Optimising Project Feature Weights for Analogy-based Software Cost Estimation using the Mantel Correlation

Jacky W. Keung and Barbara Kitchenham
NICTA Ltd., Sydney, Australia
{Jacky.Keung, Barbara.Kitchenham}@nicta.com.au

Abstract

Software cost estimation using analogy is an important area in software engineering research. Previous research has demonstrated that analogy is a viable alternative to other conventional estimation methods in terms of predictive accuracy.

One of the important research areas for analogy is how to determine suitable project feature weights. This can be achieved by using an extensive project feature weights search, where the quality measure is optimised. However, this approach suffers similar issues as the brute-force feature selection approach in analogy.

We propose a novel method to deal with this issue based upon the use of the Mantel randomisation test. Specifically, we determine project feature weights based on the strength of correlation between the distance matrix of project features and the distance matrix of known effort values of the dataset.

We demonstrate the procedure on a specific dataset, showing the use of the Mantel correlation to identify whether analogy is appropriate, and whether the project feature weights can be determined by statistical inference. Our results also show improved prediction accuracy when multiple project features are used with determined weights. Our method, thus, provides a sound statistical basis for analogy.

1. Introduction

Data-intensive analogy for software effort estimation has been proposed as a viable alternative to other prediction methods such as linear regression. In many cases, researchers found analogy outperformed algorithmic methods [1]. However, the overall performance of analogy depends on a number of factors that influence prediction accuracy of analogy. They include the dataset quality, the relevance of the project cases to the target project, the feature subset selection technique, and the distance measures used to assess similarity between the source analogues and the target project [2].

Analogy-based systems such as ANGEL [1] do not consider each project feature's influence with respect to target software effort. Auer and Biffl [3] consider the importance of the project feature weights used to determine analogous projects, which will have the overall estimation accuracy and reliability impacts. However their method uses an extensive search approach similar to that of Shepperd and Schofield's brute-force approach [1] for feature subset selection, which inherits similar issues for analogy.

In this paper, we suggest that using Mantel's correlation and randomisation test, it is possible to determine a set of project feature weights which can be used to account for the influence of each project feature based on inferential statistics rather than brute-force. Also, our method can determine whether the dataset is indeed appropriate for analogy-based estimation, and detect extremely outlying cases that will ultimately distort prediction outcomes using a sensitivity analysis strategy.

Results show that our method solves some of the fundamental issues of analogy, which feature weights now can be determined statistically, and more importantly, that prediction performance improves.

Section 2 presents an overview of related work. Section 3 describes our method and the underlying theory. Section 4 provides the datasets and the analysis procedure. Section 5 presents the results, and section 6 discusses the results and provides further directions. Section 7 concludes the paper.

2. Background and Related Work

The analogy framework is based on studies by Shepperd et al. reported in the late 90's, where a data-intensive analogy-based reasoning approach (a.k.a case-based reasoning) was successfully established for the purpose of software cost estimation[1, 4]. Estimation by analogy is similar to expert judgment [5] in a way that it relies on a comparison and adjustment between previously completed projects and the proposed project. Unlike the human-based expert

1530-1362/07 $25.00 © 2007 IEEE
DOI 10.1109/ASPEC.2007.42

judgment approach, analogy is actually a data-intensive approach that heavily relies on past project data.

For the past 10 years, analogy-based software cost estimation has been considered as an alternative approach to regression-based estimation methods with the potential for software effort estimation that is particularly useful in the early stages of the software life cycle. Similarly, in many circumstances, analogy provides an alternative to other data-intensive approaches, where sufficient data needed to fit a suitable model statistically is simple unavailable, but there is at least one analogous, or similar, project for which cost and schedule data is available.

The general principle of data-intensive analogy is to reuse software development experience in the form of past project data stored in a project case repository or a database, in a form that includes important project features of those projects from the point of view of their possible effect on development effort [1, 4].

When estimating a new project, the analogy-based method uses a number of steps. First, the estimator reviews the new project and then measures the project features of the target project. The analogy-based system then searches the project database for the projects that are most similar (in terms of project features) to the target project case and selects one or more similar projects as source analogue(s). Finally, it predicts the software effort for the target case in a sense that is adapted to the new problem. The effort value of the source analogues becomes an initial estimate for the target project.

