The recommendation system of H&M products

Based on purchase history

Xuan zhang
the college of sciences
northeastern university
Boston US
zhang.xuan8@northeastern.edu

ABSTRACT

This recommender system is based on content. In this article, I make an Item file and a User file to do cosine similarity. In general, the recommender system will calculate the cosine similarity between the User file and Item file. But, to H&M data, there are lots of products that are in this data. Thus, I use t-sne and a hierarchical cluster to observe the data and use k-meaning to do good classification. This way helps the system to find recommended products quicker. Finally, each customer can get 10 recommended products. And the average of RMSE from 100 customers is not over 0.12. RMSE is very low.

.

CCS CONCEPTS

• Applied computing • Electronic commerce • Online shopping

KEYWORDS

User file, Item file, hierarchical cluster, T-sne, k-mean cluster, the cosine similarity, RMSE, Precision and Accuracy

ACM Reference format:

1 Introduce the data

H&M is a big company, and it provides a huge dataset that includes products' pictures, customers' purchase history, product detailed metadata, etc. In this project, I only use two CVS files data.

There are two CSV files Transactions_train.csv containing the purchases of each customer to date.

	t_dat	customer_id	article_id	price	sales_channel_id
0	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	663713001	0.050831	2
1	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	541518023	0.030492	2
2	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	505221004	0.015237	2
3	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687003	0.016932	2
4	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687004	0.016932	2

Articles.csv contains detailed metadata for each article_id available. In the articles.csv, there are nine numerical attributes that were selected to do the K-mean cluster and calculate cosine similarity. There are product_type_no, graphical _ appearance _ no, colour _group _ code,perceived_colour_value_id,perceived_colour_master_id,department_no,index_group_no,section_no,garment_group_no. These nine attributes will distinct products of H&M from different aspects which include colour, type, graphs and garment etc.

	article_id	product_code	product_type_no	graphical_appearance_no	colour_group_code	perceived_colour_value_id	perceived_colour_master_id	department_
0	108775015	108775	253	1010016	9	4	5	1
1	108775044	108775	253	1010016	10	3	9	1
2	108775051	108775	253	1010017	11	1	9	1
3	110065001	110065	306	1010016	9	4	5	1
4	110065002	110065	306	1010016	10	3	9	1
5	110065011	110065	306	1010016	12	1	11	1
6	111565001	111565	304	1010016	9	4	5	3
7	111565003	111565	302	1010016	13	2	11	3
8	111586001	111586	273	1010016	9	4	5	3

2 Methods and algorithm

2.1 Introduction about assumption and definition

In this project, I assume that some potential customers in the same area (like Boston) once bought H&M group's products. They may browse H&M group's website and APP. When they browse its website and APP, we will provide some recommendations about H&M group's products. These products are being stored in warehouses which are around Boston. All products are seasonal clothes. Thus, I just use a part of customers' information and a part of the products' information from this data which provides all customers' purchase history and product information.

Making a good Item file and User file is very important for us to do a good recommendation to customers. I wash the data and analysis its attributes of the data. For each attribute of customers, I will do a weight depending on purchase history.

Then, to classification, this step is to reduce space and time when we recommend products to customers. K-means can get a good classification result. But we still need to watch some results from T-sne and hierarchical cluster for helping us to do a good K-mean.

2.2 Item file and User file

Item file: using articles.csv nine attributes to do vectors of 0's and 1's, where a 1 represents the occurrence of one value of the attribute.

product_type_no	graphical_appearance_no	colour_group_code	perceived_colour_value_id	perceived_colour_master_id	department_no	index_group_no	section_
[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0]	[1, 0, 0 0, 0, 0 0, 0, 0 0, 0, 0
[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0]	[1, 0, 0 0, 0, 0 0, 0, 0 0, 0, 0
[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 1, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0]	[1, 0, 0 0, 0, 0 0, 0, 0 0, 0, 0
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0]	[0, 1, 0 0, 0, 0 0, 0, 0 0, 0, 0
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0]	[0, 1, 0 0, 0, 0 0, 0, 0 0, 0, 0

User file: Use transactions_train.csv to do a user matrix of 0 to 1, and will add 1 when one value appears, then divide the number of this customer's purchase.

