Pre-publication draft. *Proceedings of the International Conference on Computers in Education*, (Chiba, Japan), IOS Press, Amsterdam, vol. 1, pp. 111-118

Dual-model: An Architecture for Utilizing Temporal Information in Student Modeling

Bark Cheung Chiu and Geoffrey I. Webb

School of Computing and Mathematics, Deakin University, Geelong, Australia
e-mail: {chiu, webb}@deakin.edu.au

A modeling system may be required to predict an agent's future actions even when confronted by inadequate or contradictory relevant evidence from observations of past actions. This can result in low prediction accuracy, or otherwise, low prediction rates, leaving a set of cases for which no predictions are made. This raises two issues. First, when maximizing prediction rate is preferable, what mechanisms can be employed such that a system can make more predictions without severely degrading prediction accuracy? Second, for contexts in which accuracy is of primary importance, how can we further improve prediction accuracy? A recently proposed Dual-model approach, which takes models' temporal characteristics into account, suggests a solution to the first problem, but leaves room for further improvement. This paper presents two classes of Dual-model variant. Each aims to achieve one of the above objectives. With the performance of the original system as a baseline, which does not utilize the temporal information, empirical evaluations in the domain of elementary subtraction show that one class of variant outperforms the baseline in prediction rate while the other does so in prediction accuracy, without significantly affecting other overall measures of the original performance.

Keywords: **Agent modeling, Student modeling, Temporal model, Decision tree.**

1 Introduction

A major difficulty confronts any student modeling system operating in an educational environment. While the system seeks to form a model of a subject's competencies based on observations of applications of those competencies, by the time it has collected sufficient observations to form a model, the competencies that generated the initial observations are likely to have changed. When a student model is formed based on historical observations of the student's behavior, older evidence is likely to be less pertinent than more recent evidence for accurately predicting the student's future actions. A learning system that treats all historical data with equal weight may produce inaccurate student models, reducing their practical utility and degrading their prediction performance. However, it is rarely possible to accurately determine which observations pertain to the student's current competencies. Consequently, to discard older observations is to risk discarding valuable pertinent data. When a modeling system is required to predict a student's future action, it may face a dilemma: to make an unreliable prediction or to make no prediction at all. This can result in low prediction accuracy, or otherwise, low prediction rates, leaving a set of cases for which no predictions are made. This raises two issues. First, when maximizing prediction rate is preferable, what mechanisms can increase prediction rate without severely degrading prediction accuracy? Second, for contexts in which accuracy is of primary importance, how can we further improve prediction accuracy without degrading prediction rate?

Previous attempts at tackling the first issue by using various conflict resolution methods have reported trade-offs between prediction rate and prediction accuracy, where prediction rate improved at the expense of reducing prediction accuracy [2, 6]. Chiu and Webb [1] propose a Dual-model approach (using two temporally divided models) to alleviate this problem. These two models, namely the *fresh* and *extended* models, once inferred, can not only better predict students' future actions, but also enrich the description of a student's mastery of a domain. Initial evaluation of Dual-model in the context of modeling 3-digit subtraction skills has demonstrated that the augmented system can provide significantly more predictions without significantly affecting the levels of prediction accuracy [1]. However, only one of a number of possible strategies to this end was explored. Further, this 'fresh-first' strategy was based on the assumption that the predictions of a model based on recent evidence would be more accurate than an 'extended' model based on all observations of the student. Subsequent testing of this hypothesis suggested that it was incorrect, however [X – need to fix up reference numbering]. In addition, this method was evaluated with a data set for which varying methods were used to collect data from a moderate size of population (73 students from three schools). In order to confirm that the original Dual-model can yield better performance, a re-evaluation of this method with new data is desirable.

Moreover, previous studies have not dealt with the second main issue. Techniques for improving prediction accuracy while maintaining prediction rates remain an important goal. This research investigates a new class of techniques for tackling this problem. Considering the simplicity of the Dual-model approach, which imposes only a one-fold increase on the original computational cost, we explored new alternatives to this method and developed three new variants of the Dual-model method, aimed at improving accuracy without adversely affecting prediction rate. This paper presents an evaluation of these alternatives derived from the Dual-model approach in a domain of modeling students' subtraction skills.

2 The use of temporal information in student modeling systems

A student's knowledge and beliefs may alter over time. Consequently, a static student model may not truly reflect the student's current knowledge. While Giangrandi and Tasso [4] propose a temporal management mechanism, such that contradicting hypotheses about a student can co-exist within a model, Webb and Kuzmycz [9] embedded a data aging mechanism in the FBM (Feature Based Modeling) system [8]. Both of these approaches share a belief that historical data, which is based on direct observations, is a valuable source of information for student models, and should not be overlooked.

