The Consistent Measurement of External Subjective Software Quality Attributes

Alireza Ataei

Department of Computer Science, Ardabil Branch, Islamic Azad University, Ardabil, Iran.

Abstract: Most external software quality attributes are conceptually subjective. For example, maintainability is an external software quality attribute, and it is subjective because interpersonally agreed definitions for the attribute include the phrase ‘the ease with which maintenance tasks can be performed’. Subjectivity clearly makes measurement of the attributes and validation of prediction systems for the attributes problematic. In fact, in spite of the definitions, few statistically valid attempts at determining the predictive capability of prediction systems for external quality attributes have been published. However, attempts at validating prediction systems for external subjective quality attributes have tended to rely on experts indicating that the values provided by the prediction systems informally agree with the experts' intuition about the attribute. These attempts are undertaken without a pre-defined scale on which it is known that the attribute can be measured consistently. In this paper we use Bayesian inference, Markov chain Monte Carlo simulation and missing data imputation to develop statistical tests for consistent measurement of subjective ordinal scale attributes. In so doing, it would be possible to decide whether consistent independent estimates of the true values of software quality attributes can be assigned and prediction systems for quality attributes developed.

Key words: Subjective Software Quality Attributes, Direct Measurement, Consistent Measurement, Distribution Principle, Minimum Rejection Principle, Bayesian Inference, Error Rates.

INTRODUCTION

In order to assist software quality management and development the International Standards Organisation (ISO) identifies six external quality attributes, which are expected to be measured and controlled. At least three of these attributes, namely maintainability, usability and portability, are unequivocally subjective because they require 'ease of change, use or transfer to be assessed (ISO/IEC 9126-1:2001 2001). Furthermore, in the case of maintainability, Rosenberg states that ‘‘….we would like to be able to tell from the source code how difficult it will be to make a change…’’ (Rosenberg, 1997). He then suggests that ‘‘the difficulty is measured both in the amount of time needed to make the change, and the probability of successfully making the change on the first attempt’’. In general, the ability to express the semantics of code changes is weak and the relationship between code and change difficulty is not completely understood consequently there is a lack of generally applicable valid prediction systems (Rosenberg, 1997). Hence, at the code stage, both time and probability are only likely to be subjective estimates (unless the change has been made before and the same set of circumstances that influenced that change will apply in the future). If estimates of the two measures can be obtained from experts the measures could be expressed as an expected time with some confidence limits for a change, giving a subjective range of possible values. An equally valid alternative would be to express the difficulty of a change as a subjective value on an ordinal scale. However, it is important to distinguish between the external attribute maintainability of a software module or system and a prediction of maintenance difficulty for a specific change. Maintainability of a software module is clearly intended to indicate potential maintenance difficulties inherent in the module code in general and not the difficulty (or duration) of making a specific maintenance change.

The subjective measurement of external software quality attributes, including maintainability, readability, understandability and testability of modules, has also been proposed as part of one of several alternative strategies for improving safety related software engineering standards (Fenton and Neil, 1998). Furthermore, recently developed approaches for use in software management and control also require ordinal scale measurement, e.g. Bayesian Belief Networks and Naïve Bayes Networks. These approaches require that external attributes (e.g. reliability, software development effort) and internal product and process attributes (e.g. testing effort, problem complexity, code complexity) can be measured consistently on ordinal scales by their users (Fenton and Neil, 1999) and (Pendharkar et al., 2005). Melton et al., (1990), when proposing the foundations of a methodology for software measurement research, clearly distinguished between ordered objective consistent measures that can be made on a software abstraction (an attribute of a software entity) and external psychological attributes that we should like to associate with the abstractions. For example, Melton et al. (loc cit.) distinguish between an abstracted structural measure from a software document and the psychological complexity of the document. They also indicate that some researchers had (circa. 1990 and...
previously) not made clear their underlying assumptions concerning the nature of the expected relationship between the external attribute and the abstracted measure they were proposing. For example, researchers had attempted to equate structural software complexity with psychological complexity of the document without unequivocal evidence for the assumption. Such proposals had been made even though it is clear that the psychological complexity of a document is subjective and should be expected to vary with the person assessing the document, (e.g. a programmer).

