Cold Start Purchase Prediction with Budgets Constraints*

Ke Hu[1], Xiangyang Li[2], Chaotian Wu[3]

[1]Microsoft Corporation
[2]Beijing Didi Infinity Technology and Development Co.,Ltd,China
[3]Southwest Jiaotong University, China

Abstract.IJCAI-16 Contest Brick-and-Mortar Store Recommendation with Budget Constraints is about buyer nearby brick-and-mortar stores recommendation. The main task of this competition focuses on predicting nearby store buying action when users enter new areas they rarely visited in the past. The contest has two novelties: first, given huge amount of online user behavior with on-site shopping record of moderate size, we could investigate whether their correlation helps in recommending nearby stores. Second, every merchant's budget constraints are imposed on prediction.

We develop a set of useful features to capture the underlying structure of the data, including merchant-related feature, merchant-location feature and so forth. We also propose to learn topic features with embedding, which represent the user on-site shopping behavior pattern and nearby stores behavior pattern in a shared feature space.

We use gradient boosting decision tree as the purchase prediction base model, and use isotonic regression as the purchase probability calibrated model. We use cascade ensemble method, and then develop a merchant recommend framework to decouple the budget constrain. Finally, our team MCMC ranked #2 in the competition.

1 Introduction

With the development of technology of mobile devices, more and more people use their smart phones to enjoy location based services (LBS). People are getting more comfortable sharing their real-time locations with various location-based services, such as navigation, car ride hailing, restaurant/hotel booking, etc. Using the technology of data mining we may explore the value of user's behavior to help people finding their demand in location based services more conveniently.

So in this contest, Thesponsor provides the data accumulated on Tmall.com/Taobao.com and the app Alipay and we try to use these data on nearby store recommendation when users enter new areas they rarely visited in the past.

The dataset of users' shopping records at brick-and-mortar stores, named Koubei Table, provides 230496 users' 1081724 purchase records between June 2015 and November 2015. The Tianchi Platform also provides another 492927 users' purchase

records only for analysis but not for download.The dataset of Tmall.com/Taobao.com provides 963923 users' 44528127 purchase records.

There are two main problems in this contest：The first of them is how to build the relationship between user’s online behavior and user’s nearby storesbehavior. As the areas are completely new for the users, it's quite difficult to find use's interest. However, the users' data accumulated on Tmall.com/Taobao.com, named Taobao Table, is very rich and these data is surely be a reflection of user's interest. So we investigate to build the relationship between user’s online behavior and user’s nearby stores recommendation.The second problem is a set of budget constraints imposed on the recommender system. For instance, the service capacity or number of coupons is limited at the stores.

In order to solve the above problem, we consider this problem as a Binary classification problem firstly. Training target of the model is to find out whether a user would like to purchase a certain merchant's product in a certain location. At the beginning, we extract features including merchant-ralated features, interactive user features. In phase of extracting cold start users features, we extract topic embedding features of users from Taobao table, and then transfer these embedding features to Koubei table with the same feature space. Besides, we cluster the user topic embedding with K-Means model and use these categories' buy preference as features, thus improving the generalization ability. With extracted features, we use GBDT to model the purchase probability.

To solve the problem of budget constraints, we try to use GBDT prediction as the purchase probability at first. For a specificmerchant, the sum of a large number of user-location pairs can reflect the budget occupying to some extent. With the budget estimate, we could recommend a specific user number without exceeding the budget of merchant. However, with the increase of tree number, GBDT would shift of the predicted values away from 0 and 1, hurting calibration. Besides, numerous merchant does not have some many customers so the estimated probability is significant. Then we utilize Isotonic Regression to calibrate probabilities, and design a recommend framework, solving the problem of budget constraints.

2 Framework

2.1 Data Analysis

The test set have 473533 user-location pair, however, only 9.6% of user-location pairs appear in Koubei Table before November. It may caused by increasing dramatically of Koubei apps and sampling of data. As the user behavior pattern is different for cold start users and interactive users, we construct two train sets of these different type of user to make the data sample independent identically distributed(iid). Moreover, compared with cold start user, the user who has actions in the Koubei has much lower probability to purchase the non-interactive merchants. Therefore, we also remove these parts of user from cold start user train sets.