To utilize temporal information in historical data, Giangrandi and Tasso introduce a time variable to be used in a truth maintenance system, which aims to generate models to explain students' explicit and implicit beliefs. In contrast, aiming at building shallow but accurate models, FBM's data aging mechanism does not require an explicit time variable, utilizing instead only the temporal order of the observations. This is used to discount old data by a set factor. Unfortunately, however, empirical evaluation of data aging reveals poor performance in practice [9]. Chiu and Webb [1] propose an alternative simple method, called Dual-model, to cater for the temporal factor. This method creates a temporal model, namely *fresh model* (built using data from the most recent observations), in addition to the conventional model, which is referred to as an *extended model* (inferred from data of all historical observations). When a Dual-model system predicts a student's future actions, both models will be consulted.

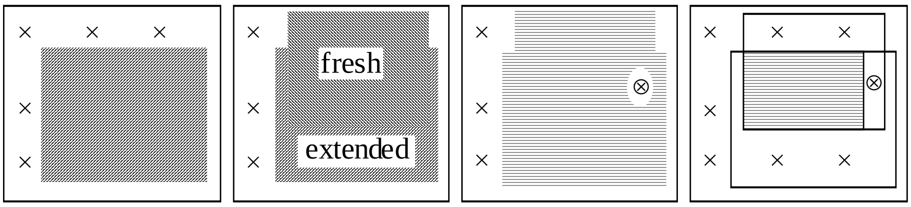
The advantages of the Dual-model method include: (1) it is simple to implement; and (2) it can be applied to other modeling systems for which training examples can be grouped by temporal characteristics. With respect to model interpretation, the fresh model provides additional information that the users may compare with the extended model to interpret which aspects of a student's behavior have changed. Cook and Kay [3] have argued that students will benefit from viewing their own models. To this end, utilities that simplify models' knowledge representation can be employed in Dual-model, improving readability and hence ensuring that those novice users may receive greater educational benefit.

The previous study of the Dual-model approach explored various sequential consultation strategies over the underlying models and favored a *fresh-first* strategy where the fresh model was consulted first. Under a sequential strategy, the models are consulted in turn until one makes a prediction, or all are exhausted. As already stated, this strategy was motivated by the assumption that the fresh model would be more accurate (albeit with lower prediction rate) than the extended model, an assumption that subsequent research has drawn into question [X]. However, Dual-models could also operate under various combinations of parallel consultation, and this paper seeks to evaluate the utility of these alternative Dual-model variants. For ease of exposition, *Dual-model*, hereafter, will refer to the generic method without specifying which consultation strategy is used, and *Dual-model1* will refer to the Dual-model method using the original fresh-first strategy, in the remaining sections of this paper.

3 New variants of Dual-model

Dual-model was originally designed to improve the prediction rate of FBM-C4.5, a variant of FBM. The FBM systems build a black-box model of an agent by capturing the relationships between the inputs and outputs of the agent's cognitive system. The context in which an action is performed is characterized by a set of attribute values called context features. An agent's action is characterized by a set of attribute values called action features. For each attribute with action features as values, a sub-model is inferred that predicts a specific action feature for any given combination of context features. These disparate sub-models can be considered in isolation, to examine different aspects of the agent being modeled. Alternatively, the predictions of each sub-model can be aggregated to make detailed predictions about specific behavior. FBM-C4.5 is an implementation that adopts the FBM approach, using C4.5 [7] as an induction engine to infer a decision tree for each sub-model. Placing accuracy as the first priority, all the FBM systems make no prediction, by default, when there exists ambiguity among the specific predictions of individual sub-models (a sub-model's prediction predicts whether an action feature will be present, while its specific prediction predicts the outcome of that action feature). For FBM-C4.5, this is implemented by employing a consensus resolver to determine a final prediction.