Measurement Agreement:

The Kappa statistic is frequently used to assess inter-rater agreement about subjective classifications of ordinal scale attributes. Kappa and weighted Kappa statistics provide a subjective rating of agreement between raters (Altman, 1999).

In addition, whilst Kappa and Alpha and similar statistics provide means of assessing agreement or the relevance of questions to the attribute of interest none of them formally consider whether the measurement that may be taking place might be consistent and represent the attribute from a measurement theory perspective. Even Alpha, which is used to examine internal consistency of the questions asked and provides a score value for an entity’s subjective attribute, does not indicate whether the Alpha scores actually represent the subjective attribute.

In general, the approach is used to improve the understanding of a complex multivariate dataset by reducing the number of dimensions of the dataset and hence clarifying the main elements of the underlying structure of the data. The latent trait approach can be used when it is desired to assign a single value to a multi-dimensional set of variables. The variables may be categorical, ordinal or binary and the trait might be continuous or categorical. An example of an unobserved concept underlying observed data is that of physical and mental disability. Physical and mental disability is defined in terms of inability to perform what can be considered to be, the normal activities of day-to-day life and is measured in several dimensions (Congdon, 2001). This is a similar situation to the ones in which Alpha can be used, and like Alpha it does not enable us to provide a demonstration of the representation condition of measurement for subjective ordinal scale classification. One reason for this is that it does not provide us with information about the rejection rates. Another reason is that the latent trait approach is expected to be used when there are several recognised and measurable dimensions to the latent trait. This situation does not apply when raters are asked to directly classify an entity on an ordinal scale. Hence, this approach too has a similar purpose to Cronbach’s Alpha. However, an item response curve can be used to measure the probability that an individual respondent answers correctly given their trait score. Note that once again it is necessary to have several items about which responses must be analysed and that the response curve will not facilitate the estimation of measurement error for an ordinal scale with more than two classes. Thus item response models are for these reasons not an appropriate means of examining rejection rates.

What Is Consistent Measurement:

In the latter paper a discussion of why structural software measures such as McCabe’s cyclomatic complexity metric cannot be considered to represent the software attribute complexity are given by using the axioms for ordinal scale measurement. In particular, McCabe’s metric can be shown not to be considered to represent external complexity because there exists no known consensus of agreement that equal values of McCabe’s metric for two different program modules, with different control flow structure, represent equal levels of complexity. Strictly we might also want to make sure that we could find a group of experienced software engineers who could agree that the level of complexity was different for the two modules. Nevertheless, it is almost certain that there will be disagreement about complexity for some program modules because there is an element of subjectivity in the interpretation of external complexity and no agreed definition for program module complexity.

Consistent objective measurement is often considered to include the notion of being able to repeatedly assign the same number to an attribute. However, when measurement of length on a ratio scale using a rule is undertaken the true value of the length of an object may never be known or physically measured and according to Kyburg, (1984): “What makes the measurement of length objective is not that people always agree, but that, although they (almost) never agree, they come very close to agreeing almost all of the time.” Hence, even for objective consistent measurement repeatability may only apply within certain limits of error.

Furthermore, lack of bias is a necessary condition for consistent measurement. But it is not sufficient because there would be no evidence of bias if raters randomly allocated entities to classes. Kyburg ensures that bias is minimised by applying the Distribution Principle (DP) (Kyburg, 1984). The DP states: “Given a set of statements including both incorrigible observation statements and prima facie observation statements which is consistent with the axioms of our language, and given that the minimum rejection principle is satisfied, the number of prima facie observation statements of each kind to be rejected is that number that makes the rejection rates for the various kinds of observation statements as nearly uniform as possible.” Thus such rejection rates
will give the least bias that can be inferred from the set of measurement observations made by the raters. In summary, the more agreement and the more uniform the rejection rates the more consistent the measurement.

