Profit Estimation Error Analysis in Recommender Systems based on Association Rules

Gurdal Ertek, Xu Chi, Gabriel Yee, Ong Boon Yong, Byung-Geun Choi

Dr. Gürdal Ertek's Publications

Ertek, G., Chi, X., Yee, G., Yong, O. B., Choi, B.-G., "Profit Estimation Error Analysis in Recommender Systems based on Association Rules". In Proceedings of 2015 IEEE International Conference on Big Data (Big Data). (2015) 2138 - 2142.

Profit Estimation Error Analysis in Recommender Systems based on Association Rules

Gurdal Ertek
Rochester Institute of Technology - Dubai
Dubai Silicon Oasis, Dubai, UAE
Email: gurdalertek@gmail.com

Xu Chi
Gabriel Yee
Ong Boon Yong
Byung-Geun Choi
Singapore Institute of Manufacturing Technology
Singapore
Email: cxu@simtech.a-star.edu.sg
Email: qmyee@simtech.a-star.edu.sg
Email: ongby@simtech.a-star.edu.sg
Email: byung-geun.choi@epfl.ch

***Abstract*—It is a challenge to estimate expected benefits from recommender systems based on association rule mining. This paper aims to address this challenge and presents a study of buying preferences of a sample of retail customers. It reveals a monotonic, non-linear relationship between the expected profits (as a function of information loss) and minimum support threshold levels, when considering transactions for a recommender system based on association rules. This finding is significant for recommender systems that utilize potential profits as a decision-making criterion.**

***Keywords*—recommender systems, association mining, association rules, profit estimation, retail industry.**

I. Introduction

Association mining is one of the popular data mining methodologies used in practice [1]–[8], especially in constructing and implementing recommender systems. In retail and merchandising, it is used to identify rules or relationships between products and then make recommendations for new products to consumers based on the products that they have previously selected. Such recommender systems can thus be viewed as informal "word-of-mouth" systems that are adjusted by consumer profiles, behaviors, and matrices decided by its owner [9]–[11]. A well-known example of such a system is one employed by the online retailer Amazon and recommendations are exposed to users under the header "Customers Who Bought This Item Also Bought". Offline retailers can also employ association mining for developing rules and policies that can increase sales and profits [12]–[14] such as decisions on shelf layout, bundles or last minute offerings near the cashiers can be made based on items which are bought together frequently, such as peanut butter and grape jelly [15], can be found as neighboring items on the shelves because they complement each other.

Despite the vast literature on association mining and recommender systems [1], [2], [16]–[21], we found a lack in research efforts to quantify the effects of information loss when varying minimum support thresholds. To illustrate the significance of this issue, consider the association graph given in Figure 1. The association graph uses results calculated for the dataset presented in Ertek, Demiriz and Cakmak [23] and

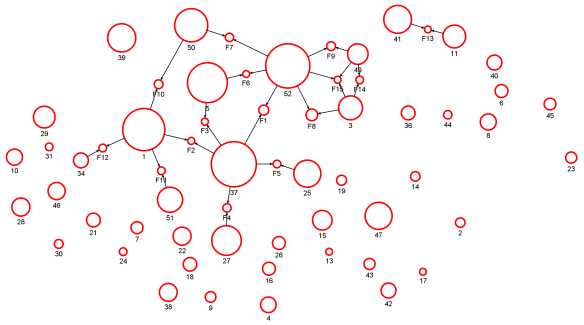

Fig. 1. Association graph for the Ertek, Demiriz and Cakmak [23] dataset, when $minsup = 0.01$