2.2 Data Processing

We split the training sets into two sets, 70% and 30% respectively. One is for local training and the other is for validation. Moreover, there is 56206 duplicate samples record in the raw data sets. We remove the duplicate data sets to avoid influencing the distribution.

2.3 Feature Extraction

We extract the feature from four aspects, including merchant-related features, merchant-location features, interactive user features, interactive user-merchant features. We mainly focused on cold start related features. We then analysis the feature and perform more feature engineering like smoothing. We then conduct feature selection by Random Forest.

2.4 Model Training

We build up two models for cold start users and interactive users. We use LDA and K-Means to model the cold start users, use GBDT to predict the user-merchant-location triplepurchase rate. Further, we useisotonic regression to perform the probability calibration, which could improve the accuracy due to the budget constraints.

2.5 Recommendation Framework

From isotonic regression, every user-merchant-location triplehas prediction probability. We first rank the pair according to the prediction. Moreover, we pass the pair one by one according to the order, accumulating the probability of the same merchant, and then filter the following pairs if the accumulated probability exceeds the budget constraints. Therefore, we could get the high probability user-merchant-location pairs and avoid exceeding the constraint to a large extent. The final results show that this prediction framework with probability calibration can effectively improve the accuracy.

3 Features

One important task of this competition was to predict the possibility that a user would purchase a merchant in a given location. Each instance of training and test set was a {user_id, location_id, merchant_id} triple. From our statistics, the user history preference, merchant-related characters, location-related characters would influence the use purchase action in the future. Thus, we designed features mainly from four aspects: merchant-related features, merchant-location features, interactive user features and interactive user-merchant features.

3.1 Merchant-Related Feature

Merchant features contained the properties of the merchant's information over 120 days. In the shopping logs, a merchant was bought by some different users in one or more locations. We described a merchant by many degrees such as different time periods, different locations, different user behavior pattern. We list some key features and illustrate our intention.

- Sales Time Interval Statistics

Firstly, we extracted use purchase time interval of a merchant. Secondly, we calculated statistic of the time interval of a merchant such as median, minimum, maximum, and standard deviation. As users might purchase a merchant many times, so we can easily extract the mean of time interval of a merchant for a user and calculated its statistics. These features characterize the behavior pattern of the buying users of merchant.

- Lifecycle

In the last 120 days,the lifecycle of a merchant could reflect the number of days which the merchant was bought. So we extracted the number of days which a merchant was bought and the number of days, the length of the time segment before first time when the merchant was bought to the day before the forecast time, and some other statistics.

- Merchant Budget

There was a table of merchant budget info which was very useful for prediction. The budget of a merchant represented for the planed cost of a merchant for its merchandise. The large budget merchantsmainly have high and stable purchase value.

- Lifespan

The first time and the last time from the purchase of users can both affect the subsequent purchase. For instance, a user will have a very small probability to purchase because he only had a single purchase four months ago. Thus, we need to obtain the median of the interval between the first purchase and the last day, the interval between the last purchase and the last day,and the interval between all the purchase time and the last day.

3.2 Merchant-Location Feature

To further character the merchant, we also design merchant-location feature. Merchant-location feature is the most importance information for our prediction. By our experiment of feature selection, we found that most of Merchant-Location features have a relatively high sensitivity. There are four parts that is count feature, ratio feature, rank feature and daily level feature.

- Count Feature

The number of records in a specific merchant-location by user level and purchase record level in different days. In general, the more of the purchase number, the precision of recommend would be higher. This is because users, especially cold start users, prefer to purchase popular merchants.

- Rate Feature

This is calculated by the user purchase numbers of merchant-locationdividing the user purchase numbers of the location. By the experiment, we find the rate feature is thestrongest feature in our predication. In fact, our target is to predict the rate in the December, thus this feature is significant. Moreover, we found that some of data have smaller purchase record but rather high rate, which could affect the recommendation. So we reduce the effects via simple smooth(1), and the smooth variables are different for different pro-portion.

$$\mathbf{Rate} = \frac{\boldsymbol{merchant} - \boldsymbol{location}_{user_{num}}}{\boldsymbol{location}_{user_{num}} + \boldsymbol{\partial}} (\boldsymbol{\partial}\mathbf{isvariable}) (1)$$

- Rank Feature

This includes the rank of the merchant and the location in the whole data, the rank of the merchant in the geographical location and the rank of the location in the merchant. The location fixed rank feature could character the preference when user comparing the merchants in the same location. The merchant fixed rank feature could weaken the prediction that the merchant-location is weak in the specific merchant.

