Similarity Based Classification: Determining the Books that A Costumer Would Buy in The Future Using Past Purchases and Similarities Between Different Costumers and Books

A Data Mining Course Project

Faraz Moghimi :faraz.moghimi@umb.edu

Instructor : Dr. Josephine Namayanja

Abstract

Companies like that have huge data bases with a variety of features assigned to each of their user profile like types, ratings, views, etc. However, in its core there is much more sophisticated and complicated algorithm can be used to predict future behavior regarding product purchases by costumers if you happen to have a lot of costumers and a lot of products in your data. This algorithm would be called Similarity Based Classification. Similarity based classification estimate the class label of a test sample using the labeled training sample and the similarities between them. In this research project we will focus on developing and optimizing a machine learning similarity based classification algorithms to make predictions about the books that consumers will buy in the future. This would help ad design and suggestion, promotion campaigns and price settings for the book companies. We conduct variety of experiments and achieve a 68% percent accuracy with a decent validity. This means that we are able to predict future costumer behavior without any actual features using only the similarities between books and users to certain degree only 15,000 sample to train the model. This means that under different circumstances with more data and features these numbers could much more accurate.

Keywords—Similarity Based Classification, Consumer Behavior Prediction, SVM, Logistical Regression style, styling, insert (key words)

1. Introduction

Costumers do not always follow a completely well defined logic in their purchases. However, with enough data one can determine general patterns regarding consumer's purchase preferences. (Mitra, n.d.). We can see this phenomenon happening at its finest in the giant tech companies that have a lot user data like Amazon, Netflix, and Spotify. Considering that they have huge amount of data regarding different aspects of their users, it is understandable that they this good in predicting these patterns and making recommendations to their users. Hence, making more money and saving on their marketing budget by doing it smarter. Companies like that have huge data bases with a variety of features assigned to each of their user profile like types, ratings, views, etc. However, in its core, there is much more sophisticated and complicated algorithm can be used to predict future behavior regarding product purchases by costumers if you happen to have many costumers and many products in your data. This approach is called Similarity Based Classification.

Similarity based classification estimate the class label of a test sample using the labeled training sample and the similarities between them (Yihua Chen, 2009).. For example, in our case, this would mean be able to predict if a costumer would buy a particular product by assessing whether the most similar costumers to him have bought that product. In this research project we will focus on developing and optimizing a machine learning similarity based classification algorithms to make predictions about the books that consumers will buy in the future. This would help ad design and suggestion, promotion campaigns and price settings for the book companies.

2. Potential Contributions of the project

As for the potential contributions of this project, first I should mention that this particular task and data source was based on Kaggle competition. So, it would be fair to assume classifying this particular data set might not have much novelty per say. However, this study could contribute in providing a base algorithm

or approach that might be useful in various other applications as well as generating insights about the general classification approach that can be used in similarity based classifications.

- This project can help provide a framework that can be utilized for various similarity based classification endeavors. For instance, the most obvious application of this work could be in improving user experience and user advertisement. However, can also be potentially utilized in sales/revenue prediction, growth prediction, etc. Particularly, it would be really interesting for me personally to see the utilization of this framework in a financial markets setting where investors are the users and the books are the equities in a possible similarity based classification of investors and stocks.
- How do logistical regression and SVM's difference in theory plays out in their performance as a similarity based classifier? This work could potentially add to the numerous empirical comparisons between SVM and logistical regression
- How do different similarity patterns among data points such as book/user popularity, book/user jacardian similarity, etc. affect the performance of the model. Insights generated from this experiment could potentially help build more efficient future models.

3. Related work

This project relates to the general field of similarity based classification in terms of the theory and algorithms used. The field is not necessarily new, but it has been growing, modified and implemented in various ways. (Yihua Chen, 2009) provides a comprehensive review of the field before 2009. Since then there has various explanations in the field in terms both application and theory. For instance, (Gao, 2019) provides robust implementation of similarity based classification for symmetric positive defenite matrices.

Numerous other forms of similarity based classification have been used for image recognition as well. (Auch, 2020) provides more recent applications of this approach.