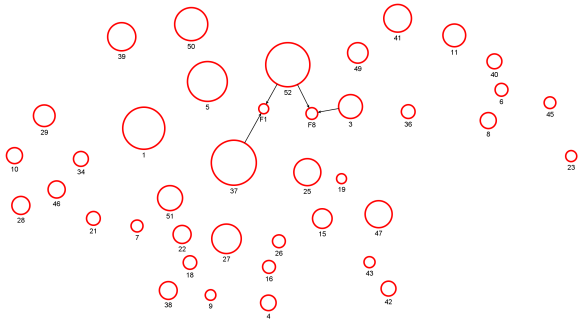

Fig. 2. Association graph for the Ertek, Demiriz and Cakmak [23] dataset, when $minsup = 0.02$

is constructed using a visualization scheme introduced in [22] and improved in [23]. The software tool used is AssocMiner [25]. In Figure 1, the graph visualizes the results where the minimum support value is set to $minsup = 0.01$. Now, also consider Figure 2 where the minimum support value is set to $minsup = 0.02$. It is easy to observe that the number of connected nodes (items and itemsets) and arcs (frequent itemset relations) becomes less, and thus it can also be expected that any profit estimate derived from the considered nodes and arcs in both graphs would give different results. Therefore, it is a challenge to strike a balance between minimum support

threshold levels and computational time needed to yield a best recommendation for a user.

In this paper, we present a mathematical model to address the challenge of estimation of expected profits. Then we present the outcomes of the estimation against varying levels of minimum support thresholds.

II. Background

When developing recommender systems, a common problem that arises would be finding the "balance" in the tradeoff between minimum support threshold levels and the computational time needed to yield the best recommendation for a user. Using a $minsup = 0$ value allows for a scenario of *complete information* in which all items and itemsets in the dataset are considered amounting to a need to make many evaluations before a recommendation can be computed. In practice, $minsup > 0$ values are often used as a compromise to reduce computational time. The obvious drawback is therefore that *incomplete information* is used to make a recommendation. However, using *incomplete information* is not necessarily a bad thing as it allows the recommender system to neglect the many rare-occurrence transactions that exist in the long tail of demand and revenue [26].

III. Mathematical Model

In order to evaluate the effects of varying $minsup$ values and expected profits, we first develop a mathematical model to compute the case of $minsup = 0$ (Case A) and when $minsup > 0$ (Case B). Note that the estimations in both cases are for the profits before any recommendations are made. That is, the formulas do not account for the possible additional sales due to recommendations of a recommender system.

A. Sets

Let $\mathcal{N}$ be a set of items, where $\mathcal{N} = \{1, \cdots, N\}$. The index used is $n = 1, \cdots, N$, where n is integer. Let $\mathcal{M}$ be a subset of $\mathcal{N}$.

B. Parameters

Let $P(n)$ or P_n denote the probability of the customer buying item n only, without buying any other items. Let $P(\mathcal{M})$ or $P_{\mathcal{M}}$ denote the probability of buying only the items $n \in \mathcal{M}$ in an itemset $\mathcal{M} \subseteq \mathcal{N}$, but no other items. From basic probability theory

$$\sum_{\mathcal{M} \subseteq \mathcal{N}} P_{\mathcal{M}} = 1.$$

Let $S(n)$ or S_n denote the *support*, i.e., probability of the customer buying item n, possibly also buying other items. Let $S(\mathcal{M})$ or $S_{\mathcal{M}}$ denote the support of the itemset $\mathcal{M}$, i.e., the probability that the items $n \in \mathcal{M}$ in the itemset $\mathcal{M} \subseteq \mathcal{N}$ are bought, possibly with other items.

The relation between the probabilities and support values can be expressed as follows:

$$S_{\mathcal{M}} = \sum_{\mathcal{M} \subseteq \mathcal{M}'} P_{\mathcal{M}'}. \tag{1}$$

where the support value of an itemset is equal to the summation of the probability values of all the supersets of that itemset.

Let $\gamma(n)$ or γ_n denote the profit from selling one unit of item n to the customer.

C. Assumptions

An important assumption made is that each customer will only choose one unit of item in a transaction, allowing us to perform binary association mining. For the dataset in Ertek, Demiriz and Cakmak [23], [24] this is valid as the customers are making the transactions for personal consumption. In such an arrangement, purchasing more than one unit would not be meaningful. In more extreme examples, such as in the purchase of digital audio or videos, online retailers could also restrict the ability to buy multiple units of the same item.