Given a domain, for which predictions are required, Figure 1(a) illustrates the cover of predictions made by a hypothetical system (for example, FBM-C4.5), where a symbol "×" stands for a case for which no prediction is made. Figure 1(b) depicts a possible cover by Dual-model. The predictions of a new base model (fresh) are superimposed on those of the original extended model. Note that the cover by the extended model alone is the default setting of the original system. The predictions made by the fresh model overtake part of the extended model's cover. Informed by recent findings that both fresh and extended models appear to have comparable general accuracy, , an alternative strategy is suggested, which refrains from making predictions whenever the two models make different predictions. This strategy leads to a Dual-model variant, DualConsensus, which consults fresh and extended models in parallel and uses a consensus resolver to handle conflicting predictions from the two models. It aims to improve the prediction rate of a single-model system while taking greater precautions against making incorrect predictions. Figure 1(c) illustrates a possible cover by DualConsensus, where a symbol "⊗" stands for a case for which no prediction is made due to conflicting predictions from the two models.



(a) Single-model (b) Dual-model1 (c) DualConsensus (d) DualConsensus-strict

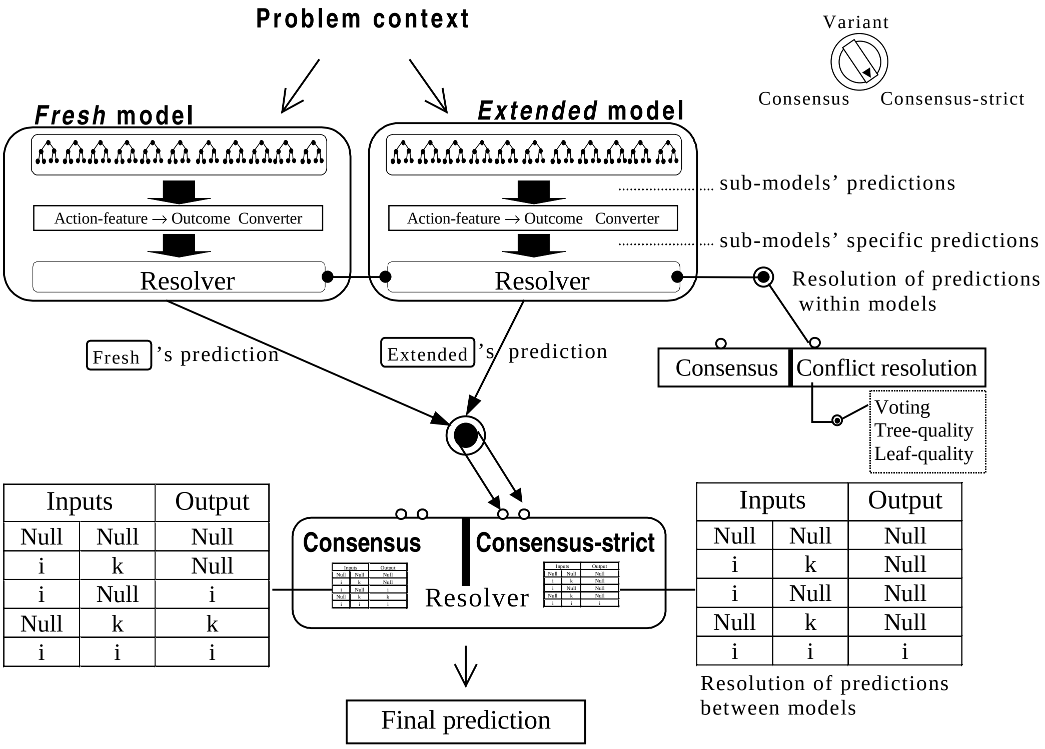
These illustrations plot hypothetical models on a two dimensional space. Each dimension represents a hypothetical context feature (independent variable). Points would normally be labeled by predicted values, but are here used to depict the combinations of context feature values for which a system makes a prediction (covered by the shaded area), or makes no prediction (represented by a symbol "×" or "⊗").

Figure 1. Predictions covered by a single-model system and variants of Dual-model.

Alternatively, the system can output a prediction only when both models make predictions and suggest the same specific prediction. Otherwise, it outputs no prediction. The use of this *Consensus-strict* resolution mechanism, leading to a variant called DualConsensus-strict, should provide high prediction accuracy because the predictions are supported by both models. However, the system's prediction rate will be very likely to drop, as illustrated in Figure 1(d), where the cover area is dramatically reduced. To solve this potential problem, a novel method is to use base models augmented by prediction promoters to form the Dual-model system. Chiu and Webb [1] have studied three conflict resolvers, namely Voting, Tree-quality, and Leaf-quality, respectively, for FBM-C4.5. All of these resolution methods exhibited an improvement in prediction rate, albeit at a small cost in reduced accuracy. These resolution methods can be considered as candidate prediction promoter techniques to be used within fresh and extended models. As Dual-model trades-off increased accuracy for decreased prediction rate, it was anticipated that the use of these internal resolution techniques within Dual-model might augment each other, the negative trade-off in one measure from one technique being more than compensated for by the positive effect on that measure from the other technique. The operations of these resolvers are summarized as follows.