The similarity between each potential source analogue and the target project is measured using a distance measure as the proximity in *n*-dimensional space, where each dimension corresponds to a different feature. An unweighted Euclidean distance measure is commonly used[1]. The unweighted Euclidean distance does not account for the impact different project features in the dataset.

An important research area in analogy has been the impact of different project features. Auer and Biffl [3] report on the importance of the project feature weights used to determine analogous projects, which will have the overall estimation accuracy and reliability impacts and was not adequately addressed by traditional approaches. The different impact or weighting of a project's various features is a crucial aspect of the analogy-based method. The weighted Euclidean distance ∂ over the features $(d_1,...,d_n)$ and can be expressed as:

$$\partial(p, p') = \sqrt{\sum_{i=1}^{n} w_i (d_i - d'_i)^2} \qquad (1)$$

w_i is the weight for its respective feature i for n projects. Based on the above equation, a small distance indicates a high degree of similarity. The most similar projects can be then used as source analogues for the new cost estimate.

The question in here is how to determine the values of w_i, because they have direct influence on the overall similarity measure ∂. These values are usually determined by experts according to their experience, or using the method developed by Shepperd and Schofield [1], where each project feature weight is set to either 0 or 1 so that an estimation quality metric such as MMRE can be optimised using a brute force approach to select only relevant project features. For example, if feature 1 is selected then w_1 is set to 1, if feature 2 is not selected then w_2 is set to 0 and so on.

Auer and his colleagues [6] have argued that searching for a subset of important features fails to account for each feature's individual influence on project similarity, and for the volatility of the resulting feature weights over the lifetime of a growing project database, which makes analogy-based method less acceptable. They have proposed a brute-force approach for weighting project feature dimensions for analogy, which has been implemented as a Java command line tool named AMBER to facilitate batch processing.

The basic principle of AMBER's extensive feature weighting approach is much alike the brute-force feature selection algorithm in the original ANGEL tool developed by Shepperd and Schofield [1]. AMBER extensively select the optimal subset of features with respect to overall estimation performance measures, that is, each feature's weight is optimized based on MMRE or other quality measures. Their three main results can be summarized as [6]:

1. Optimal feature weights approach improves estimation accuracy and reliability.
2. Improves volatility leading to greater acceptance by practitioners
3. Obtained estimates represent upper limits of analogy-based estimation quality as measured by standard metrics

Nevertheless, Auer's brute-force approach has a number of drawbacks similar to that of ANGEL. First, it is computationally expensive. Although it is claimed that model calibration only takes place when the historical feature database is updated and once completed, estimates are obtained in real-time. Second, Auer's brute-force approach is similar to the optimal feature selection developed in [1], and it inherits similar issues identified in [7]. There is no method to measure the appropriateness of the analogy approach for a specific dataset. That is, a target estimate will be produced regardless of the usefulness of the database. Also there is no method to identify abnormal projects. Therefore the quality of the dataset and its relevance are not considered. This introduces the potential risk of

misleading estimates as a result of spurious effects. In statistics, a spurious relationship is a mathematical relationship in which two occurrences have no causal connection, yet it may be inferred that they do, due to a certain third, unseen factor, often referred as a confounding factor [8]. Auer and Biffl [3] also admitted that the ability of removing outliers in the portfolio is not readily available in their approach.

This paper extends Auer et al.'s extensive feature weight search approach [6] and Shepperd and Schofield's analogy algorithm [1]. It proposes to use an alternative algorithm to brute-force search to account for individual influences of project features based on a statistical relationship between project feature similarities. The paper also evaluates our approach on a real-world project dataset, and compares its performance with the conventional analogy approach.

3. Method

Although usually unstated, the fundamental hypothesis underlying the use of data-intensive analogy-based reasoning for software project effort estimation is:

"Projects that are similar with respect to project and product factors such as size and complexities will be similar with respect to project effort."

Based on this assumption, analogy-based tools such as ANGEL, compute a similarity measure using project and product features between a new target project and projects in an historical database. An effort estimate for the new project is then based on the actual effort of the k most similar projects in the database [1]. The similarity between projects is determined by unweighted Euclidean distance measure.