- Daily Level Feature

From our statistic, 11.8% of merchant-location pair first appears in the last two weeks. With the dramatic dramatically increase of both user and merchant, it is significant to feed some daily level feature to learn. We design the feature as how many days the merchant is online, how many average active users every day.

3.3 Interactive User Feature

A user that bought from a merchant could purchase from it in the future. And it also worked in our contest. The accuracy rate of a record that a user bought from a merchant nearly all exceeded 20 percent by our experiment, which is rather high for recommendation. Here we list some of the representative ones.

- Purchase Time Interval

The statistical information of purchasing time from users reflects the frequency of their purchasing time, which could provide more information to show the case of them in the future for purchasing. In order to represent their frequency of time better,

we adopt the average,maximum,minimum, median and the standard deviation of time interval from the purchasing process.

- Repeat Purchase Behavior

We define the repeat purchase behavior as purchasing the same commodity for users in different time, and they can proceed to purchase some commodity later when they have the trend of repeat purchase. For a user, we first compute the average time interval for repeat purchase of a good. Then, we compute their statistical information, such as the average,maximum,minimum, median and the standard deviation.

- Recently Purchase Behavior

The purchase behavior of users in a period can influence the subsequent behavior and the effects from different periods are different. For instance, it was distinguished between the behavior of the purchase of a good the day before and 30 days before. So we need to obtain the number of records, the quantity of the type and days for the purchase of users 1 day before, 3 days before, 7 days before,10 days before, 15 days before,30 days before, 60 days before, 90 days before and 120 days before.

3.4 Interactive User-Merchant Feature

This computes the compositional features from a user and a good, which can show their statistical information when a user and a good appear simultaneously and can reflect their correlation better. The probability of purchasing it for a user in the subsequent time can be predicted by the correlation. Then we will show the key features as follow.

- Recently Purchase

Behavior Similar to the definition of user recently purchase behavior.

- Lifespan

Similar to the definition of merchant lifespan. Besides, we introduce the days between the first purchase and the last one when the user purchase it in this section.

- Purchase Rank

This indicates the rank of behaviors for the purchase of user. To some degree the time of in the recent purchase can affect the subsequent purchase.

- Purchase Category Density

This indicates the case of the purchase process of the user for the commodity. We obtain the categories of the previous day, the day and the day after when he can purchase it, which could show the density of the purchase. For the user, the probability of his purchase willbe lower if the number of category is bigger when he can purchase it in a certain time.

3.5 Feature Selection

We use random forest[Breiman et al., 2001] to perform feature selection. We shuffle every feature among samples in the offline testing test. It can show the sensitivity of every feature through random forest. Besides, it can reflect the nonlinear relationship between feature and label. Finally, we choose 62 out of 120 features to train GBDT model.

4 Base Models

We build GBDT model on cold start users and interactive users, respectively. In this topic, we mainly focus on cold start users purchase prediction. First, we build LDA and K-Means model on the Taobao table, and process the output as the feature of cold start users and non-interactive user-merchant pair. We then build GBDT on the merchant-related features extracted from Taobao Table combined with user-related features from Koubei Table. The optimized model was then used to make prediction for every user-merchant-location triple.

4.1 LDA

The base Latent Dirichlet Allocation(LDA)[Blei, et al., 2003] generates latent topics through document and word collinearity. LDA assumes a fixed number of topics. The topic distribution is drawn from a Dirichlei prior, and learn from likelihood of each word in the document. As a result, a topic is drawn from the topic distribution and the word is drawn from a topic-wise word distribution, which can be used to inference the topic-wise document distribution too [Grithsˊ*et al.*, 2004].

In our contest, as most of the user is cold start user, we have no user interactive data in Koubei table. Thus, it is difficult to predict the buying action of theseusers. However, 73% of these part user has action record in Taobao table. The chanllendge is to use transfer learning to explore the user pattern in Koubei target table from Taobao source table[Faisal*et al.*, 2012].

We could apply LDA to model the user latent space inTaobao table. The item in the Taobao can be treated as a word and the user can be treated as a document. Therefore, we use the LDA to model the user in our Koubei Table, We assume the following generative process for the user as d and Table1 illustrate LDA.