The disagreements should occur in an unbiased or uniform manner and therefore Kyburg requires that the rejection rates are uniform for consistent measurement to have taken place (Kyburg, 1984).

Uniform rejection rates with maximum agreement will ensure minimum error. Thus to be able to infer consistent measurement of an attribute it is necessary to show agreement and uniform rejection rates for a set of empirical measurement statements or classifications. The agreement and the uniformity of the rejection rates obtained from a set of classifications might be sufficient, depending on what use is to be made of the measurement. In summary - the more agreement and the more uniform the rejection rates the more consistent the measure. Furthermore, the best estimates of the true values for the sample of entities measured by a group of raters will occur when measurement is consistent. That is, the true values are best estimates in the sense that they are chosen to maximise agreement between raters, whilst ensuring that the rejection rates are unbiased, and thus minimised, given the set of measurement observations made by the raters.

Consistent Measurement for Subjective Attributes:
Consistent measurement gives minimum error (given the raters’ set of classifications) and hence consistent measurement of a subjective attribute will give the best estimate of the true value of an entity’s quality attribute from the raters’ observed classifications. Furthermore, the best estimates of the true classes can then be used in order to determine the predictive capability of a prediction system for an external quality attribute. Such prediction systems being of most valuable at architectural and detailed design.

We note that the concept of consistency in consistent measurement is not the same as that in consistent estimator of a statistic. A consistent estimator is a sample estimate of a population statistic, which obeys the following condition. For an estimator to be considered consistent requires that the probability of the difference between the estimator and the statistic converges to zero as m the number in the sample becomes large (Kendall and Stuart, 1973). For example, the sample mean of a set of data is a consistent estimator of the population mean.

Furthermore, the concept of consistent measurement on an objective ordinal scale is defined by its associated axioms (Roberts, 1979). These axioms are mathematically demonstrable for objective ordinal scale measurements, but clearly this will not necessarily be the case for subjective ordinal scale measurements, cf. (Fenton, 1994).

A demonstration of consistent measurement ensures minimum error given the raters’ observed classifications. Furthermore, when the raters’ measurement is consistent the estimated true classes will be consistent and can be used to determine the predictive capability of a prediction system for an external subjective software attribute.

Tests for Consistent Measurement on an Ordinal Scale:
The tests given in this paper address four basic sets of questions. The first set of questions concerns whether there are differences between raters’ error raters. The second set addresses whether agreement has taken place. The third set questions whether pairs of error rates can be considered equal. The fourth set considers whether differences in the size of the error rates with respect to distance from the true class are acceptable. Together the answers to these questions are sufficient to decide if consistent subjective measurement on an ordinal scale has taken place.

It is straightforward to provide estimates of the true classes by applying the MRP. However, when an equal number of raters choose two or more different classes for an entity a tie occurs.

Discussion:
In addition, there are other difficult problems to consider when attempting to assign definitive and widely acceptable quality attribute values to software systems or systems’ artefacts. These difficulties clearly concern interpersonal agreement amongst large groups of experts and users across different organisations. The problem of agreement between many experts would be difficult to solve (should a solution be sought) and may only be soluble through the auspices of large organisations with an interest in software quality, such as the BCS, ISO, ACM and IEEE. Such organisations are likely to be the only ones capable of gathering together sufficient experts and users in order to gain agreement about definitive values for external quality attributes. Furthermore, the definitions for quality attributes, such as maintainability, may yet need to be made more precise before consistent measurement can be achieved across groups of experts and users.
However, when consistent measurement for a group of raters can be demonstrated the best estimates derive from the classifications for the true classes of the entities will also have been identified. These true value estimates can then be used to provide an independent and unbiased assessment of the predictive capability of a prediction system for a quality attribute. Prediction systems are not often intended to provide definitive values for quality attributes of software entities and tend to be used at an early stage of software development. For example, in the case of maintainability, in order to alert developers during design or coding to potential maintenance problems and to indicate to them that the modules might best be modified at design or coding to facilitate maintenance. Hence, in this case, whether a value is definitive or not is less important than the ability to identify modules (with a quantified level of confidence) that the raters would find difficult to maintain. Clearly, however, such prediction systems are likely to be of most value within the system development department in which they were developed, where agreement and measurement consistency between a department’s experts is likely to be more easily obtained.

