IEMS308

HW2 - Association Rules

Yujia Zhai

Executive Summary

Given the point-of-sale data provided by the retailer, information about how different SKUs are associated with each other is available. Since the retailer wants to improve their entire chain structure to raise the sales volume, they're interested in how they should rearrange the floors. Namely, limited to the maximum number of moves to make, they want to know what SKUs are highly associated with each other, and furthermore, to position highly associated ones as close as possible. Therefore, the purpose is to identify most important SKUs among about totally 1.56 million ones, based on over 5.5 million order records.

Association rule is a data mining technique used to discover co-occurrence relationships among individuals. It seeks to analyze conditional probability of an individual, given presence of certain other individuals, and then use the "lift" criterion to measure how these individuals are correlated. In applications, it's often used in purchase pattern analysis and recommendation system. For example, based on discovered association between items, purchase of one item would suggest high probability of buying the other one, and thus, recommending the other one might lead to higher profit for the retailer.

In this case, the retailer would like to apply association rule technique to find out important SKUs so that rearrangement of those SKUs would lead to higher sales volume. Because of limited manpower, only 20 moves at most can be conducted. Therefore, rearranging all the SKUs would be impossible. Therefore, the retailer would like to find out the top 100 SKUs from the top 50 association rules with highest lift. Then, these 100 SKUs are the most important ones that the retailer wants to focus on.

Problem Statement

The retailer wants to re-position their SKUs to potentially increase sales volume. However, the limitation of manpower makes it impossible to make moves more than 20 times. Therefore, the retailer would like to find out the "most important" 100 SKUs that are worth being moved. The criterion for "most important" is that these SKUs should be highly associated with others (i.e., are from association rules with highest lift), so positioning these SKUs close to each other would produce potentially largest sales volume.

Assumptions

- The data is comprehensive and general. Namely, it fully represents the purchase behavior, and the behavior is not biased.

- The randomly selected subset of data also fully represents the whole data.
- Only simple order information is considered in this case. Other behaviors, such as returning or replacing items, are not considered.

Methodology

The methodology can be summarized and categorized as:

(1) The dataset is cleaned and transformed to get order information. Specifically, if transaction numbers, transaction dates, and store numbers are the same, then it's considered to be a single order. Among these orders, the ones with at least 3 items purchased are selected. This step includes the order information that is needed for association rule analysis.

(2) A subset of the order information data is selected. In this case, 200,000 records are randomly selected from all the order data. A simple data exploration is conducted for this subset.

(3) Association rule analysis is conducted on the 200,000-record subset data. A minimum support of 0.001 is set, because supports that are too small might make it unworthy to emphasize too much on. Based on the minimum support, the top 50 association rules with largest lift are selected. Namely, the top 100 SKUs are selected from the subset.

(4) Visualization of association rule analysis conducted. It presents the plot of confidence versus support. Thus, the results are more straightforward to observe.

Data Exploration

Among the whole dataset, totally 5.5 million orders are derived, with 1.56 million SKUs. After the subset is randomly selected, 200,000 orders are kept with about 0.96 million SKUs.

Analysis

The top 50 association rules are listed in Table 1 below.

```
result.head(50) # check top 50 records order by lift value
```