Table 1.LDA Algorithm

1. Choose $\theta_d \sim Dir(\alpha)$,$\alpha$Dirichlet distribution with parameter;
2. For each of the N words in d:
   (1) Choose a topic Z_n as Multinomial (d), a multinomial distribution of parameterθ_d;
   (2) Choose a word w_n from p($w_n|z_n,\beta$), a multinomial probability conditioned on the topic z_n

We can obtain the probability of all the documents as (2).

$$\mathrm{p}(\mathrm{D}|\alpha,\beta) = \prod_{1}^{n} \int p(\theta_d|\alpha) \quad * \\ (\prod_{n=1} N_d \sum_{z_d,n} p(z_d,n|\theta_d)p(w_d,n|z_d,n,\beta))\mathrm{d}\theta_\mathrm{d} (2)$$

We use the Gibbs sampling method as solver to get the topic vector for user d. Then, vector θ_d is used as the features of the user in Koubei table.

With the modeling method described above, the buying action and click action is counted with different weight to present co-linearity matrix of user and item. As the user-item matrix is very sparse, we use the user and item's category co-linear information instead. We use LDA to model the user of Taobao table, and the latent factor can be used to infer user buying behavior of Koubei table. Finally, we choose eight topics to be embedded as user feature vector in Koubei table. GBDT model can learn the association of user feature vector and specific merchant, and thus the interactive user preferences can be applied to cold start users through the features vector. Besides, the merchant can be represented by the feature vector of the bought users, and we calculate the cosine similarity of user and merchant to describe the user preference, which is also added as an additional feature.

4.2 K-Means

K-means clustering is a algorithms of vector quantization. It is used to partition n samples into k clusters, and each sample belongs to the cluster whose central sample is nearest. This results in a partitioning of the sample space into Voronoi cells.

To enforce the performance of user topic embedding, we also use K-Means to cluster to user to five clusters. In addition, we output every cluster's purchase rate of specific merchant-location to the GBDT input. This method is combined of user topic features and purchase rate features, it can model the cold start user purchase probability directly and improve the generalization ability, thereby it improves the ability of representation.

However, K-means clustering is difficult to choose the cluster center at the first round(NP-hard question), and this would result in unstable of choose k cluster center of user topic vectors. We utilize K-means++[Davidet al.,2007] to choose the initialized samples as a optimized trainer.

4.3 GBDT

Gradient Boosting Decision Tree (GBDT)[Friedman et al., 1999] is a machine learning algorithm which ensembles numerous weak decision tree predictors to make a final prediction. It builds the model in a stage-wise fashion like other boosting methods do. Learning the loss of previous tree, every tree of each iterative could enforce

the learning of previous wrong predicted samples. Finally, GBDT add the predicted value of interactive trees as the final prediction[Bertoni et al., 1997].

As the behavior pattern of cold start users and interactive users is different, we build different model on cold start users and interactive users, respective. From our data study, the purchase rate of merchant-location of cold start users and interactive users is quite different. Cold start users tend to choose popular or discount merchant, while interactive users prefer the merchant they have bought or similar to they have bought. Thus, we train two different models to improve accuracy.

To get the optimized GBDT model,we mainly tuned three parameters: the depth of the trees, the number of trees, as well as the shrinkage parameter. With the increase of features,we increase the depth of the trees gradually. At last, we choose to increase the numbers of trees and decrease the shrinkage, which helps it converge to a better optimizing point. In the final, the optimized GBDT parameters are trees=1000, shrinkage parameters=0.01 and depth=8.

5 Ensemble Model

In order to boost the performance, we propose a cascaded ensemble method to perform probability calibration model. The trained GBDT model is used as a base model, and the Isotonic Regression model is used as a calibrated model. Then, we design a merchant recommend framework to merge the predict result and decouple the budget constrain.

5.1 Probability Calibration

Every merchant in Koubei has a budget constraints, we should know how many user-merchant-location triples fill the budgetsin addition to the probability ranking. GBDT typically yields good accuracy and AUC metrics. However, boosting may yield poor squared error with the iteration. This will influence theprecision in our topic as the model could not predict whether the recommendation buyer fills the budget [Friedmanet al., 2000].

