vol. 12, No. 1, 2007, pp 65-106.
- [4] Li, J.Z., Ruhe, G., "Decision Support Analysis for Software Effort Estimation by Analogy", *Proceedings of ICSE 2007 Workshop on Predictor Models in Software Engineering (PROMISE'07)*, USA, May 2007.
- [5] Li, J.Z., Ruhe, G., "A Comparative Study of Attribute Weighting Heuristics for Effort Estimation by Analogy", *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering (ISESE'06)*, September 2006, Brazil.
- [6] Conte, S.D., Dunsmore, H., and Shen, V.Y., *Software engineering metrics and models*, Benjamin-Cummings Publishing Co. Inc., 1986.
- [7] Figueira, J., Mousseau, V., and Roy, B., "ELECTRE methods", In: Figueira, J., Greco, S., and Ehrgott, M., (Eds.), *Multiple Criteria Decision Analysis: State of the Art Surveys*, Springer, New York, 2005, pp 133-162.
- [8] Ehrgott, M., *Multi-criteria Optimization*, Springer 2005.
- [9] Efron, B., and Gong, G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation, *The American Statistician*. 37(1983): 36-48.
- [10] ISBSG, Data R8, *International Software Benchmark and Standards Group*, www.isbsg.org, October 18, 2005.
- [11] Kemerer, C.F., "An Empirical Validation of Software Cost Estimation Models", *Communication of the ACM*, Vol. 30, No. 5, 1987, pp 436-445.
- [12] Menzies, T., Chen, Z.H., J. Hihn, and K. Lum, "Selecting Best Practices for Effort Estimation", *IEEE Transactions on Software Engineering*, Vol. 32, No. 11, 2006, pp 1-13.
- [13] Boetticher, G., Menzies, T., Ostrand, T., "PROMISE Repository of Empirical Software Engineering Data", *West Virginia University, Department of Computer Science*, 2008, available at: <http://promisedata.org/?cat=11>.
- [14] Pawlak, Z., *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, 1991.
- [15] Witten, I.H., Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques (2<sup>nd</sup> Edition)*, USA: Morgan Kaufmann Publishers, 2005.
- [16] RapidMiner, available at: [www.rapidminer.com](http://www.rapidminer.com)
- [17] Ullah, M.I., and Ruhe, G., "One product versus product line: Decision support based on customer needs analysis", *Doctoral Symposium at 11<sup>th</sup> International Software Product Line Conference*, 10 - 14 September 2007, Kyoto Japan.
- [18] ELECTRE IS, available at: <http://www.lamsade.dauphine.fr/english/software.html>
- [19] Li, J.Z., Ruhe, G., "Analysis of Attribute Weighting Heuristics for Analogy-Based Software Effort Estimation Method AQUA<sup>+</sup>", *Empirical Software Engineering*, Vol. 13, No. 1, 2008, pp 63-96.