- Voting resolver: outputs a prediction with a majority of votes from competing specific predictions. If two or more specific predictions tie for the first place, no output is made.
- Tree-quality resolver: sequentially consults the trees, which were ranked by quality measures using a ten-fold cross-validation [5] technique. When a tree provides a positive prediction, its specific prediction is adopted. If all trees are exhausted, the resolver outputs no prediction.
- Leaf-quality resolver: receives data pairs, <Tree's prediction, Leaf's quality>, from the trees, where Leaf is the decision node of a tree for a test instant and its quality is estimated based on the homogeneity of training examples reaching the node. If none of the trees provides a positive prediction, the resolver makes no prediction. Otherwise, among the data pairs for which the first element is positive, the resolver adopts a tree associated with the Leaf with the highest quality value, and outputs the tree's specific prediction.

Figure 2 summarizes all the new strategies applied to Dual-model, mentioned above. A test problem, which is described by a set of context features (Problem context), from a domain, requires the system to predict a student's answer. The current setting illustrates that DualConsensus-strict is in operation, where a Voting resolver is selected as the prediction promotor within individual models.



Inputs		Output
Null	Null	Null
i	k	Null
i	Null	i
Null	k	k
i	i	i

Inputs		Output
Null	Null	Null
i	k	Null
i	Null	Null
Null	k	Null
i	i	i

Figure 2. A prediction made by a variant of Dual-model.

4 Evaluation

4.1 Experiments

224 nine-to-ten year old students from eight primary schools participated in five rounds of tests, where Tests 1 and 2 were conducted with a one day interval in week 1 and the remaining successive tests were administered at weekly intervals. For each test, each student was administered a test sheet consisting of forty 3-digit subtraction problems. Twenty randomly pre-generated repeated problems were intermixed with a further twenty problems generated randomly and uniquely for each sheet. These repeated problems, which covered a general spectrum of problem context, were included to allow evaluation of students' consistency in problem solution, the discussion of which falls outside the scope of this paper. For all random problems, the minuend was greater than or equal to the subtrahend, ensuring that the correct solutions were not negative. For each test sheet, all of the 40 problems were randomly ordered. This might have reduced the external influence on individual students by other students, for example, through discussion of the problem solving methods after a test, which might affect their performance in further tests.

The answered test sheets from each student were entered into a corresponding data file. To minimize typing errors, each data sheet was input twice and cross-validated to resolve discrepancies. The data files from those students who have been absent from any test and from those who answered less than 30 problems on any test sheet were excluded from the final data set. The resulting data set consists of 172 student files. Based on the pre-generated repeated problems, a validation program was used to trace any typing errors on those repeated problems. Seven 3-digit items, out of the 34400 data items, 172 (files) × 5 (tests) × 20 (repeated problems) × 2 (minuend and subtrahend), were discovered as mistyped and amended. We obtained a data-entry error rate of 0.0002. Assuming that the rate of errors detected in these values was similar to that in the remaining values, we estimate that the probability of any single randomly generated problem or a student answer being incorrectly entered is less than 0.0005.

The FBM-C4.5 subtraction modeller (henceforth FBM-C4.5) predicts each digit of a solution independently. In consequence, the following discussion uses the column of a subtraction problem as the unit of measure. Four metrics are useful for evaluating a modeling system:

- *Prediction rate*: the percentage of columns for which a prediction was made;
- *Prediction error*: the percentage of predictions that were incorrect;
- *Error prediction proportion*: the proportion of predictions that predicted an erroneous answer; and
- *Error prediction error*: the percentage of error predictions that were incorrect.

For evaluation, a Dual-model system predicted a student's answers from Tests 3 to 5 based on the models inferred from the prior round data. Let $M(1, \ldots, i)$ be a model built from observations from Tests 1 to i. Each of the following experiments was conducted by forming two models, *extended*$(1, \ldots, n - 1)$ and *fresh*$(n - 1)$, from a sequence of tests, $1, \ldots, n - 1$, which were used to predict a student's precise answers (on a column by column basis) for the 20 non-repeated randomly generated 3-digit questions in Test n. The reason for using those 20 questions in a test instead of all 40 questions is that the excluded questions are repeated questions, which have been seen in prior tests, and used to infer the fresh and extended models. The randomly generated questions served to provide unseen cases to test the systems studied.

