Comparison of Two Approaches to Recommender Systems with Anonymous Purchase Data*

Yuri Zhuravlev[1], Alexander Dokukin[1], Oleg Senko[1], Dmitry Stefanovsky[2], Ivan Saenko[3], and Nikolay Korolev[3]

[1] FRC CSC RAS, Moscow, Russia
zhur@ccas.ru, dalex@ccas.ru, senkoov@mail.ru
[2] IRIAS, Moscow, Russia
dstefanovskiy@gmail.com
[3] Moscow State University, Moscow, Russia
i.a.saenko@mail.ru
[4] Moscow State University, Moscow, Russia
nikolay.korolev.s@gmail.com

Abstract. The paper discusses problem of assessing customer preferences profiles for goods offered by retail store. A profile is assessed on the condition that a certain set of goods was previously purchased by the buyer. It is supposed that information about previously purchased products is contained in sales receipt. The trivial technique where preference rating is calculated as common frequency of occurrence of products in sales receipts is compared with a technique using combinations of association rules and a new method involving product clustering at initial stage. Preference rating is at that evaluated by combined feature description that includes the description of the evaluated product and the description of the set of goods present in sales receipt. Feature description of evaluated product is a vector of proximity measures to clusters received by agglomerative hierarchical clustering. Feature description of a set of goods is vector of mean proximity measures between clusters and goods from this set. Gradient boosting method was used to distinguish goods which were and were not actually purchased by combined feature description. At that purchasing probabilities estimates that are returned by recognition algorithm are considered as preference ratings. ROC analysis is used to compare efficiency of three techniques.

Keywords: Recommender system · Machine learning · Feature extraction · Gradient boosting.

1 Introduction

The purpose of this article is to study the effective use of data to personalize offers to retail customers, mainly with online purchases. The study presented in

* This work was supported in part by the Russian Foundation for Basic Research, projects no. 18-29-03151, 18-01-00557.

this work is a continuation of the studies previously presented in the works [1] and [2]. The discussed method is aimed at taking into account the entire content of the sales receipt, and not only information about the most popular products as in [2]. In general, the model of the relationship between the retailer and the buyer is as follows: the appearance of the client is considered consistent, so the seller offers each buyer a range of products, then the client decides whether to make a purchase. The retailer may encounter restrictions in terms of display or capacity, which limit the number of products in the offer. The retailer's goal is to maximize the expected total revenue for the sales season. Recommender systems are a popular tool aimed to give a customer an advice which good in the best way corresponds to his/her demands [3]. Many techniques can be used to implement one. Context based systems use some supplementary information about customers or goods. However such information is often hard to achieve. Another approach employs information about customer's preferences expressed by them in one way or another. In the latter case some very efficient mathematical tools involving matrix decomposition can be used. But getting client's preference data is associated with additional costs in offline shopping. Finally, recommendations can be based on digital traces left by anonymous customers, i.e. the set of customers' receipts registered up to a certain point. Additionally in the simplest case all identification is based on a set of goods being bought. Segmentation or clustering is the key to effective personalization and identifying preferences.

We illustrate the practical value of a clustering policy in real conditions using a dataset from a major Russian retailer. The data set consists of roughly ten thousand cosmetic and related goods purchased in different combinations in about one hundred thousand transactions over two months period. We compare the effectiveness of the proposed policy with a data-intensive policy that ignores any potential similarity of preferences in different profiles and, thus, evaluates the product preferences for each profile individually. Namely we discussed as baseline approaches using association rules. The simplest baseline method is frequency based algorithm that will be referred to as A_F. Algorithm A_F calculates receipt owner preference ratings for item Y as fraction of receipts with Y among all receipts.

2 Data Set

Each receipt Z can be described as a binary vector of length N corresponding to a total number of products sold by a certain retail organization. The vector's elements corresponding to products sold in that particular transaction are marked as ones, and the rest are marked zeros.New customer's recommendations are made with the help of algorithm that was generated on the basis of the previously collected receipt data and the currently performed transaction which can be described in the same binary form. Efficiency of algorithms was studied at data set that is collection of 98500 receipts. Total number of products N was about 6000.

3 Ensembles of Association Rules

Exactness of A_F is not high because it ignores all information about previously purchased goods. So more complicated method using ensembles of association rules was also considered. Association rule is a way of measuring consequence like relationships between objects [4]. In this case the relation between a product X being bought, i.e. $Z_0(X) = 1$, and a product Y to be recommended. Here, each receipt Z is described as a binary vector of length N corresponding to a total number of products. New customer's recommendations are made on the basis of the previously collected receipt data and the currently performed transaction Z_0 which can be described in the same binary form.