	item_A	item_B	countAB	supportAB	countA	supportA	countB	supportB	confidenceAtoB	confidenceBtoA	lift
256	6032521	6072521	760	0.022757	264	0.007905	330	0.009881	2.878788	2.303030	291.333333
258	6062521	6072521	807	0.024165	295	0.008833	330	0.009881	2.735593	2.445455	276.842034
55	1453503	1563503	380	0.011379	201	0.006019	237	0.007097	1.890547	1.603376	266.399647
255	6032521	6062521	604	0.018086	264	0.007905	295	0.008833	2.287879	2.047458	259.003390
300	6742521	6752521	411	0.012307	214	0.006408	248	0.007426	1.920561	1.657258	258.625188
281	6470353	6490353	449	0.013445	221	0.006618	267	0.007995	2.031674	1.681648	254.119071
292	6642521	6742521	348	0.010420	214	0.006408	214	0.006408	1.626168	1.626168	253.773430
306	6972521	7232521	367	0.010989	240	0.007186	205	0.006138	1.529167	1.790244	249.112439
200	4648362	6208362	348	0.010420	201	0.006019	237	0.007097	1.731343	1.468354	243.965993
265	6300353	6340353	333	0.009971	208	0.006228	226	0.006767	1.600962	1.473451	236.573945
293	6642521	6752521	364	0.010900	214	0.006408	248	0.007426	1.700935	1.467742	229.050045
185	4318362	4648362	329	0.009851	281	0.008414	201	0.006019	1.170819	1.636816	194.530621
111	3348362	7228362	649	0.019433	510	0.015271	221	0.006618	1.272549	2.936652	192.298855
206	4898362	6738362	302	0.009043	224	0.006707	241	0.007216	1.348214	1.253112	186.825578
188	4318362	5938362	444	0.013295	281	0.008414	339	0.010151	1.580071	1.309735	155.657985
228	5278362	7228362	696	0.020841	730	0.021859	221	0.006618	0.953425	3.149321	144.074977
286	6560353	6580353	392	0.011738	212	0.006348	442	0.013235	1.849057	0.886878	139.708358
263	6208362	6998362	194	0.005809	237	0.007097	201	0.006019	0.818565	0.965174	136.004030
107	3348362	5278362	1338	0.040065	510	0.015271	730	0.021859	2.623529	1.832877	120.021080
289	6580353	8522644	270	0.008085	442	0.013235	201	0.006019	0.610860	1.343284	101.493888
191	4318362	7228362	188	0.005629	281	0.008414	221	0.006618	0.669039	0.850679	101.100594
249	5938362	6208362	229	0.006857	339	0.010151	237	0.007097	0.675516	0.966245	95.187932
250	5938362	6738362	217	0.006498	339	0.010151	241	0.007216	0.640118	0.900415	88.702824
252	5938362	7228362	196	0.005869	339	0.010151	221	0.006618	0.578171	0.886878	87.369239
104	3348362	4318362	373	0.011169	510	0.015271	281	0.008414	0.731373	1.327402	86.921415
186	4318362	4898362	160	0.004791	281	0.008414	224	0.006707	0.569395	0.714286	84.890696
262	6208362	6738362	141	0.004222	237	0.007097	241	0.007216	0.594937	0.585062	82.441935
211	5079905	5739904	249	0.007456	273	0.008175	374	0.011199	0.912088	0.665775	81.444085
202	4648362	6998362	98	0.002934	201	0.006019	201	0.006019	0.487562	0.487562	81.008094
273	6340353	8520723	108	0.003234	226	0.006767	203	0.006079	0.477876	0.532020	78.616505
197	4648362	4898362	104	0.003114	201	0.006019	224	0.006707	0.517413	0.464286	77.140725
331	8520723	8522644	93	0.002785	203	0.006079	201	0.006019	0.458128	0.462687	76.117638
239	5379905	6409904	148	0.004432	240	0.007186	275	0.008235	0.616667	0.538182	74.888000
105	3348362	4648362	227	0.006797	510	0.015271	201	0.006019	0.445098	1.129353	73.952707
297	6738362	6998362	102	0.003054	241	0.007216	201	0.006019	0.423237	0.507463	70.320431
52	957662	5600671	135	0.004042	249	0.007456	269	0.008055	0.542169	0.501859	67.309536
189	4318362	6208362	125	0.003743	281	0.008414	237	0.007097	0.444840	0.527426	62.683004
233	5329905	7039904	146	0.004372	222	0.006648	358	0.010720	0.657658	0.407821	61.349539
251	5938362	6998362	116	0.003473	339	0.010151	201	0.006019	0.342183	0.577114	56.853432
108	3348362	5938362	281	0.008414	510	0.015271	339	0.010151	0.550980	0.828909	54.278883
218	5129905	6939904	170	0.005090	271	0.008115	417	0.012487	0.627306	0.407674	50.238658
190	4318362	6998362	84	0.002515	281	0.008414	201	0.006019	0.298932	0.417910	49.667393
199	4648362	5938362	101	0.003024	201	0.006019	339	0.010151	0.502488	0.297935	49.501695
205	4898362	6208362	78	0.002336	224	0.006707	237	0.007097	0.348214	0.329114	49.067360
236	5369905	6349904	179	0.005360	313	0.009372	390	0.011678	0.571885	0.458974	48.970951
223	5278362	5938362	355	0.010630	730	0.021859	339	0.010151	0.486301	1.047198	47.907140
295	6656135	7596135	281	0.008414	680	0.020362	308	0.009223	0.413235	0.912338	44.806513
215	5109905	5749904	236	0.007067	388	0.011618	459	0.013744	0.608247	0.514161	44.254969
141	3672270	7783784	61	0.001827	204	0.006109	233	0.006977	0.299020	0.261803	42.858622
298	6738362	7228362	64	0.001916	241	0.007216	221	0.006618	0.265560	0.289593	40.129626

