1. The introduction of the research

This research is based on the KDD 2010 CUP competition with the datasets from Intelligent Tutoring Systems (ITS) for thousands of students in US with the spanning time of 2008-2009 school year. The competition was to predict the students’ future performance of solving algebra problems. The variable showing the ability of the student for the specific problem is *Correct\_First\_Attemp* which has binary classes with ‘0’ for correct and ‘1’ for incorrect. All the participants of the competition need to build the prediction model with two training datasets and two development datasets which are for the further prediction. The participants’ finished models will be tested with a test data and the evaluation of their prediction accuracy will be compared with the real output with RMSE method.

Dataset: The datasets for this competition have the problems of large volume and high dimensional attributes. Some attribute also has very high categorical distribution. List as:

|  |  |  |
| --- | --- | --- |
| Datasets | Students | Steps |
| Development Data Sets |  |  |
| Algebra I 2005-2006 | 575 | 813,661 |
| Algebra I 2006-2007 | 1840 | 2,289,726 |
| Bridge to Algebra 2006-2007 | 1146 | 3,656,871 |
| Challenge Data Sets |  |  |
| Algebra I 2008-2009 | 3310 | 9,426,966 |
| Bridge to Algebra 2008-2009 | 6043 | 20,768,884 |

Rules:

Goal:

1. The winners’ concerns

Size of the dataset:

Performance:

1. Distribution computing solutions

Workflow:

Algorithm:

1. Introduction of the methodology