Let's denote $S(\{X\})$ and $S(\{Y\})$ the subsets of all receipts S containing product X or Y correspondingly, whereas $S(\{X, Y\})$ will denote a subset containing both goods. The support of X is then defined as $Sup(X) = \frac{|S(\{X\})|}{|S|}$, whereas confidence of the X, Y pair is $Conf(X, Y) = \frac{|S(\{X,Y\})|}{|S(\{X\})|}$. Pairs of objects with large enough values of both criteria form association rules which can be used to estimate probability of buying product Y subject to product X purchase. The conclusion may be made based on a single best association rule or by their ensemble. Support and confidence are calculated for all products from the training set S_T and their pairs respectively.

The associative rules preference ratings for the owner of Z_j receipt can then be calculated as

$$A_{AR}(Y, Z_j) = \frac{1}{|\{X_1^j, \ldots, X_r^j \mid Sup(X_i^j) > 0\}|} \times \times \sum_{X \in \{X_1^j, \ldots, X_r^j \mid Sup(X_i^j) > 0\}} Conf(X_i^j, Y)\,.$$

In situation when sales receipt is absent preference rating for item Y may be evaluated as support value only and association rules algorithm is reduced to A_F, i.e.

$$A_F(Y, Z_j) = Sup(Y)\,.$$

4 Methods Based on Clustering

When using clustering techniques the set of binary receipt vectors is divided into several groups or clusters in which the digital traces are considered close to each other in terms of a selected metric. Then the Y product's preference rating can be calculated by frequencies present in the cluster containing the Z_0 trace.

Clustering methods are used in recommendation systems to select groups of customers with similar preference profiles [5]. Here we suggest another technique where clustering is used to select groups of complementary products. The derived set of clusters is further used to generate multidimentional feature description of products. Such descriptions allow effective application of machine learning tools.

The authors of the present research have already shown that agglomerative hierarchical grouping method applied to the described binary data produces well interpreted set of product clusters [1]. At that the chi-squared metric [6] was used to evaluate proximity measures between goods that are described by binary vectors.

Chi-squared metrics. Let's consider two arbitrary binary vectors x, y of the same length.Let's denote by a the number of positions in which both vectors are ones, i.e. xy^T.Let's define b, c, d in the similar fashion, i.e. $b = x(\bar{1} - y)^T$, $c = (\bar{1} - x)y^T$, $d = (\bar{1} - x)(\bar{1} - y)^T$, where $\bar{1}$ is a vector of ones of the same length as both x and y.The chi-squared metric ρ is then defined as

$$\rho(x, y) = \frac{(ab - cd)^2 sign(ab - cd)}{\sqrt{(a + b)(b + c)(c + d)(d + a)}}.$$

To evaluate similarity measure between clusters C_i and C_j sum $\rho_{UA}(C_i, C_j)+\rho_{CL}(C_i, C_j)$ is used. Metric

$$\rho_{UA}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{X \in C_i} \sum_{Y \in C_j} \rho(X, Y)$$

corresponds to unweighted average linkage clustering while metric $\rho_{CL}(C_i, C_j) = \min_{X \in C_i, Y \in C_j} \rho(X, Y)$ implements complete linkage clustering. Using $\rho_{CL}(C_i, C_j)$ prevents merging big clusters. Combining of two metrics allows to control distribution of clusters by size and thus to receive optimal set of clusters providing exact estimation of preference rating.

Let we have L non-intersecting clusters $C_1, \ldots, C_L$ in the S_T set. The distance of product Y to the i-th cluster is calculated as

$$P(Y, C_i) = \frac{1}{|C_i|} \sum_{X \in C_i} \rho(Y, X). \quad (1)$$

Vector

$$\mathbf{P}(Y) = [P(Y, C_1), \ldots, P(Y, C_L)] \quad (2)$$

can serve as a good feature description of the product Y since it is continuous and it reflects customer's interest to the product in terms of his interest to different clusters of products.