D. Expected Profit for Case A

Let the expected profit per customer be $E^A(P, \gamma)$, when the minimum support threshold $minsup$ is set equal to 0. If we had only $N = 2$ items, the expected profit would be

$$E^A(P, \gamma) = \gamma_1\Big(P_{\{1\}} + P_{\{1,2\}}\Big) + \gamma_2\Big(P_{\{2\}} + P_{\{1,2\}}\Big).$$

If we had $N = 3$ items, the expected profit would be

$$\begin{aligned} E^A(P, \gamma) = \ & \gamma_1\Big(P_{\{1\}} + P_{\{1,2\}} + P_{\{1,3\}} + P_{\{1,2,3\}}\Big) \\ & + \gamma_2\Big(P_{\{2\}} + P_{\{1,2\}} + P_{\{2,3\}} + P_{\{1,2,3\}}\Big) \\ & + \gamma_3\Big(P_{\{3\}} + P_{\{1,3\}} + P_{\{2,3\}} + P_{\{1,2,3\}}\Big). \end{aligned}$$

By using (1),

$$E^A(P, \gamma) = \gamma_1 S_{\{1\}} + \gamma_2 S_{\{2\}} + \gamma_3 S_{\{3\}}.$$

In the general case N items, we would have

$$\begin{aligned} E^A(P, \gamma) &= \sum_{m=1}^{N} \sum_{\mathcal{M} \vdash C_{1,m}} \gamma_m P_{\mathcal{M}} \\ &= \sum_{m=1}^{N} \gamma_m S_m \end{aligned} \tag{2}$$

where $\mathcal{M}$ satisfies the condition $C_{1,m}$:

$$\mathcal{M} \vdash C_{1,m} \Leftrightarrow (\mathcal{M} \subseteq \mathcal{N}) \wedge (m \in \mathcal{M})$$

E. Expected Profit for Case B

When the minimum support threshold is set to a positive value $minsup > 0$, the expression will be very similar, except for the condition under the second summation. The condition will change to $C_{2,m}$ to now include only the probabilities and joint probabilities whereby their respective *support* are greater than or equal to $minsup$. More formally presented, the expected profit per customer $E^B(P, \gamma)$, when the minimum support threshold $minsup$ is positive, is given as

$$E^B(P, \gamma) = \sum_{m=1}^{N} \sum_{\mathcal{M} \vdash C_{2,m}} \gamma_m P_{\mathcal{M}} \tag{3}$$

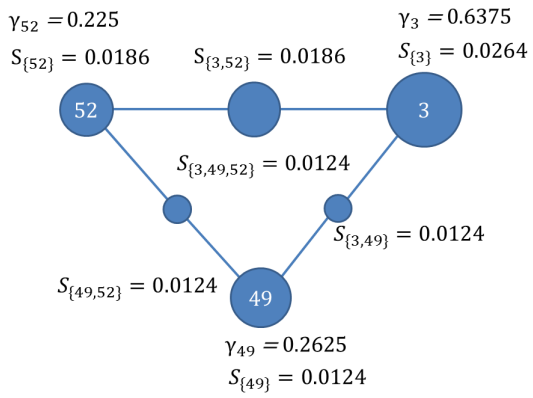

Fig. 3. Association graph for the numerical example, where support values for items and itemsets, and profits for items are shown.

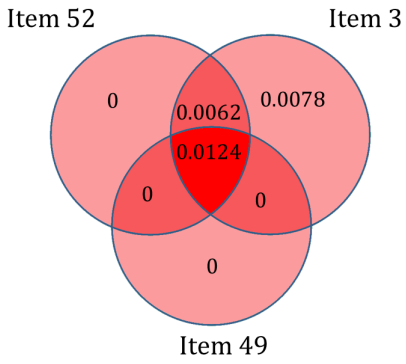

Fig. 4. Venn diagram for the numerical example, where probabilities used in profit calculation are shown for items and itemsets.

where $\mathcal{M}$ satisfies the condition $C_{2,m}$:

$$\mathcal{M} \vdash C_{2,m} \Leftrightarrow (\mathcal{M} \subseteq \mathcal{N}) \wedge (m \in \mathcal{M}) \wedge (S_{\mathcal{M}} \geq minsup)$$

Since some of the subsets $\mathcal{M}$ are eliminated in E^B, $E^A \geq E^B$.