Similar to the approach of Auer et al. [6], where it allocates separate feature weights w_i to the n project features d_i(see Equation 1). Our approach is based on the identification on well-defined statistical techniques to test the hypothesis of analogy, and based on the strength of the relationship between each feature and the target feature similarity to determine suitable project feature weights w_i. To statistically test this hypothesis, a measure of correlation or relationship with its significance, between the distance matrix based on project features and the distance matrix based on *ActualEffort* is required.

3.1 Weighting Project Features Using Mantel's Correlation

Mantel's correlation for comparing two distance or dissimilarity matrices was first introduced as a solution to the problem of detecting space and time clustering of diseases for cancer research [9]. It has since been widely adopted in ecology, biology and psychology researches to address this kind of problem [10].

A classical example in ecology is attempting to explain the distribution of species based on constraints of their environmental variables. The operative question in these ecology experiments is: "Do samples that are close with respect to *X*s (environmental variables) also tend to be close with respect to *Y*s (species variables)?" The question is analogous to the questions we want to ask in analogy-based software cost estimation approach i.e. "Do projects that are close with respect to *X*s (project and product features) also tend to be close to *Y*s (development effort)?"

Although Mantel discussed more general situations and findings in his original study, Manly provides more comprehensive examples of Mantel's method [10], [11]. The basic principle of Mantel's method is to measure the association between the corresponding elements in two distance matrices by a suitable statistic, usually the Pearson correlation. The significance of the correlation is then determined by a permutation procedure in which the original value of the test statistic is compared with the distribution of the statistics found by randomly re-ordering the elements in one of the distance matrices. The normal statistical tests for the Pearson correlation coefficient are inappropriate in this case because the elements in a distance matrix are not independent.

Distance matrices A for predictor variables and B for response variables are constructed as follows:

$$A=\begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix} \quad B=\begin{bmatrix} 0 & b_{12} & \cdots & b_{1n} \\ b_{21} & 0 & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & 0 \end{bmatrix} \quad (2)$$

The distance matrix is a matrix of n cases (e.g. projects). Each case has a distance measure constructed from p features (variables). Thus, for example, the distance element between case 1 ($x1$) and case 2 ($x2$) is calculated using the simple Euclidean distance:

$$a_{21}=\sqrt{\sum_{i=1}^{p}(x1_i-x2_i)^2} \quad (3)$$

Equation (3) considers the values of all p variables for each pair of cases. Note that before the diagonal elements can be constructed the variables have to be standardized by transformation so that they are all equally weighted and comparable. The usual transformation is to divide each value by the difference between the maximum and minimum value.

Because of its symmetrical nature, only the lower diagonal elements in the above matrices (Equation 2) need to be considered when constructing and testing the Mantel correlation. The Mantel correlation coefficient is:

$$R_M = \frac{\sum a_{ij} b_{ij} - \sum a_{ij} \sum \frac{b_{ij}}{m}}{\sqrt{\left[\left\{\frac{\sum a_{ij}^2 - \left(\sum a_{ij}\right)^2}{m}\right\} \times \left\{\frac{\sum b_{ij}^2 - \left(\sum b_{ij}\right)^2}{m}\right\}\right]}} \quad (4)$$

Where m is the number of diagonal elements in the distance matrix and is given by:

$$m = \frac{n(n-1)}{2} \quad (5)$$

For the randomisation test the distance matrix elements randomly permuted for one of the matrices, matrix A (Equation 2) say. A random order matrix A_{Random} (Equation 6) can be constructed based on the random ordering of elements. For example one randomisation of the elements of A, gives the matrix A_{Random}:

$$A_{Random} = \begin{bmatrix} 0 & a_{68} & a_{18} & \cdots & a_{38} \\ a_{68} & 0 & a_{16} & \cdots & a_{36} \\ a_{18} & a_{16} & 0 & \cdots & a_{31} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{38} & a_{36} & a_{31} & \cdots & 0 \end{bmatrix} \quad (6)$$

The entry in column 1, row 2 is the distance between data items 8 and 6; the entry in column 2, row 3 is the distance between data items 6 and 1 and so on. The value of the Mantel correlation is then computed using matrix B (Equation 2) and A_{Random} (Equation 6).

