Combining Different Classifiers in Educational Data Mining

He Chuan, Li Ruifan, and Zhong Yixin

School of Computer Science,
Beijing University of Posts and Telecommunications,
Beijing, China
hcl258@yeah.net

Abstract. Educational data mining is a crucial application of machine learning.The KDD Cup 2010 Challenge is a supervised learning problem on educational data from computer-aided tutoring. The task is to learn a model from students' historical behavior and then predict their future performance. This paper describes our solution to this problem. We use different classification algorithms, such as KNN, SVD and logistic regression for all the data to generate different results, and then combine these to obtainthe final result. It is shown that our resultsarecomparable to the top-ranked ones in leader board of KDD Cup 2010.

Keywords: data mining, logistic regression, k-nearest neighbor, singular value decomposition, classifiers combination.

1 Introduction

In KDD Cup 2010, the task is to predictstudent algebraic problem performance giveninformation regarding pasperformance. This prediction task presents not only technical challenges for researchers,but is also of practical importance, as accurate predictions can be used, for instance, tobetter understand and ultimately optimize the student learning process.Specifically, participants were provided with summaries of the logs of student interactionwith intelligent tutoring systems. Two data sets are available: algebra 2008-2009 andbridge to algebra 2008-2009. In the rest of this paper, we refer to them as A89 and B89, respectively. Each data set contains logs for a large number of interaction steps. Some interaction log fields areincluded in both training and testing sets, such as student ID, problem hierarchy includingstep name, problem name, unit name, section name, as well as knowledge components (KC)used in the problem and the number of times a problem has been viewed. However, somelog fields are only available in the training set: whether the student was correct on the firstattempt for this step (CFA), number of hints requested (hint) and step duration information. The details are list in Table 1.

J. Zhang (Ed.): ICAIC 2011, Part V, CCIS 228, pp. 467–473, 2011.

Table 1. Dataset statistics

Datasets	Algebra 2008-2009	Bridge to Algebra 2008-2009
Lines (train)	8,918,054	20,012,498
Students (train)	3,310	6,043
Steps (train)	1,357,180	603,176
Problems (train)	211,529	63,200
Section (train)	165	186
Units (train)	42	50
KC (train)	2,097	1,699
Steps (new on test)	4,390	9,807

The competition regards CFA, which could be 0 (i.e., incorrect on the first attempt) or 1, as the label in the classification task. For each data set, a training set with knownCFA is available to participants, but a testing set of unknown CFA is left for evaluation. The evaluationcriterion used is the root mean squared error (RMSE). In the competition, participantssubmitted prediction results on the testing set to a web server, where the RMSE generatedbased on a small subset of the testing data is publicly shown. This web page of displayingparticipants' results is called the "leader board."

Facing such a complicated problem, we use different classification algorithms such as KNN, SVD [2, 3] and logistic regression [2] for all the data to generate different results, and then combine these to get final result. In particular, logistic regression needs many proper features to work well, so feature engineering is a necessary step. KNN and SVD, which are transferred from collaborative filtering community, will exploit the basic information in the given data. In the following sections, we make the arrangement: Section 2 shows our method in details, including some preprocessing, data grouping, feature engineering, logistic regression and trust region optimization, and combining methods. Section 3 gives the final result and the discussion for our method.

2 Our Method

This section is the main part of our paper. It explains the whole procedure of our method.

A. Validation Set Generation

Because we do not have all the ground truth labels for test data, we have to generate validation sets by ourselves. Table 2 shows the number of samples in validation set of Algebra 2008-2009 and Bridge to algebra 2008-2009.

Table 2. The number of samples in validation and training set of Algebra2008-2009 and Bridge to algebra 2008-2009

Datasets	algebra 2008-2009	bridge to algebra 2008-2009
Training(V)	8,407,752	19,264,097
Validation()	510,303	748,402
Training(T)	8,918,055	20,012,499
Training()	508,913	756,387