F. Estimation Error for the Profit Function

The estimation error is introduced when a minimum support threshold can be found through evaluating the *profit ratio*

$$\rho = E^B(P,\gamma)/E^A(P,\gamma) \tag{4}$$

where $0 \leq \rho \leq 1$, and $\rho < 1$ if the profit is underestimated in recommender systems.

IV. Numerical Example

Consider the numerical example given in Figure 3. The numerical example is a subset of the data from Ertek, Demiriz and Cakmak [23]. Assume that this is a data instance with $N = 3$ (Item 3, Item 49, Item 52), the *support* is adjusted using (1). The Venn diagram that shows the probabilities is also provided in Figure 4. Let us assume that we run association mining for this data with $minsup = 0.015$.

The profit values under Case A and Case B are

$$\begin{aligned} E^A = \; & \gamma_3\Big(P_{\{3\}} + P_{\{3,49\}} + P_{\{3,52\}} + P_{\{3,49,52\}}\Big) \\ & + \gamma_{49}\Big(P_{\{49\}} + P_{\{3,49\}} + P_{\{49,52\}} + P_{\{3,49,52\}}\Big) \\ & + \gamma_{52}\Big(P_{\{52\}} + P_{\{3,52\}} + P_{\{49,52\}} + P_{\{3,49,52\}}\Big) \\ & = 0.6375(0.0078 + 0 + 0.0062 + 0.0124) \\ & + 0.2625(0 + 0 + 0 + 0.0124) \\ & + 0.225(0 + 0.0062 + 0 + 0.0124) \\ & = 0.0243 \end{aligned}$$

or

$$\begin{aligned} E^A = \; & \gamma_3 S_{\{3\}} + \gamma_{49} S_{\{49\}} + \gamma_{52} S_{\{52\}} \\ & = 0.6375 \times 0.0264 + 0.2625 \times 0.0124 + 0.225 \times 0.0186 \\ & = 0.0243. \end{aligned}$$

and

$$\begin{aligned} E^B = \; & \gamma_3\Big(P_{\{3\}} + P_{\{3,52\}}\Big) \\ & + \gamma_{52}\Big(P_{\{52\}} + P_{\{3,52\}}\Big) \\ & = 0.6375(0.0078 + 0.0062) \\ & + 0.225(0 + 0.0062) \\ & = 0.0103 \end{aligned}$$

Notice that the computation of E^B is based on the probabilities $P_{\{3\}}, P_{\{52\}}$ and $P_{\{3,52\}}$.

Therefore

$$\rho = E^B/E^A = 0.0103/0.0243 \approx 0.4239$$

In this case, the profit is underestimated by about 58%.

V. Case Study

The dataset in Ertek, Demiriz and Cakmak [23], [24] is a case study of the global hot drinks industry which grew 4.6% to reach a value of USD$ 94 billion USD in 2012 [27]. The dataset contains the personal attributes and purchasing preferences of 644 Turkish customers if given a purchase budget of 15 TL (approximately USD$ 8 at the time of the survey). In this study, we treat the purchasing preferences information as purchase transactions.

After running the dataset through AssocMiner [25], we found 52 unique items and a total of 659 transactions or itemsets with nonzero support values. The largest support value for an item was found to be at 0.118 and the smallest at 0.005. The profits for each transaction were then computed via a spreadsheet using real prices of the items (in TL, as of 2009) and assuming a profit margin of 15%, a value that was derived by averaging the profit margins of Starbucks globally [28] and in China [29]. The dataset was then evaluated in 60 scenarios of varying $minsup$ values ($0.0020 \leq minsup \leq 0.1200$) and the computational results plotted in Figure 5.