Repeating the same procedure many times produces the randomisation statistic distribution. Using this randomisation distribution we can test whether the value of the Mantel correlation derived from the original pair of distance matrices is significantly different from zero. If the Mantel correlation is significantly different from zero, we can be sure that projects that are close together with respect to project features are close together with respect to effort and that analogy-based estimation is an appropriate method for the dataset under investigation.

Based on Marriott [12], Manly [10] notes that it is generally practical to determine the full randomisation distribution using 1,000 randomisations, because this value is a realistic minimum for estimating a significance level of about 0.05 and 5,000 randomisations is a realistic minimum for estimating a significance level of about 0.01.

This procedure is critical as it removes the independence problem that violates the basic principle of correlation tests, and this makes the Mantel correlation coefficient a suitable measure of relevance for assessing the relationship between distance matrices. And more importantly, based on the fundamental hypothesis of analogy, suitable feature weights can be determined using the Mantel correlation coefficients.

3.2 Sensitivity Analysis - Outlier Detection

It is important to realize that the analogy-based method has no means of assessing the datasets quality and will always endeavour to predict no matter under what circumstances. The procedure described in this section avoids the problem of always predicting even if the dataset is inappropriate while also providing a mechanism to detect outlying project cases in the dataset using a sensitivity analysis procedure.

We have developed the Mantel Leverage Metric [7] to support sensitivity analysis for analogy, based on the same principle as the Jackknife method [13] and the properties of the standard normal distribution. The principle of Mantel Leverage metric (LM) is based on calculating the Mantel correlation excluding each (project) in run. This indicates the extend to which the Mantel correlation for the complete dataset is influenced by each individual case.

We use the Jackknife method to provide an unbiased estimator for the Mantel R. The Jackknife estimator $\hat{R}$ of Mantel R can be calculated as:

$$\hat{R} = \frac{\sum_{i=1}^{n} R_i}{n} \quad (7)$$

Where n is the total number of cases, and R_i is the Mantel correlation of all cases excluding i^{th} project case in turn. The Jackknife estimator $\hat{R}$ will be normally distributed (approximately) with an unknown variance S_M^2, which can be estimated as:

$$S_M^2 = \frac{\sum_{i=1}^{n} (R_i - \hat{R})^2}{n-1} \quad (8)$$

To calculate the leverage metric (LM_i) for each case i, let R_i be the Mantel correlation for the dataset excluding case i, and $\hat{R}$ be the Jackknife estimator of overall Mantel's R, then

$$LM_i = R_i - \hat{R} \quad (9)$$

LM_i is the difference (residual) between R_i and the overall Mantel estimator $\hat{R}$, indicating the impact of the specific case i on the overall correlation. Under the null hypothesis that case i is NOT abnormal, R_i will be an unbiased estimator of $\hat{R}$ will be approximately $N(0, S^2)$. The following z test provides a mechanism to formally verify whether the value of R_i is an abnormal one. For each case i, LM_i can be converted to its standard normal form:

$$z_i = \frac{LM_i}{S_M} \quad (10)$$

If $|z_i|$ is greater than 2 the case significantly deviates from 0 at the 0.05 (approximately) significance level. If $|z_i|$ is greater than 4, the significance level is 0.001 (approximately).

If the relationship between the distance matrices is resilient to the removal of the abnormal case(s) measured by the *p*-value of Mantel's correlation, and we can be confident that analogy is appropriate for the reduced dataset, and the strength of the correlation can be used to indicate the explanatory power of analogy model for the dataset.