Starting from $n = 3$, from where the fresh and extended models first differ, the above process was repeated up to $n = 5$. For each student, the total numbers of answers and errors collected, and the total numbers of predictions and correct predictions made by a system in these tests, were determined. The relevant grand totals for the whole data set were used to compute overall performance.

The first experiment evaluated two variants of the Dual-model method, namely Dual-model1 and DualConsensus, against the baseline FBM-C4.5. Where Dual-model1 uses a fresh-first strategy to make a prediction, DualConsensus uses a consensus resolver to determine a final prediction. The second experiment evaluated three new variants of Dual-model, namely DualVstrict, DualTstrict and DualLstrict, against the baseline FBM-C4.5. The letters V, T and L in these names represent the use of Voting resolvers, Tree-quality resolvers and Leaf-quality resolvers, respectively, in fresh and extended models. All of these new variants employ a *Consensus-strict* resolver to determine a final prediction.

4.2 Results

In Tests 3 to 5, the 172 students contributed 30,882 answers (digits), of which 3,500 were incorrect answers. FBM-C4.5 made 29,483 predictions (95.5% prediction rate). Of these predictions, 6.3% (27,639) were accurate. For predicting the students' errors, the system suggested 1,980 digits, of which 1,467 were accurate. These accounted for an error prediction proportion and error prediction accuracy of 6.7% and 74.1%, respectively. With these measures as a baseline, the corresponding performance measures of Dual-model1 and DualConsensus are presented in Table 1. We used two-tailed paired t-tests to evaluate the statistical significance of these systems' performance against the baseline. The statistical outcomes, p values and t values, are also listed in the corresponding columns. Where a performance difference is significant (at the 0.05 level), the corresponding value is shown in a different font, bold or underlined, to indicate that its is significantly better or worse, respectively. The degrees of freedom of these t-tests were 171 unless otherwise specified.

Table 1. Performance of two Dual-model variants against the baseline (two-tailed t-test).

	Baseline	Dual-model1	DualConsensus
Total predictions made	29,483	30,315	30,091
Prediction rate (%)	95.5	**98.1**	**97.4**
p value [*t* value]		< 0.0001 [9.01]	<0.0001 [7.33]
Total predictions that were incorrect	1,844	2,084	1,960
Prediction error (%)	6.3	6.9	6.5
p value [*t* value]		0.0067 [-2.75]	0.1018 [-1.65]
Error predictions made	1,980	2,144	2,031
Error prediction proportion (%)	6.7	**7.1**	6.7
p value [*t* value]		0.0358 [2.12]	0.3561 [0.93]
Error predictions that were incorrect	513	616	553
Error prediction error (%)	25.9	28.7	27.2
p value [*t* value]		0.4636 [-0.74]	0.7886 [0.27]
degree of freedom*		45	45

*For error prediction accuracy, only pairs in which error predictions have been made by both systems can be compared.

As can be seen, Dual-model1 achieved an overall prediction rate of 98.1%, and an error prediction proportion of 7.1%, which is significantly higher than that of FBM-C4.5 (95.5% and 6.7% respectively). However, a small but significant increase in prediction error (6.9) is also observed. For DualConsensus, a significant improvement (by 1.9%) in its prediction rate has been achieved while there is a slight but not significant increase in prediction error, in comparison to FBM-C4.5. DualConsensus comes closest to our objective of increasing prediction rate without degrading accuracy.

The performance of the second group of Dual-model variants is summarized in Table 2. *DualVstrict* exhibits a trade-off effect across the four prediction performance categories. *DualLstrict* and *DualTstrict* achieve a significant improvement in overall prediction error without affecting the overall prediction rate. Note that there is still a trade-off effect on the performance of error prediction for these two systems. While a substantial gain in error prediction error is observed for each of them (3.9% and 4.5%, DualTstrick and DualLstrick, respectively), each trades off a considerable loss in error prediction proportion (0.7% and 0.8%, respectively). It appears that the technique of integrating an external Consensus-strict resolver and internal quality-oriented conflict resolvers is effective in reducing the risk of making incorrect predictions while a high level of overall prediction rate can still be maintained. In the context of overall performance, these variants with hybrid combination of conflict resolvers achieve our second objective. In the context of error prediction, the effectiveness of this class of Dual-model is subject to how these utilities are used.

Table 2. Performance of three Dual-model variants against the baseline (two-tailed t-test).