product_type_no	graphical_appearance_no	colour_group_code	perceived_colour_value_id	perceived_colour_master_id	department_no	index_group_no	section_
[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 1, 0, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[2, 0, 0, 0, 0]	[0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
[0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 2, 0, 0, 0, 0, ...	[2, 0, 0, 2, 0, 1, 0, 0]	[0, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[3, 0, 0, 0, 2]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
[0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, ...	[4, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...	[2, 0, 2, 1, 0, 0, 0, 0]	[1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[5, 0, 0, 0, 0]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...	[1, 0, 0, 1, 0, 0, 0, 0]	[0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 2]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
[8, 0, 0, 0, 0, 4, 2, 0, 0, 0, 0, 0, 0, 0, 0, ...	[13, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0,...	[11, 9, 0, 0, 0, 2, 0, 4, 0, 0, 0, 2, 0, 0, 0,...	[13, 9, 6, 0, 2, 0, 0, 0]	[11, 9, 0, 6, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0,...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[12, 0, 0, 0, 18]	[0, 0, 0, 0, 0, 0, 0, 3, 0, 2, 0, 0,
...	...	...	...	...	...	...	
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 1, 0, 0, 0, 0, 0, 0]	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[0, 0, 0, 1, 0, 0, 0, 0]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	[1, 0, 0, 0, 0]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

The number of the customer purchase:

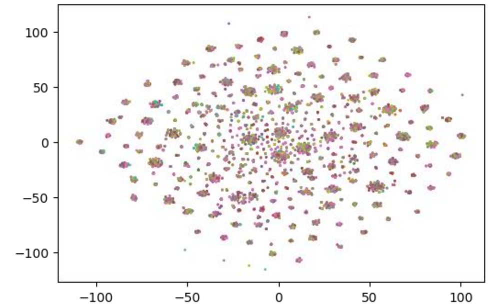
```
[79]: sta
[79]: {'000058a12d5b43e67d225668fa1f8d618c13dc232df0cad8ffe7ad4a1091e318' : 2,
 '00007d2de826758b65a93dd24ce629ed66842531df6699338c5570910a014cc2' : 5,
 '00083cda041544b2fbb0e0d2905ad17da7cf1007526fb4c73235dccbbc132280' : 5,
 '0008968c0d451dbc5a9968da03196fe20051965edde7413775c4eb3be9abe9c2' : 2,
 '000aa7f0dc06cd7174389e76c9e132a67860c5f65f970699daccc14425ac31a8' : 30,
 '001127bffdda108579e6cb16080440e89bf1250a776c6e55f56e35e9ee029a8d' : 1,
 '001ea4e9c54f7e9c88811260d954edc059d596147e1cf8adc73323aebf571fd8' : 2,
 '001fd23db1109a94bba1319bb73df0b479059027c182da490e1161b34cd3af61' : 2,
 '0021da829b898f82269fc51feded4eac2129058ee95bd75bb1591e2eb14ecc79' : 4,
 '00228762ecff5b8d1ea6a2e52b96dafa198febddbc3bf350eb611f28e61ea6ce' : 1,
 '002b3c0a44a22c45a8d62ea9d2b88d1a89e335f8b8400335a85c3b5845794cb9' : 3,
 '003ac6eff3ca20cbe4e14a106a5e28b7551616005de8f9dd0f9455e5543b31ac' : 4,
 '00401a367c5ac085cb9d4b77c56f3edcabf25153615db982fcc7991d182e10a9' : 11,
 '00402f4463c8dc1b3ee54abfdea280e96cd87320449eca8953eb06769a5c20d4' : 2,
 '0045c79125b4dc958579f902b49eacd8598f9eeaa1220506d1cae1275f122084' : 4,
 '004888668b9458668aa3234d17337671673599f7f84d36f77fd366b5139058c5' : 1,
 '0051a1d00ec9a4ff1a2f9aa32a4bb054cc9ba08e80be590aec39391ff3dc56e8' : 1,
 '005c9fb2ba6c49b2098a662f64a9124ef95cbec5fcf4ebdb4dcbaaf83f979c51' : 4,
```

The products' ID for each customer:

```
30]: goodid
30]: {'000058a12d5b43e67d225668fa1f8d618c13dc232df0cad8ffe7ad4a1091e318' : [663713001,
  541518023],
 '00007d2de826758b65a93dd24ce629ed66842531df6699338c5570910a014cc2' : [505221004,
  685687003,
  685687004,
  685687001,
  505221001],
 '00083cda041544b2fbb0e0d2905ad17da7cf1007526fb4c73235dccbbc132280' : [688873012,
  501323011,
  598859003,
  688873020,
  688873011],
 '0008968c0d451dbc5a9968da03196fe20051965edde7413775c4eb3be9abe9c2' : [531310002,
  529841001],
 '000aa7f0dc06cd7174389e76c9e132a67860c5f65f970699daccc14425ac31a8' : [501820043,
  501820043,
  674681001,
  671505001,
  671505001,
```

2.3 Hierarchical cluster, T-sne and K-mean cluster

T-sne is one good way to observe data. It will reduce the dimension of data for us to observe. This is a picture from T-sne.