4. Dataset and Analysis Procedure

The Desharnais dataset is used in this study. This dataset comprises 77 completed software project data from a Canadian Software house. It was first reported in Desharnais [14] and was used in Shepperd and Schofield [1] to compare regression models and analogy, and in [6] to analyse the different impacts of project feature weights using an extensive search algorithm. The Desharnais dataset is one of the most well-known and complete datasets publicly available in software effort estimation research. The original version of the dataset had 81 projects but four of the projects had missing values and were excluded from our analysis [14]. This dataset has 8 independent variables, which are shown in Table 1. The response variable is *ActualEffort.*

Proj. Feature	Description	Type
Adj.FPs	Adjusted Function Points	Continuous
RawFPs	Raw Function Points	Continuous
Transactions	No. of Transactions	Continuous
Entities	No. of Entities	Continuous
Adj.Factor	Technology Adj. Factor	Continuous
ExpProjMan	Experience of Proj. Mgmt.	Continuous
ExpEquip	Experience of Equipment	Continuous
Dev.Env	Development Environment	Categorical

Table 1 Desharnais Dataset Project Features

The Desharnais dataset is then divided on the basis of differing development environments, because it is unlikely that a company would have access to such large volumes of data, and because smaller, more homogenous datasets are more useful for effort estimation. This dataset grouping approach is similar to the study of Shepperd and Schofield [1], where the Desharnais dataset is divided into Desharnais-1(44 cases), Desharnais-2(23 cases) and Desharnais-3(10 cases). This leaves 7 usable independent project features excluding *Dev.Env* (see Table 1) in each homogenous subset.

An essential issue in analogy is to identify which features are important for the purpose of case retrieval. Mantel's method allows us to show the relevance of each project feature to our target feature software effort. We then apply weights derived from the Mantel correlation tests to these relevant project features. The procedure is as follow:

1. Identify project features that are significant using Mantel's correlation, where confidence is 95% or p-value < 0.05.
2. Apply our sensitivity analysis on the selected project features to determine any outlying data points.
3. Construct a weighted distance matrix for the dataset based on the Mantel correlation coefficients.
4. Use the Jackknife validation approach to obtain quality measures for the predictions.
5. Compare the result with the control group (conventional method) where weights are not applied to the features selected in step 1.

The question in here is whether the inclusion of project feature weights based on the Mantel correlation coefficients will improve estimation accuracy. We use conventional quality measures, such as the mean magnitude of relative error (MMRE), and Pred(25). The MMRE is given by the following equation:

$$MMRE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|X_i - \hat{X}_i\right|}{X_i} \qquad (11)$$

Pred(25) is defined as the percentage of predictions falling within 25% of the actual known value. X_i is the actual effort and $\hat{X}$ is the predicted effort, given that there are *n* project cases to evaluate.

5. Results

In this section, we demonstrate our method on the Deshnarnais77 dataset described in the previous section.

5.1 Identify Useful Project Features and Dataset Sensitivity Analysis

Applying the Mantel correlation on each project feature to the Desharnais-1 dataset, we found the RawFPs distance matrix, the *Adj.FPs* distance matrix, the *Transactions* distance matrix, and the *Entities* distance matrix were significantly correlated with the *ActualEffort* distance matrix (see Table 2).

Feature	Mantel-R	p-value	Use
RawFPs	0.705	0.001	YES
Adj.FPs	0.697	0.001	YES
Transactions	0.648	0.002	YES
Entities	0.248	0.037	YES
Adj.Factor	0.113	0.114	NO
ExpEquip	-0.033	0.729	NO
ExpProjMan	-0.021	0.558	NO

Table 2: Mantel correlation for each project feature separately on the Desharnais-1 dataset (n=44).

To ensure the variables selected are not an artefact of any abnormal case, we use a sensitivity analysis as outlined in Section 3.2 to identify abnormal cases based on all 44 cases and with selected features:

RawFPs, AdjFPs, Transactions, and Entities, these features are used to construct a distance matrix to correlate with the distance matrix of *ActualEffort*. The Mantel correlation based on these two variables is 0.645, and its p-value is 0.001.

The sensitivity analysis is continued, by correlating all cases excluding the ith project case in turn, denoted "**Mantel-Ri**" in Table 3. The overall Jackknife estimator $\hat{R}$ of Mantel-R is 0.641 (normally distributed).

