

KDD CUP 2010

Hemar Team

Writed by Cheng Lei

Our Solution

- ø0. Some New Challenges
- ø1. VS. 2007 Netflix Prize
- ø2. Main Thoughts
- ø3. Results
- ø4. Conclusion

0. Some New Challenges

- 1.The data released by the organizer have no valid information about time.What we do some
- 2.The dimensionality curse. When the number of dimensionality comes to large,it seems like nothing to do for us,a poor group with poor facilities.
- 3.Weka.we confidently took it as an omnipotence tool. Finally,we abandon it,though we have a try to do some secondary development to adapt to the data.
- 4.Original thought is to do some ensemble.In our plan,with different particular data preprocessing technologies,these methods will bring into effect,including classification(logistic regression,CART),forgetting curve,clustering(k-means)...

5. MatLab.SVD (Singular value decomposition) was thought to do some PCA(Principal Component Analysis) for dimension decreasing. However, unexcepted, when the dimension is “cursed”, the machine, with 4G memory and 8 CPUs(actually, we can not do parallelized programing by MatLab),works out. Finally,we do nothing about PCA and dimension decreasing.

In a word,due to the limited time, our knowledge and human resources reasons, we only apply k-means algorithm for clustering—cluster students and clustering problem-step pairs.

Some pre-processing technologies are applied and parameters are probed many times to get our best answer.

Detail program is described as slides follows:

VS. 2007 Netflix Prize

We severely suggest you to know 2007's competition systematically.

The KDD Cup of the year 2007 dealt with the datasets of the Netflix prize. The training data were collected between October, 1998 and December, 2005. The ratings are on a scale from 1 to 5 (integral) stars.

The competition included two tasks:

The first task is to predict for a given User/Movie-combination (both given as IDs) whether the rating was done in 2006 or not.

The second task asks the participants to estimate the number of ratings for a subset of the 17770 movies given by the original Netflix Prize.

Judging Rule : the naive RMSE-scale.

For the reason of 1 million dollars, many thoughts and methods came out in that year. Learning many of them, connecting to the problem of 2010's cup, thinking by all of us, we get two believes:

Film-viewers(Users) have different hobbies to make different behaviors to rate different Movies.

Movies have different properties to be loved and rated by different Users.

Main Thoughts

With what we get from 2007's competition, We focus on one belief for 2010's:

Different students expert at different problem-steps(explained in the next silde),and the degree of the experting is marked by accuracy.

It contains two aspects:

First,students are different: they are separated into parts by properties themselves and their performance on the problem-steps which they tried to solve.

Second,problem-steps are different: they are separated into parts by their properties and their accuracy of being solved by student.

Some Assumptions

problem-step: Different problem may have same step. Only the combination of problem and step is unique. Take problem-step as notation.

Accuracy: whenever mentioned in these slides, the accuracy is means giving the problem-step accurate answer at first attempt.

Different Students(1): different properties themselves

Students have their own properties. As to time, students spend different time on a problem-step. The length of time represent their habits and proficiency which means whether the student is very good at the problem.

Unfortunately, last data we get at the competition, have no valid information about time. Therefore, this property has to be aborted.

Different Students(2): different performance

Different student have different performance on these (in our file student_table.txt):

ARTTRIBUTE	MEANING
Anon_Student_Id	Student id
All_Step_Count	The count of problem-steps the student take
All_Rate_Correct_First	The rate* of student get right answer to the problems
All_Std_Dev_First	Standard Deviation of correct first attempt
Sum_Hints	Sum number of hints
Ave_Hints	Average number of hints
Max_Hints	Max Number of hints

Different Problem-steps

They are separated into parts by their properties and their accuracy of being solved by student.

ARTTRIBUTE	MEANING
Problem-Step Name	...
Correct_First_Attempt_Rate	The Problem-Step is done by all its students' Correct_First_Attempt_Rate
Step_Std_Dev	Correct_First_Attempt_Rate of standard deviation
Step_Ave_Hints	Average number of hints of all the students who do the problem-step needing
Step_Unit i	The number of steps in unit i
KCi,j	l(th) knowledge standard, j(th) knowledge component

About the rate(I)

As the web of the 2010 KDD CUP says, one student do not take all problem-step and one problem-step is not done by all students, i.e. the data matrix of student and problem-step is sparse. Therefore, for example, when we calculate the rate of one student getting current answers for problem-steps in unit i, if the student do nothing able the unit, how we get the rate? 0 or 1? Obviously, neither of them are not accurate.

Additionally, if both the rate of one student getting one current answer on one problem-step and the rate of another student getting 100 current answers on 100 problem-steps equal 100 percent. Obviously, that is not fair.

About the rate(II)

Similarly, this is another typically example—shooting. Short to say one man shoots once and gets one hit, the other shoots 100 times and gets 100 hits, who's better? According to Laplace's law of succession(proving Bayes' statistics /theorem) ,based on the former's performance, the rate he gets one more hit is $(1+1)/(1+2)$, while the later's is $(100+1)/(100+2)$.

Commonly, based on x hits of n times shots, we estimate the rate getting one next hit of next shot is $x+1/n+2$.

For more background knowledge,we severely recommend Bayes's Theory/Estimation. It deserve your deep study.