	Baseline	DualVstrict	DualTstrict	DualLstrict
Total predictions made	29,483	29,230	29,582	29,485
Prediction rate (%)	95.5	94.7	95.8	95.5
p value [*t* value]		0.0377 [-2.09]	0.7673 [0.30]	0.9868 [0.02]
Total predictions that were incorrect	1,844	1,563	1,597	1,576
Prediction error (%)	6.3	**5.3**	**5.4**	**5.3**
p value [*t* value]		0.0011 [3.32]	0.0028 [3.04]	0.0008 [3.43]
Error predictions made	1,980	1,731	1,770	1,727
Error prediction proportion (%)	6.7	5.9	6.0	5.9
p value [*t* value]		0.0001 [-4.02]	0.0006 [-3.50]	<0.0001 [-4.01]
Error predictions that were incorrect	513	378	390	370
Error prediction error (%)	25.9	**21.8**	**22.0**	**21.4**
p value [*t* value]		0.0251 [2.32]	0.0084 [2.77]	0.0489 [2.03]
degree of freedom		43	43	42

5 Conclusions

Two classes of Dual-model variants have been evaluated in this study. Each attempts to solve one of the problems stated at the commencing section of this paper: (1) how to improve prediction rate without degrading prediction accuracy; and (2) how to improve prediction accuracy without affecting prediction rate.

For the first objective, experimental results show that Dual-model1 is effective for improving the prediction rate. However, prediction accuracy is slightly but significantly affected. DualConsensus achieves a significantly higher prediction rate without significantly degrading the prediction accuracy of the original single-model system, in the domain of studied.

For the first objective, the employment of a Tree- or Leaf- quality resolver internally and a Consensus-strict resolver externally enables a Dual-model system to achieve significant improvement in prediction accuracy without significantly affecting the overall prediction rate. Of these two resultant hybrids, DualLstrict runs much faster than DualTstrict, due to the fact that a leaf quality estimate can be directly accessed from a tree while a tree's quality is estimated through a computing-intensive process. This observation suggests that DualLstrict is to be preferred among variants of this class.

Our experimental results reveal, in situations where a series of observations are collected over an extended period of time, the Dual-model approach, which takes temporal factors into account, provides two directions for improving overall prediction performance. Dual-consensus improves the overall prediction rate, while DualLstrict improves the overall prediction accuracy, albeit without improvement in error prediction proportion, of the original single-model system. While the Dual-model technique has been developed and evaluated in the context of FBM-C4.5, it should be equally applicable to any student modeling system that constructs models from multiple observations over time.

References

[X] Chiu, B. C, "Predictive Modeling of Student Competency" PhD thesis, Deakin University, Geelong, Aust., 1999. (Or other reference if there is one).

[1] Chiu, B. C., and Webb, G. I., "Using decision trees for agent modeling: Improving prediction performance", *User Modeling and User-Adapted Interaction*, vol.8**,** no. 1-2, pp. 131-152 (1998).

[2] Chiu, B. C., Webb, G. I., and Zijian, Z., "Using decision trees for agent modeling: A study on resolving conflicting predictions", *Proceedings of 10th Australian Joint Conference on Artificial Intelligence*, AI97, Perth, pp. 349-358 (1997).

[3] Cook, R. and Kay, J., "The justified user model: A viewable, explained user model", *Proceedings of Fourth International Conference on User Modeling*, Hyannis, Mass., pp. 145-150 (1994).

[4] Giangrandi, P., Tasso, C., "Temporal reasoning in student modelling", *Proceedings of AI-Ed 97 World Conference on Artificial Intelligence in Education* Kobe, Japan. IOS Press, pp. 514-521 (1997).

[5] Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection", *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, pp. 1137-1145 (1995).

[6] Kuzmycz, M., "Resolving conflicting knowledge in student models", *Proceedings of the Eighth World Conference on Artificial Intelligence in Education.* Amsterdam: IOS Press, pp. 522-529 (1997).

[7] Quinlan, J. R., "*C4.5: Programs for Machine Learning*", Morgan Kaufmann (1993).

[8] Webb, G. I., and Kuzmycz, M., "Feature Based Modeling: A methodology for producing coherent, dynamically changing models of agent's competencies", *User Modeling and User-Adapted Interaction* vol. 5, no. 2, pp. 117-150 (1996).

[9] Webb, G. I., and Kuzmycz, M., "Evaluation of data aging: A technique for discounting old data during student modeling", *Proceedings of the Fourth International Conference on Intelligent Tutoring System* (1998).