Table 3 Sensitivity Analysis based on selected features in Table 2 for Desharnais-1 Dataset ($\hat{R}$=0.641 *n=44*)

Case	Mantel-R	Mantel-R$_i$	p-value$_i$	LM$_i$	|z$_i$|
44	0.645	0.277	0.013	-0.364	6.299
38	0.645	0.700	0.001	0.059	1.025
23	0.645	0.696	0.001	0.055	0.955
25	0.645	0.677	0.001	0.036	0.629
33	0.645	0.669	0.001	0.028	0.495
…	…	…	…	…	…

Table 3 shows the largest 5 leverage statistics for the Desharnais-1 dataset. It also shows project case 44 is an extremely influential data point in the Desharnais-1 dataset. The exclusion of case 44 causes the Mantel correlation to be reduced from 0.645 to 0.277 (R_{-44}), although the Mantel correlation is still significantly greater than zero (p-value=0.01). This clearly demonstrated the effect of spurious correlation, which will have an impact on the overall estimation accuracy. Although the results are strongly influenced by the specific data value, that there is an underlying predictive relationship that can be used for analogy-based estimation.

The impact of case 44 is further illustrated in Figure 1. It is clear that case 44 in the Desharnais-1 dataset has a significant impact on the distance measures of project features and target feature *ActualEffort*.

Figure 1 Distance Matrix Correlation, before/after the removal of case 44.

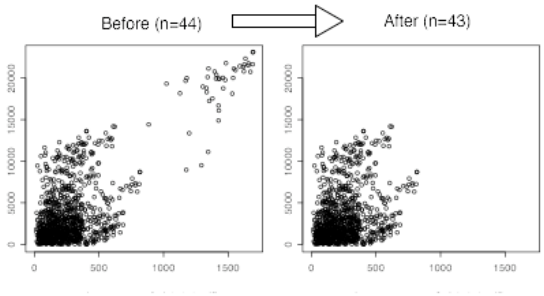

Figure 1 shows the distance matrix correlation plot between project feature similarities and target feature *ActualEffort* similarities. The removal of case 44 in the dataset causes all distance measures associated with case 44 to be removed, which is illustrated on the right-hand side of Figure 1. Note that project case 44 in the homogenised Desharnais-1 dataset is equivalent to project case 77 in the original Desharnaise77 dataset.

This sensitivity analysis will reduce the likelihood of building spurious prediction systems in a manner analogous to a sensitivity analysis used for evaluating regression results. Next we remove project case 44 to ensure the dataset is stable (free of extreme cases).

Feature	Mantel-R	p-value	Use
RawFPs	0.291	0.008	YES
Adj.FPs	0.286	0.006	YES
Adj.Factor	0.235	0.021	YES
Entities	0.188	0.054	NO
Transactions	0.124	0.141	NO
ExpEquip	-0.062	0.878	NO
ExpProjMan	-0.059	0.823	NO

Table 4: Mantel correlation for each project feature separately on the Desharnais-1 dataset (n=43, case 44 removed).

Repeating the same procedure above for the reduced Desharnais-1 dataset shows that only project features *RawFPs*, *Adj.FPs* and *Adj.Factor* have a significant relationship with target feature *ActualEffort* (based on similarity measures), given that their p-values are still significant (p-value<0.05). We will use these three selected project features to investigate the Desharnais-1 dataset.

Feature	Mantel-R	p-value	Use
Adj.FPs	0.781	0.001	YES
RawFPs	0.730	0.001	YES
Transactions	0.636	0.001	YES
Adj.Factor	0.327	0.014	YES
Entities	0.272	0.026	YES
ExpEquip	0.089	0.089	NO
ExpProjMan	0.043	0.339	NO

Table 5: Mantel correlation for each project feature separately on the Desharnais-2 dataset (n=23).

Subsequent investigations of the Desharnais-2 and Deshnaris-3 datasets show that they also have predictive relationships and can be used with analogy-based prediction methods.

Based on the Deshnaris-2 dataset, Table 5 shows that features *Adj.FPs*, *RawFPs* and *Transactions* express strong correlation with target feature *ActualEffort*. Features *Adj.Factors* and *Entities* also have some degrees of relationship with *ActualEffort,* given that the p-values < 0.05. Sensitivity Analysis applied on the Desharnais-2 dataset, in this case, finds no extreme outlying case.