In Figure 5, we can observe that the decrease of ρ is monotonic and non-linear. It also can be understood that increasing the $minsup$ thresholds would increase the underestimation of

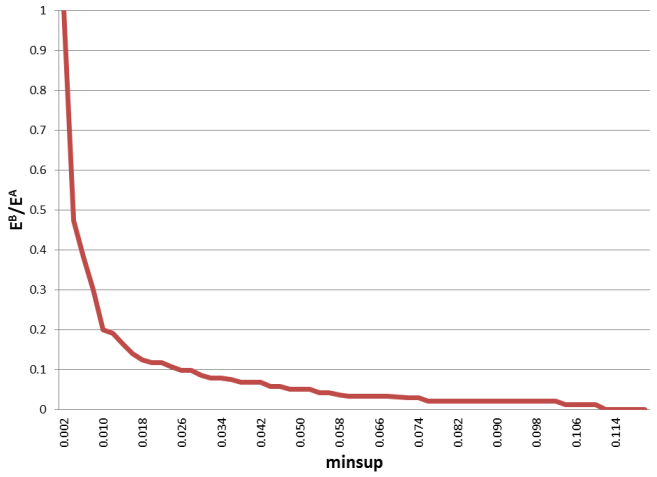

Fig. 5. $\rho = E^B/E^A$ vs. varying values of $minsup$, for profit margin of 15% in Ertek, Demiriz and Cakmak [23].

profit function. Another important observation is also that the profit ratio decreases significantly even at very low $minsup$ thresholds, suggesting that it is necessary to set the $minsup$ to very low values to be able to obtain an accurate estimation for the expected profit.

The transaction data, the assumed item prices and profit margin, and the results of the analysis are shared online as a reference dataset for future research in this field [24].

VI. Conclusions

In this paper, we looked into the issue of profit estimation error due to information loss in recommender systems based on association mining. We investigated this by analyzing two cases: Case A where $minsup = 0$ (complete information) and Case B where $minsup > 0$ (incomplete information). The developed concepts and formulas are illustrated with a numerical example, and applied to real world data. We find that there is a monotonic, non-linear relationship between the expected profits (as a function of information loss) and support threshold levels.

As a final word, we would like to once again draw attention to the importance of accurate estimation of the profit function in recommender systems. This issue is highly relevant and important in recommender system projects, which are increasing in number and scope in today's data and information driven world.

Acknowledgment

The authors thank Ayhan Demiriz and Fatih Çakmakçı for their contribution in an earlier paper, as well as the contributors to that paper for collecting the data used in the case study.

References

[1] S.-H. Liao, P.-H. Chu and P.-Y Hsiao, "Data mining techniques and applications: A decade review from 2000 to 2011", Expert Systems with Applications, Volume 39, Issue 12, Pages 11303-11311, 2012.

[2] A.K. Choudhary, J.A. Harding and M.K. Tiwari, "Data mining in manufacturing: a review based on the kind of knowledge", Journal of Intelligent Manufacturing, 20(5), 501-521, 2009.

[3] J. Rong, H.Q. Vu, R. Law, G. Li, "A behavioral analysis of web sharers and browsers in Hong Kong using targeted association rule mining", Tourism Management, Volume 33, Issue 4, Pages 731-740, 2012.

[4] C.K.H. Lee, K.L. Choy, G.T.S. Ho, K.S. Chin, K.M.Y. Law and Y.K. Tse, "A hybrid OLAP association rule mining based quality management system for extracting defect patterns in the garment industry", Expert Systems with Applications, Volume 40, Issue 7, Pages 2435-2446, 2013.

[5] M.-J. Shih, D.-R. Liu and M.-L. Hsu, "Discovering competitive intelligence by mining changes in patent trends", Expert Systems with Applications, Volume 37, Issue 4, Pages 2882-2890, 2010.

[6] R.-S. Wu, C.S. Ou, H.-Y. Lin, S.-I. Chang and D.C. Yen, "Using data mining technique to enhance tax evasion detection performance", Expert Systems with Applications, Volume 39, Issue 10, Pages 8769-8777, 2012.

[7] A.M. Cruz, "Evaluating record history of medical devices using association discovery and clustering techniques", Expert Systems with Applications, Volume 40, Issue 13, Pages 5292-5305, 2013.

[8] C.-W. Cheng, C.-C. Lin and S.-S. Leu, "Use of association rules to explore cause effect relationships in occupational accidents in the Taiwan construction industry", Safety Science, Volume 48, Issue 4, Pages 436-444, 2010.