GBDT uses boosting as the ensemble method of the base learner dicision tree. In thetreatment of boosting as a max margin classifier like SVM,Schapire [Schapire et al., 1998] observed that to getthe max marginnear the decision surface, boosting related method would sacrifice the margin of the easier cases. This would shift the prediction away from 0 and 1, thereby hurting calibration. Besides, in our experimants, with the increased iteration of boosting trees, these shift becomes more significant and the predicted values tend to close to either side of the dicison surface.

To improve GBDT 's poor calibraton, we use Isotonic Regression to calibrate the base predictions made by GBDT model[Niculescuet al., 2005]. Isotonic regressoin assumes only that:

$$y_i = m(f_i) + \varepsilon_i \quad (3)$$

$$\hat{m} = \arg\min_z \sum \left(y_i - z(f_i)\right)^2 \quad (4)$$

The objective function can be solved by pair-adjacent violators (PAV) algorithm[Ayer et al., 1955] presented in Table 2.

Table 2.PAV algorithm for calibrating posterior probabilities from base model predictions

1. Input: training set (f_i, y_i) sorted according to f_i;
2. Initialize $\widehat{m_{i,i}} = y_i$, $w_{i,i} = 1$;
3. While ∃Is.t.$\widehat{m_{k,i-1}} \geq \widehat{m_{i,l}}$
   Set $w_{k,l} = w_{k,i-1} + w_{i,l}$
   Set $\widehat{m_{k,l}} = (w_{k,i-1}\widehat{m_{k,i-1}} + w_{i,l}\widehat{m_{k,i-1}})/w_{k,l}$
   Replace $\widehat{m_{k,i-1}}$ and $\widehat{m_{i,l}}$ with $\widehat{m_{k,l}}$
4. Output the stepwise constant function:
$\hat{m}(f) = \widehat{m_{i,j}}$, for $f_i < f \leq f_j$

We split the data a training set and validation set to 70% and 30% respectively. After GBDT is trained on the training set, the calibrated model is trained on validation set to fit Isotonic Regression. We use isotonic regression model afforded by sklearn package.

The advantage of Isotonic Regression model is that it does not assume any form for the target function. Thus, it could calibrate the baseprobabilitieseven predicted by strong model to posterior probabilities without losing the ability of representation.

5.2 Merchant Recommendation Framework

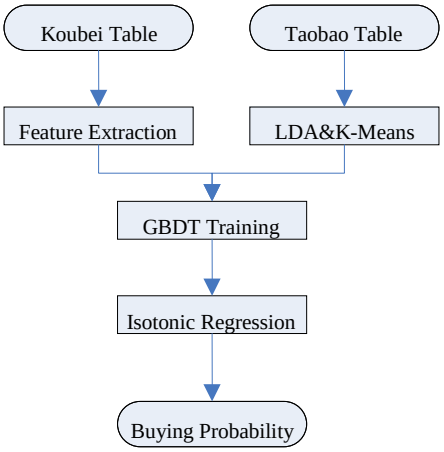

Fig. 1.

As described above, we extract merchant-related features from Koubei table, and combine with the user feature extracted from Taobao table. Then, we train GBDT and the base model, and train Isotonic Regression as the calibrated model. We use the

cascaded ensemble to perform the probability calibration. With this pipeline, we could get the purchase probability of every user-merchant-location triple.

Ensemble Models from isotonic regression, every user-merchant-location triplehas prediction probability. Firstly, we rank the tripleaccording to the prediction. Secondly, we pass the user-merchant-location tripleone by one according to the order. The corresponding merchant's budget will minus the purchase probability, and following merchant triples will be filtered if this merchant budget is exceeded. Finally, the user-merchant-location recommendation will be ended if the estimated probability is lower than the estimated precision. Therefore, we could get the high probability user-merchant-location triples and avoid exceed the budget constraint to a large extent. The final results show that this prediction framework with probability calibration can effectively improve the accuracy. We use table 3 to illustrate the recommendation framework pseudo-code.