Feature	Mantel-R	p-value	Use
Adj.FPs	0.616	0.006	YES
Entities	0.503	0.006	YES
RawFPs	0.470	0.052	NO
Transactions	0.400	0.052	NO
ExpEquip	0.309	0.009	YES
Adj.Factor	0.132	0.097	NO
ExpProjMan	-0.028	0.685	NO

Table 6: Mantel correlation for each project feature separately on the Desharnais-3 dataset (n=10).

The last column shown in Table 4, Table 5 and Table 6 shows the project feature selected for that particular dataset.

Similar to the Desharnais-2 dataset, Desharnais-3 (see Table 6) also exhibits a stable predictive relationship with no extremely outlying case, and only three project features are included in the feature vector for effort estimation.

5.2 Weighting Project Features and Effort Estimation

We first construct a weighted distance matrix for each of three datasets (Desharnais-1, Desharnaise-2, and Desharniase-3). The weights are the strength of the correlation of each project feature identified in the previous analysis of each dataset (Table 4, and Table 5 and Table 6). Based on Mantel's distance matrix correlation, only project features are used that are significantly correlated with target feature *Act.Effort*.

Dataset (No. of features)	Weight = 1	Weight = Mantel-R	Difference
	MMRE	MMRE	MMRE -%
Desh1.(Fea.=3)	0.394	0.394	00.00
Desh2.(Fea.=5)	0.424	0.359	15.12
Desh3.(Fea.=3)	0.257	0.257	00.00
	Pred(25)	Pred(25)	Pred(25)+%
Desh1.(Fea.=3)	0.326	0.394	21.15
Desh2.(Fea.=5)	0.261	0.391	50.00
Desh3.(Fea.=3)	0.700	0.700	00.00

Table 7: MMRE, Pred(25) improvement in percent.

Table 7 compares the effort prediction results generated using two different approaches, showing improved prediction accuracy using the quality measures MMRE and Pred(25). The first approach is the conventional approach, which uses the identified relevant project features and treats them with equal weights (i.e. weight = 1 for each selected feature and weight = 0 otherwise). The second approach sets project feature weights with their respective Mantel correlation coefficient. Note that positive percentage value denotes improvement of an accuracy metric, i.e. the lower the MMRE the higher the prediction accuracy, and conversely higher the Pred(25) value the higher the prediction accuracy.

We use these two different accuracy metrics to show the prediction performance, because MMRE is one of the most commonly used quality metrics for this purpose. However, it has been subject to many criticisms for its unbalanced measure and penalises overestimates more than underestimates, for example in [15].

At first, by using the new approach to set feature weights for the Desharnais-1 dataset, the MMRE exhibited zero improvement, while using Pred(25) shows a 21.15% improvement in prediction accuracy. Desharnise-2 shows a 15.12% improvement based on the MMRE measure, against the 50% improvement shown by Pred(25). Both prediction accuracy measures for Desharnis-3 remain the same.

6. Discussions

We have successfully employed a new approach to identify optimal feature weights for analogy-based software cost estimation using Mantel's correlation and randomisation test that correlate project features and effort similarity.

Our results show that applying project feature weights is an effective strategy to account for the influence of each project feature's impact to the overall estimate. Prediction accuracy outperformed the conventional approach where weights are not applied. Although project feature weights identified by using extensive search seem more straightforward than our proposed approach, our approach is based on established statistical analysis techniques. The technique developed in this study would only be effective if more than one project feature is used. This is because the feature weight for a single project feature is not necessary.

Based on the Mantel statistics, we are also able to identify whether the analogy-based approach is appropriate for the dataset under investigation. This was not possible with the brute-force approach.

Estimation accuracy measures generated in this study may not be the same or comparable to other studies using the Desharnais dataset such as [1] [3] [6]. This is mainly due to that fact that the outlying data point (detected in this study) was not removed in other studies. Outlying data point detection was considered a problem. There was no simple way to achieve this. A spurious relationship was detected using our sensitivity analysis method on the Desharnais-1 dataset.

The feature weights identified in this case are based on statistical inference between project features and effort similarity, rather than on a trial-and-error approach on all combinations [3]. Computation for a large dataset that currently requires hours or even days can now be effectively reduced to a few seconds in most cases. This allows the analogy-based method to scale up to larger datasets.