[9] Y.S. Kim and B.-J. Yum, "Recommender system based on click stream data using association rule mining", Expert Systems with Applications, Volume 38, Issue 10, Pages 13320-13327, 2011.

[10] J.B. Schafer, J. Konstan and J. Riedi, "Recommender systems in e-commerce", in Proceedings of the 1st ACM conference on Electronic commerce (pp. 158-166). ACM, 1999.

[11] J.A. Konstan and J. Ried, "Deconstructing recommender systems: Recommended for you", IEEE Spectrum, 49-56, 2012.

[12] B. Shim, K. Choi and Y. Suh, "CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns", Expert Systems with Applications, Volume 39, Issue 9, Pages 7736-7742, 2012.

[13] I. Cil, "Consumption universes based supermarket layout through association rule mining and multidimensional scaling", Expert Systems with Applications, Volume 39, Issue 10, Pages 8611-8625, 2012.

[14] A. Demiriz, G. Ertek, T. Atan and U. Kula, "Re-mining item associations: Methodology and a case study in apparel retailing", Decision Support Systems, 52(1), Pages 284-293, 2011.

[15] S. Sangelkar, N. Cowen and D. McAdams, "User activity-product function association based design rules for universal products", Design Studies, 33(1), Pages 85-110, 2012.

[16] D.H. Park, H.K. Kim, I.Y. Choi and J.K. Kim, "A literature review and classification of recommender systems research", Expert Systems with Applications, Volume 39, Issue 11, Pages 10059-10072, 2012.

[17] E.W.T. Ngai, L. Xi and D.C.K. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications, Volume 36, Issue 2, Part 2, Pages 2592-2602, 2009.

[18] X. Amatriain, A. Jaimes, N. Oliver and J.M. Pujol, "Data mining methods for recommender systems", in Recommender Systems Handbook (pp. 39-71). Springer US. 2011.

[19] W. Lin, S.A. Alvarez amd C. Ruiz, "Efficient adaptive-support association rule mining for recommender systems", Data Mining and Knowledge Discovery, 6(1), Pages 83-105, 2012.

[20] J.L. Herlocker, J.A. Konstan, L.G. Terveen and J.T. Riedl, "Evaluating collaborative filtering recommender systems", ACM Transactions on Information Systems (TOIS), 22(1), Pages 5-53, 2004.

[21] P. Pu, L. Chen and R. Hu, "Evaluating recommender systems from the users perspective: survey of the state of the art", User Modeling and User-Adapted Interaction, 22(4-5), Pages 317-355, 2012.

[22] G. Ertek and A. Demiriz, "A framework for visualizing association mining results", in Computer and Information Sciences – ISCIS 2006 (pp. 593-602). Springer Berlin Heidelberg. 2006.

[23] G. Ertek, A. Demiriz and F. Cakmak, "Linking behavioral patterns to personal attributes through data re-mining", in Behavior Computing (pp. 197-214). Springer London. 2012.

[24] http://ertekprojects.com/ftp/supp/13.xlsx. Accessed on June 20, 2014.

[25] E.N. Cinicioglu, G. Ertek, D. Demirer and H.E Yoruk, "A framework for automated association mining over multiple databases", in Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on (pp. 79-85). IEEE. 2011.

[26] Oestreicher-Singer, G., and Sundararajan, A., "Recommendation Networks and the Long Tail of Electronic Commerce", in MIS Quarterly, 36(1). 2012

[27] Marketline Industry Profile: Global Hot Drinks, November 2013. Reference Code: 0199-0803. Accessed on April 10, 2014.

[28] Starbucks Profit Margin. YCharts. https://ycharts.com/companies/SBUX/profit_margin. Accessed on April 10, 2014.

[29] Chart: The Extra-Caffeinated Cost of a Starbucks Latte in China. The Wall Street Journal, ChinaRealtime. http://blogs.wsj.com/chinarealtime/2013/09/04/chart-the-price-of-a-grande-latte-in-china/. Accessed on April 10, 2014.