Table 3. Lifting budget constraint using gbdt prediction for test data

1. Input: merchant budget table to nmap, model predict table to raw_ predict_result;
2. Initialize: sort_result = sort raw_predict_result by possibility in descending order, userLocCntMap={};
3. Merchant_budget_map(merchant_id, budget) = store table merchant budget info;
4. resultList = [], last_history_f_value = store the result of f value on line;
5. for record in sort_result:
   if (record.user,record.loc) in userLocCntMap and user LocCntMap[(record.user, record. loc)] >=10:
   continue;
   if nmap[record.merchant_id]>0:
   nmap[record.merchant_id]-= record.possibility;
   resultList.append(record);
   userLocCntMap[(record.user, record.Loc)]+= 1;
6. for record in resultList:
   if record.possibility> best F1 value in the recommend history / 2:
   recommend the mechant to the user-location pair

5.3 Experiment Result

Table 4 shows the F1 score of all methods, including single GBDT model, GBDT model with LDA and K-Means user features, GBDT model with user features cascaded with Isotonic Regression. As this topic has budget constaints, we have not used AUC or other offline metrics. We can observe that the LDA&K-Means model can improve the performance. Moreover, the cascaded ensemble method achieves the best performance.

Table 4. The F1 score of different methods

No	Method	F1
1	Single GBDT	0.4472
2	GBDT+LDA&K-Means	0.4513
3	Ensemble with Isotonic Regression	0.4638

6 Conclusion

In this paper, we introduce our solution at ICJAI 2016 Contest Brick-and-Mortar Store Recommendation with Budget Constraints. First, we design efficient merchant related feature system based on historical behavior. Moreover, we propose GBDT, LDA, K-Means, Isotonic Regression model to get purchase probability. Then wedesign a recommendation framework to decouple the budget constraint and get the high precision pair. Our experiments show that they all have a good performance. Finally, our MCMC Team ranked #2 in the competition.

Acknowledgements

We sincerely thank the IJCAI2016 and Tianchi Platform for giving us this opportunity to take part in the competition, we have learned much knowledge from this competition. We would also thank the staff of the competition for their help.Thanks to our opponents, so can we keep making progress.

References

[Bertoni *et al.,* 1997] Bertoni, A., Campadelli, P.,&Parodi, M. (1997).*Aboosting algorithm for regression*. InW.Gerstner,A.Germond,M. Hasler,&J.-D. Nicoud (Eds.), LNCS, Vol. V: Proceedings ICANN'97: Int. Conf. on Artificial Neural Networks(pp. 343–348). Berlin: Springer.

[Blei, et al., 2003] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. *Laterty. Latent dirichletallocation*.Journalof Machine Learning Research, 3:993–1022, 2003.

[David*et al.,*2007] David Arthur and Sergei Vassilviskii. *K-means++: The Advantages of Careful Seeding*

[Friedman*et al.,* 2000] J. Friedman, T. Hastie, and R. Tibshirani.*Additive logistic regression: a statistical view of boosting*. The Annals of Statistics, 38(2), 2000.

[Niculescu*et al.*, 2005] Niculescu-Mizil, A., &Caruana, R. (2005). *Obtaining calibrated probabilities from boosting.* Proc. 21th Conference on Uncertainty in Artificial Intelligence (UAI '05). AUAI Press.

[Griths*et al.*, 2004] T. L. Griths, M. Steyvers, *Finding scientific topics*, Proceedings of the National Academy of Sciences 101 (2004) 5228–5235

[Faisal*et al.*, 2012] A. Faisal, J. Gillberg, J. Peltonen, G. Leen, S. Kaski, *Sparse nonparametric topic model for transfer learning*, in: Proceedings of the 20th European Symposium on Artificial Neural networks, Computational Intelligence and Machine Learning, 2012.

[Friedman *et al.*, 1999] Friedman, J. (1999). *Greedy function approximation: a gradient boosting machine.* Technical Report, Department of Statistics, Stanford University.

[Schapire *et al.*, 1998] R. Schapire, Y. Freund, P. Bartlett, andW. Lee. *Boosting the margin: A new explanation for the effectiveness of voting methods*. Annals of Statistics, 26(5):1651–1686, 1998.

[Ayer *et al.*, 1955] M. Ayer, H. Brunk, G. Ewing,W. Reid, and E. Silverman. *An empirical distribution function for sampling with incomplete information.* Annals of Mathematical Statistics, 5(26):641–647, 1955.

[Breiman *et al.*,2001] L. Breiman. *Random Forests*. Machine Learning, pages 5-32, 2001.