Method Based on Top Products. In our previous studies[2], we consider method for calculating customer preference estimates based on clustering, where preference of a product for the owner of receipt Z is calculated by concatenating descriptions of r top products from Z and description of evaluated product Y. Let Z_j be an arbitrary receipt where $X_1^j, \ldots, X_r^j$ are the top r goods in the order of decrease of their frequency in the S_T. Preference rating for some product Y that is absent in some receipt Z_j was calculated by $X_1^j, \ldots, X_r^j$ with the help of algorithm that is ensemble of decision trees. This algorithm returns estimates of probability that the customer that received receipt Z_j will purchase

product Y. The input of algorithm was combined feature description that is $[\mathbf{P}(X_1^j), \ldots, \mathbf{P}(X_r^j), \mathbf{P}(Y)]$. The task of algorithm receiving was reduced to standard pattern recognition task with two classes. Several machine learning were tried. But the best efficiency was achieved for gradient boosting method generating decision trees ensembles. It was shown by experiments that method based on clustering and gradient boosting outperforms the trivial algorithm A_F and algorithm based on ensembles of association rules in terms of areas under ROC curves.

Method based on average proximity measures. There are two drawbacks to method based on r top products. At first trained algorithm cannot be applied when number of products in receipt less than r. At second all other information about the sales receipt besides information about top products is actually lost. In this paper method is discussed, where the preference of the product Y for the owner of receipt Z is calculated by averaging descriptions of all products from Z combined with the description of evaluated product Y.In other words the preference is calculated by combined description $[\mathbf{P}^a(\widetilde{X}), \mathbf{P}(Y)]$, where

$$\mathbf{P}^a(\widetilde{X}) = \frac{1}{k} \sum_{i=j}^{k} \mathbf{P}(X_j) \tag{3}$$

and $\widetilde{X} = \{X_1, \ldots, X_k\}$ are all goods from sales receipt Z. As in the previous work [2] the was reduced to standard pattern recognition task with two classes. The trained algorithm returns probability that evaluated product belongs to group of purchased products that are used as preference ratings of products.

Experiments. The original set of checks was randomly divided into subset S_T that was used to generate training set $\widetilde{S}_t$ and subset S_C that was used to generate test set $\widetilde{S}_c$. The represented below procedure was used to generate training set $\widetilde{S}_t$ from S_T.

- Step 1. Hierarchical grouping method is used to receive optimal clustering of products $\{C_1, \ldots, C_l\}$ by full initial training set S_T.
- Step 2. Set S_T^r is selected from initial training set S_T. Each receipt from S^r includes at least r products.

Steps 3-6 are repeated m times where m is size of training set.

- Step 3. Sales receipt Z is randomly selected from S_T without returning and $k_1 + k_2$ products $Y_1, \ldots, Y_{k_1+k_2}$ are randomly chosen from Z without returning, where k_1 must be fixed less r. At that first k_1 products $Y_1, \ldots, Y_{k_1}$ are selected from set of products that are present inside Z and rest products $Y_{k_1}, \ldots, Y_{k_1+k_2}$ are selected from set of products that are not present inside Z.

  Steps 4 and 5 are implemented for each product from the set $\{Y_1, \ldots, Y_{k_1}\}$.

- Step 4. If Y_j is inside set of products $\{X_1 \ldots, X_r\}$ that are present in Z or $j \leq k_1$ then $\widetilde{X} = \{X_1 \ldots, X_r\} \setminus Y_j$ and $\widetilde{X} = \{X_1 \ldots, X_r\}$ otherwise.

– Step 5. Vector description of product $\mathbf{P}(\mathbf{Y_j})$ is calculated according to (1,2) vector description $\mathbf{P}(\widetilde{X})$ is calculated according to (3).
– Step 6. Concatenation of $\mathbf{P}(\widetilde{X})$ was labeled by 1 if Y_j is present in Z and is labeled by 0 otherwise.
– Step 7. Labeled concatenation is added to training set $\widetilde{S}_t$.

The procedure was used to generate $\widetilde{S}_t$ for $k_1 = 1$, $k_2 = 4$ and $m = 15000$. Efficiency of several machine learning techniques in two-class pattern recognition task with training set $\widetilde{S}_t$ was evaluated using multifold cross-validation and by control set $\widetilde{S}_c$ that was generated from S_C using practically the same procedure that was used to generate $\widetilde{S}_t$ but with parameters $k_1 = 1$,$k_2 = 1000$ and $m = |S_C|$. In other words preference rating for product Y_j" in a sales receipt Z in S_C is calculated by combination of feature description of product Y_j" and feature description of set of products really purchased in Z and different from Y_j" if Y_j" was purchased also. Feature description were calculated using set clusters $\{C_1, \ldots, C_l\}$ found in S_T. Tried machine learning methods include logistic regression, support vector machines, decision forests with gradient boosting[7]. But the best result was received for Light gradient boosting (LightGBM)[9]. The algorithm based on clusterization and LightGBM will be reffered to as A_{ML2}.