Further investigation on other datasets may be required in the future. The impact of this method to different sized datasets is still unknown. There was no performance improvement on the Desharnais-3 dataset, which has only 10 project cases. The size of the dataset may be a contributing factor. The prediction accuracy used was based on the unreliable MMRE [15] commonly used in software cost estimation studies. This was used for the purpose of reference only, thus we can only speculate about the prediction performance improvement. Nevertheless, our method is based on established statistical theories, which guarantees the project analogues retrieved from the case base are statistically relevant to the target project.

7. Conclusion

The proposed approach is a set of comprehensive procedures that utilises the principles of the Mantel randomisation test to provide inferential statistics for analogy-based methods. Inspired by the Mantel method commonly used in ecology and psychology, our project feature weighting approach uses the strength of correlation between the distance matrix of project features and the distance matrix of known effort values of the dataset to (1) assess the suitability of the dataset for analogy, to (2) identify the most appropriate feature subset, to (3) remove an atypical project cases from the dataset, and (4) more importantly, to determine the most suitable feature weights for purpose of software effort estimation. Our results also indicate improved prediction performance using this approach.

Our method is thus a robust solution that provides a sound statistical basis for analogy. It is especially useful when feature weights are required to distinguish different influences of project features to the target problem. This is a major improvement to analogy-based software cost estimation.

8. Acknowledgements

NICTA (National ICT Australia Ltd.) is funded through the Australian Government's Backing Australia's Ability Initiative, in part through the Australian Research Council.

9. References

[1] M. J. Shepperd and C. Schofield, "Estimating Software Project Effort Using Analogies," *IEEE Transactions on Software Engineering,* vol. 23(12), pp. 736-743, 1997.

[2] F. Walkerden and R. Jeffery, "An Empirical Study of Analogy-based Software Effort Estimation," *Empirical Software Engineering,* vol. 4(2), pp. 135-158, June 1999.

[3] M. Auer and S. Biffl, "Increasing the accuracy and reliability of analogy-based cost estimation with extensive project feature dimension weighting," in *International Symposium on Empirical Software Engineering*, Redondo Beach, CA, 2004, pp. 147-155.

[4] M. J. Shepperd, C. Schofield, and B. Kitchenham, "Effort estimation using analogy," in *18th Intl. Conf. on Software Engineering*, Berlin, 1996.

[5] M. Jorgensen, "Practical guidelines for expert-judgment-based software effort estimation," *Software, IEEE,* vol. 22(3), pp. 57-63, 2005.

[6] M. Auer, A. Trendowicz, B. Graser, E. Haunschmid, and S. Biffl, "Optimal Project Feature Weights in Analogy-Based Cost Estimation: Improvement and Limitations," *IEEE Transactions on Software Engineering,* vol. 32(2), pp. 83-92, Feb 2006.

[7] J. W. Keung, B. Kitchenham, and D. R. Jeffery, "Analogy-X: Providing Statistical Inferences to Analogy-based Software Cost Estimation," *Submitted to IEEE Transactions on Software Engineering,* 2006.

[8] Wikipedia, "Defintion of "Spurious Relationship"," in *Wikipedia The Free Encylopedia*: www.wikipedia.org, 1997.

[9] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer Research,* vol. 27, pp. 209-220, 1967.

[10] B. F. J. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd Edition ed.: Chapman & Hall/CRC, 1997.

[11] B. F. J. Manly, *Multivariate Statistical Methods - A Primer*, Second Edition ed.: Chapman & Hall/CRC, 1998.

[12] F. H. C. Marriott, "Barnard's Monte Carlo tests: How many simulations?" *Applied Statistics,* vol. 28(75-7), 1979.

[13] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross validation," *The American Statistician,* vol. 37(pp. 36-48, 1983.

[14] J. M. Desharnais, "Analyse statistique de la productivitie des projects informatique a partie de la technique des point des function," in *University of Montreal.* vol. Masters thesis: Univ. of Montreal, 1989.

[15] T. Foss, E. Stensrud, B. Kitchenham, and I. Myrtveit, "A Simulation Study of the Model Evaluation Criterion MMRE," *IEEE Trans. on Software Engineering,* vol. 29(11), pp. 985-995, 2003.