Table 1. Top 50 association rules

Figure 1 below shows the plot of confidence(AtoB) versus support(AB) from Table 1. The ranking of association rules is presented in color from red to purple (corresponding to from high to low). We notice that the support maybe relatively low. This is because we obtain these data from a very large dataset, and number of SKUs is also large.

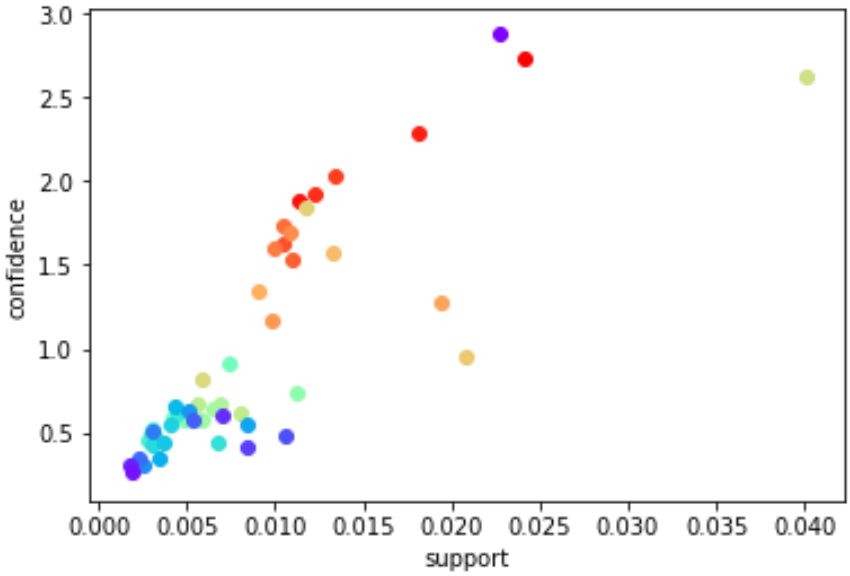

Figure 1. Confidence-Support plot

Based on this plot, we know that there are some rules that have relatively large lift and confidence, which means that they have a strong relationship. We notice that the supports are relatively low in these 50 rules. Here, the support is the joint probability of two SKUs, and thus it doesn't need to be high to make two SKUs highly associated. How two SKUs are associated with each other also depends on the marginal probability of each SKU. However, low support with high lift potentially means the marginal probabilities of the two SKUs are low, indicating that the sale volumes of the two SKUs are low. This makes repositioning of the SKUs not as worthy as it might seem to be. That's why a minimum support of 0.001 is set. This procedure filters the SKUs with high association but low sale volumes.

Conclusion

The top 50 association rules with totally 100 SKUs are obtained. The retailer can determine the 20 moves based on the top 50 association rules, since the association rule analysis provides "important" SKUs with high association with each other, so the retailer knows positioning which SKUs together could lead to the highest efficiency. These 100 SKUs will be the top consideration of the retailer when making the rearrangement plan. By moving top SKUs close to the ones highly associated with them, the efficiency can be greatly improved, and operation cost will also be reduced.

Next steps

Based on the 50-selected association rule, the retailer needs to determine which SKUs to move to position highly associated SKUs close to each other. Although the decision of repositioning depends on association rule analysis result, it also has much to do with current positions of SKUs. Therefore, to reposition SKUs, there's also one more step of optimize the repositioning process.

If more detailed information is available, deeper behaviors can be further analyzed. For example, returning and replacing behaviors might sometimes be highly associated with purchase of certain items, and information about how they are associated can be valuable.

Also, after determining these top 100 SKUs, rearrangement should be conducted. After rearrangement, feedback information, such as how sales volume is increased, needs to be collected and analyzed. This information should be kept being tracked and used to validate the effectiveness of rearrangement.