Probing Parameters

When clustering students, as well as clustering problem-steps, is uncertain. Namely, the parameter k in K-means algorithm is adjustable, the iterating times too.

Last we our best result at following conditions :

the number of students clusters is ,and...(some date is in our lab's computer/server, when I am back, I will know.)

RMSE

For a given user i and movie j

$$w_{ij} = \begin{cases} 0 & \text{if no rating given} \\ 1 & \text{otherwise} \end{cases}$$

$$\text{RMSE}^2 = \sum_{i, j} (w_{ij} - \hat{w}_{ij})^2$$

where \hat{w}_{ij} is the predicted value

Our Result


















- Team name: Heymar, which is pronounced heima(黑马) in chinese.
- To see the result at https://pslcdatashop.web.cmu.edu/KDDCup/results_full.jsp
- Our Username “Baoxing Huai” with password “ustc502”
- At time 23:22:01 of date 2010-6-8 ,about 40 minutes before the deadline, we get total/average RMSE 0.31999.

Algebra I 2008-2009	Bridge to Algebra 2008-2009	Total Score
0.332178	0.307802	0.31999

- Rank :
 - 12 in final submissions of all student teams with a fact sheet
 - 25 in Final submissions of all teams with a fact sheet

Attachment

Final submissions of all student teams with a fact sheet

Rank	Team Name	Cup Score	Leaderboard Score	Final Submission Time	Fact Sheet	Paper
1	 National Taiwan University	0.272952	0.276803	2010-06-08 23:46:50		
2	 Zach A. Pardos	0.276590	0.279695	2010-06-08 21:31:07		
3	 SCUT Data Mining	0.280476	0.284624	2010-06-08 23:25:27		
4	Y10	0.298006	0.304277	2010-06-08 23:00:54		
5	Shiraz	0.299574	0.310430	2010-06-08 10:55:22		
6	UniQ2	0.300009	0.307777	2010-06-08 19:39:52		
7	Baby	0.301864	0.311978	2010-06-08 07:16:14		
8	Atlantis	0.309573	0.316246	2010-06-08 21:52:09		
9	Green Ensemble	0.311678	0.324581	2010-06-08 10:59:42		
10	Troae	0.314485	0.326375	2010-06-08 06:49:30		
11	pozip@FRI	0.315700	0.325010	2010-06-07 12:47:44		
12	Heymar	0.323107	0.335412	2010-06-09 00:33:34		
13	MiloBing	0.344566	0.361959	2010-06-08 02:27:50		
14	ecnusei07	0.344790	0.362294	2010-06-08 04:09:14		

Final submissions of all teams with a fact sheet

Rank	Team Name	Cup Score	Leaderboard Score	Final Submission Time	Fact Sheet	Paper
1	 National Taiwan University	0.272952	0.276803	2010-06-08 23:46:50		
2	 Zhang and Su	0.273692	0.276790	2010-06-08 23:39:35		
3	 BigChaos @ KDD	0.274556	0.279046	2010-06-07 03:48:20		
4	Zach A. Pardos	0.276590	0.279695	2010-06-08 21:31:07		
5	Old Dogs With New Tricks	0.277864	0.281163	2010-06-08 23:49:11		
6	SCUT Data Mining	0.280476	0.284624	2010-06-08 23:25:27		
7	pinta	0.284550	0.289200	2010-06-08 22:14:55		
8	DMLab	0.285977	0.291296	2010-06-08 19:37:50		
9	FEG	0.288764	0.293141	2010-06-08 23:45:12		
10	FEG-K	0.289877	0.294338	2010-06-08 22:39:34		
11	Vadis	0.292382	0.297136	2010-06-07 05:05:40		
12	psweather	0.292794	0.297542	2010-06-08 04:59:54		
13	uq	0.295389	0.301525	2010-06-08 23:54:30		
14	EXL	0.296190	0.302679	2010-06-08 12:27:51		
15	Y10	0.298006	0.304277	2010-06-08 23:00:54		
16	Shiraz	0.299574	0.310430	2010-06-08 10:55:22		
17	UniQ2	0.300009	0.307777	2010-06-08 19:39:52		
18	Andreas von Hessling	0.300228	0.310936	2010-06-08 23:56:22		
19	Baby	0.301864	0.311978	2010-06-08 07:16:14		
20	grandprix	0.309506	0.316712	2010-06-07 08:49:00		
21	Atlantis	0.309573	0.316246	2010-06-08 21:52:09		
22	Green Ensemble	0.311678	0.324581	2010-06-08 10:59:42		
23	Troae	0.314485	0.326375	2010-06-08 06:49:30		
24	pozip@FRI	0.315700	0.325010	2010-06-07 12:47:44		
25	Heymar	0.323107	0.335412	2010-06-09 00:33:34		
26	MiloBing	0.344566	0.361959	2010-06-08 02:27:50		
27	DataKiller	0.344611	0.360974	2010-05-18 21:10:09		
28	ecnusei07	0.344790	0.362294	2010-06-08 04:09:14		
29	Kun Liu	0.455219	0.460331	2010-06-06 03:56:21		

Some others, not least important

- I thought our result is not good enough to post paper, though we were invited to post. However, notice the rank 29 in Final submissions of all teams with a fact sheet, he posts one. Something to think and learn.

Some detail to be ctd...

- Wait for some detail..