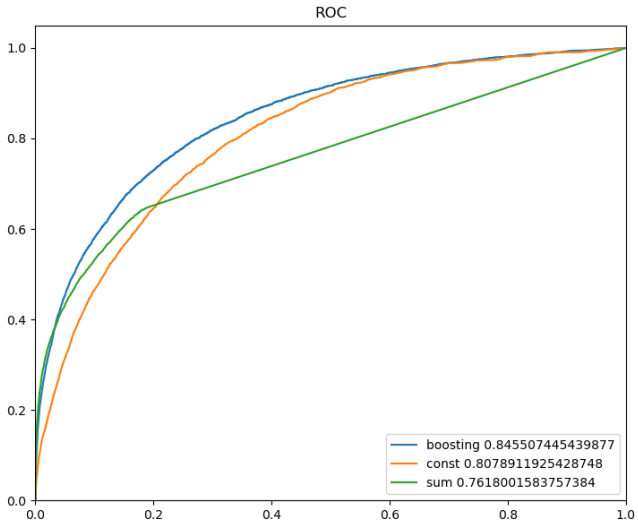

Fig. 1. ROC curves comparison. ROC AUC values for different methods are given in right low corner.

The ROC curve [8] for A_{ML2} is shown at the figure (1) together with ROC curves for A_{AR} (associative rules) and A_F (frequency based). In the legend "boostings" stands for A_{ML2}. It is seen from figure (1) that the ROC curve for A_{ML2} runs noticeably higher the ROC curve for A_F at interval for FPR from

0 up to 0.6. At that the ROC curve for A_{ML2} practically coincides with ROC curve for A_{AR} at interval from 0 to 0.07.

5 Conclusion

A new method has been developed for estimating customer preferences by the anonymous cash receipts data. The experiments indicate the prospects of the proposed approach. Firstly, the effectiveness of the proposed method turned out to be slightly higher than the effectiveness of reference methods. Evaluation was performed by ROC AUC. Secondly, it is important to mention that though the ROC AUC values of the proposed method and the frequency based algorithm are quite close the experiments showed significant differences between their recommendations in terms of goods. The machine learning algorithm former suggests rarer products which may be advantageous for the shop owner. Also, the low correlation between recommendations calculated by three technique indicates that ensemble of methods might provide some improvement in future research.

References

1. Zhuravlev, Yu., Dokukin, A., Senko, O., Stefanovskiy, D.: Use of Clasterization Technique to Highlight Groups of Related Goods by Digital Traces in Retail Trade. Proceedings of 9th International Conference on Advanced Computer Information Technologies ACIT-2019, 84–88 (2019)
2. Zhuravlev, Y., Dokukin, A., Senko, O., Stefanovsky, D., Saenko, I.: On a Novel Machine Learning Based Approach to Recommender Systems // In S. Balandin, I. Paramonov, T. Tyutina. Proceedings of the FRUCT'26, Yaroslavl, Russia, 23-25 April 2020, FRUCT Oy, Finland, 675–681 (2020)
3. Sohlberg, H.: Recommending new items to customers - a comparison between Collaborative Filtering and Association Rule Mining. Master's Thesis. Stockholm: KTH Royal institute of technology.school of computer science and communication (CSC) (2015)
4. Sun, X., Kong, F., Chen, H.: Using Quantitative Association Rules in Collaborative Filtering. In: Fan W., Wu Z., Yang J. (eds) Advances in Web-Age Information Management. WAIM 2005. Lecture Notes in Computer Science **3739**, (2005).
5. West, J.D., Wesley-Smith, I., Bergstrom, C.T.: A recommendation system based on hierarchical clustering of an article-level citation network. IEEE Transactions on Big Data **2**(2)113–123 (2020)
6. Choi, S.-S., Cha, S.-H., Tappert, C. C.: A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics **8**(1), 43–48 (2010)
7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media (2013)
8. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters **27**, 861–874 (2006)
9. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NIPS 2017), 3149–3157 (2017).