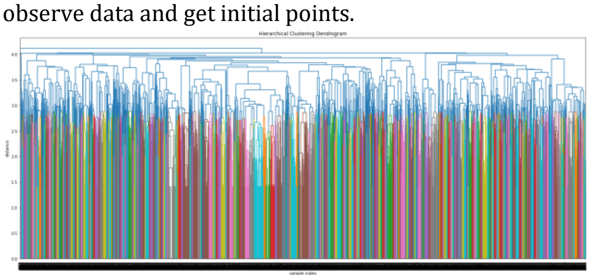


We can find lots of clusters, and the shape of the data is a parallelogram.

T-sne formula: $C = \sum_i KL\left(P_i \| Q_i\right)$

$$\frac{\partial C}{\partial y_i} = 2\sum_j \left(p_{ji} - q_{ji}\right)\left(y_i - y_j\right)\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}$$

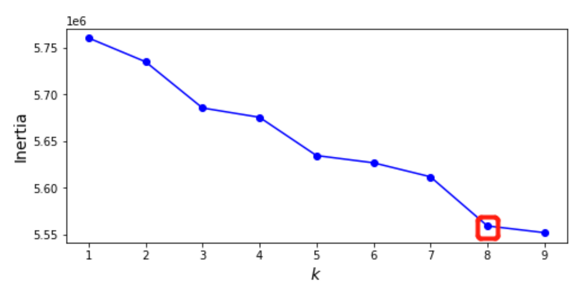
Then I used a hierarchical cluster that uses cosine distance to observe data and get initial points.



The hierarchical cluster formula:

Finally, I use K-mean's Inertia to find the elbow of the data. The elbow is 8. Then using K-mean do classification the data.

Formula: $J\left(c, \mu\right) = \sum_{i=1}^{m} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$



Final group result:

```
In [77]: p
Out[77]: array([1, 3, 0, ..., 2, 3, 3])
```

2.4 Cosine similarity

Formula: $\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$

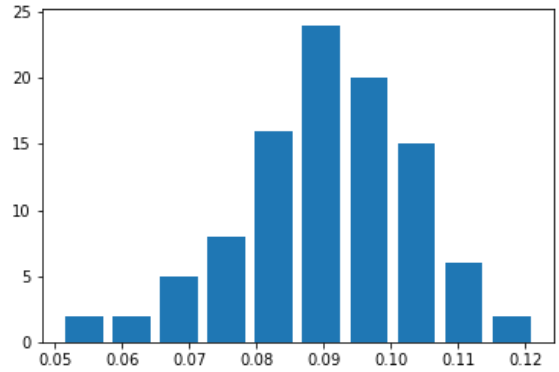
I use K-mean to group one customer from the User file. Then Use the customer data to calculate cosine similarity with the Item file and find ten similar items and return these items' id.

2.5 analysis result about RMSE

To result in the similarity of the User file and Item file, we use RMSE to calculate the difference. I select 100 customers; each customer is recommended ten products. Then I calculate the average of RMSE by the value of the customer and the value of the ten products. I choose the mean of each customer's RMSE.

Formula: $RMES\_mean\_eachcustomer = \sqrt{\sum_{i=1}^{n} \frac{\left(\hat{y}_i - y_i\right)^2}{100n}}$

I use python to do a histogram to show the distribution of the value of RMES_mean_eachcustomer.



From this table, we can find lots of the average of RMSE is about 0.09. It means the difference between attributes of recommendations and attributes of customers is very low. I get a good result from validation.

3 Procedure

1. Observe data and select useful attributes from the data
2. Making an Item file and a User file
3. Using T-sne and the hierarchical cluster to observe the distribution of the data, then using the K-mean to do classification.
   To a special customer, I select this customer's data from the User file and classify this customer into one special group, then calculate cosine similarity between items in this group and this customer and recommend products that are the biggest similarity ten.
4. Calculate RMES

4 Further research

We can try to use collaborative filtering in this project. The collaborative filtering will be based on the User file and Item file to fill out empty of these two files. I think it will make this model of this project better. In addition, there are some parts of the H&m data I did not use in this project. It is possible to extract useful features from the products' pictures. It will help this model of this project to do a better classification.

ACKNOWLEDGMENTS

Thank Professor Eliassi-Rad Tina for her help of the validation of the model and guidance of the recommender system.

REFERENCES

[1] Jure Lekovec, Anand Rajaraman, Jeffrey D.Ullman,2014. Mining of Massive datasets. DOI: http://infolab.stanford.edu/~ullman/mmds/ch9.pdf .