No assumption has been made in this paper concerning the nature of any relationship of an abstracted consistent objective measure and an external subjective or psychological quality attribute. However, the paper does seek to emphasise that empirically demonstrating that a subjective external attribute, such as maintainability, can be measured consistently prepares the way for identifying valid prediction systems for the subjective attribute. The prediction systems would use abstracted objective measures of attributes of software artefacts as inputs. Such demonstrations would also assist in helping to determine relationships between internal objective abstracted measures, intuitively thought related to an external quality attribute, and the quality attribute. Demonstrations of this kind might eventually help identify acceptable indirect measures of important subjective external attributes. Furthermore, empirical demonstrations would inevitably promote dialogue amongst software researchers resulting in greater understanding of the important external subjective quality attributes; cf. Baker et al.

A set of tests to demonstrate consistent measurement for ordinal scales has been given, which we developed from Kyburg’s Minimum Rejection and Distribution Principles. Our tests use estimates of the true classes for tied observations. We derive these estimates using Bayesian inference and MCMC missing data imputation. Bayesian inference is also used to make coherent probability statements from which inferences can be made concerning agreement, error rate equality (uniformity) and inequality.

Once consistent measurement has been demonstrated consistent estimates of the true classes can be derived from the classifications given by the raters. Further, the estimates will be unbiased with respect to the raters’ observed classifications. These estimates can then be used to provide an assessment of the predictive capability of a prediction system, which is independent of the prediction system. Prediction systems are not intended to provide definitive values for quality attributes of software entities and tend to be used at an early stage of software development. For example, in the case of maintainability, the objective would be to produce prediction systems that alert developers during design or coding to potential maintenance problems. Hence, in this case whether a value is definitive or not is less important than the ability to identify modules, with a quantified level of confidence, which the raters believe would be difficult to maintain. In addition, such prediction systems are likely to be of most value within the department in which they were developed, where rater experiences are similar and consistent measurement may be more easily obtained.

It is worth mentioning that our approach would work if we knew the true values for the entities a priori. We would be able to obtain a set of error rates for the raters that we could test for consistency. Also we can imagine a situation where we had a set of error rates, perhaps from another group of experts, which we already knew gave consistent measurement. If we then asked another set of raters to classify the same sample of modules (or another sample from the same population) we could test whether the two sets of raters were consistent with each other.

Whether or not measurement is sufficiently consistent to be useful is dependent on the purpose for which the measurement is to be used. In general, the more agreement and the more uniform the error rates the better. Clearly, agreement would need to be inferred to be much higher than in the example we use when definitive values are to be assigned to a software artefact. Higher levels of agreement would be desirable, for example, when attempting to validate an objective indirect measurement scale for maintainability. In addition, there are other difficult problems to consider when attempting to assign definitive and widely acceptable quality attribute values to software systems and their artefacts. These difficulties concern interpersonal agreement amongst large groups of experts and users across different organisations. The problem of interpersonal agreement between many experts would be difficult to solve and may only be soluble through the auspices of large organisations with an interest in software quality, such as the ISO, ACM, IEEE and the BCS. We believe that such organisations are likely to be the only ones capable of gathering together sufficient experts and users in order to gain agreement about definitive values for external quality attributes.
REFERENCES

Alireza ataei and afshin salajegheh, 2011. research journal of information techkology, 3(3).
and Sons Ltd.
Fenton, N.E., 1994. Software measurement: A necessary scientific basis. IEEE Transactions on
9126-1.
Wiley.
NTIS(Vol. 1). Springfield, VA: NTIS.
Volume 7.Addison-Wesley, Massachusetts.