Another research area that this project could relate to is predicting user behavior. This field could range from predicting user energy consumption patterns (Mosavi, A review." (2019)), analyzing student performance (Rastrollo-Guerrero, 2020) , our case which is predicting the purchasing behavior of consumers in an online book store.

4. Method

Data

The data that we will be working on is consist 200,000 rows User Ids and Book ids listing the users and the books they have read. This data has been collected from Kaggle (Kaggle, n.d.) in a project competition created by UC San Diego (Sandiego, n.d.). A few rows of data have been shown in the table below .

Table 1. raw data structure

userID	bookID	rating
u79354815	b14275065	1
u56917948	b82152306	1
u97915914	b44882292	1
u49688858	b79927466	1
u08384938	b05683889	1

Dividing the data for feature creation, train and test

As we can see in the table the only data we have is of the costumers who have read some books with the class label 1 (rating). Thus, for us to be able to a 2 class classification prediction we need labels of

costumers who would not read a particular book. In other words, we would need data rows with 0 labels to do a machine learning classification with our classes being whether a costumer would buy a particular book or not. Hence, we have to create an artificial set of data with class label 0. To do that, we will generate 10,000 rows consisting of users and the books they have not read with class labels 0 by selecting random pairings and running them against our main data base to see if they are in it or not. If the selected pair is not in our main data base, that means that we have a pair in which the user have not read the book. For our other class which will must consist of rows with 1 classes we would select the last 10,000 rows of data which has users and the books they have read. Now we have 20,000 data points with 50/50 distribution of our classes. This data would be divided to create our train and test samples.

As for the other 190,000 data points, since we do not have any actual features on this data set on the first look, we will use them to create our similarity and popularity features.

Creating 8 similarity and popularity based features for the data

In this segment we will work on creating 8 features for our data. A brief description on the methods and reasons for our feature determination is given for each feature.

(1) In this feature, we would count the number of reads each book has as a feature. This would be an indicator of the popularity of the book

(2) In this feature, we count the number total reads each user. This would illustrate the book savviness of each user in general

(3) In the third feature, we ranked those the books from most popular to least popular and added book popularity rank as a feature

(4) In the forth feature, we ranked the users based on book savviness and added that rank as a feature.

(5) In features 5 and 6 we used Jaccard similarity index (Statistic, n.d.) find similarity between users and books. Jaccard similarity index is basically:

$$J(X,Y) = |X \cap Y| \ / \ |X \cup Y|$$

In our feature 5, we will use this index to add the maximum Jaccard similarity between a query of users and other users to our feature list. As for converting this to actual useful features in the final train sample, we look at the user and book, and find what maximum Jaccard similarity for a persons other books to the current row book in question.

(6) In our feature 6, We will use Jaccard similarity index to determine the maximum similarity between a query of books and other books. We will convert this to a useful feature like the previous segment. In other words, we assign the maximum similarity of a book's other user with the examined user as a feature to our final table.

(7) As for the 7th feature, first we would count the number of times two books have been read at the same time by users for all of the pairing and then we would divide that number with total number 2 pairings which would provide on us with basically a number with a probability nature that gives us a perspective into the likelihood these books being read together. In other words, similarities between two book pairings

(8) The logic of the feature 7 will applied to user pairings to determine the similarities between two users

Machine learning and experiment

In this project, we will use Support Vector Machines (SVM) and Logistical Regression as our two training models and fine tune each separately. Then we evaluate these models based on prediction accuracy using our test samples with 2,4,6,8 parings of our features to see the contributions they are making. We would also add and remove Standard Scaler to see its contributions each model's performance.

5. Results

The results from the discussed experiments can be seen on :

Table 2. Results

Features	Details	SVM + Standard Scalar	SVM	Logistic Regression + Standard Scalar	Logistic Regression
Only with bookcount and Usercount	Simplest Model	0.6442	0.4944	0.6404	0.4944
+ ranking	We added same features, in different format of ranking	0.6452	0.4944	0.6478	0.5994
+ Jaccard Similarity	We added maximum Jaccard similarity between the query user and other users. And same things for books.	0.6788	0.4944	0.6766	0.5994
All parameters(+pair probability)	We finally added the probability of book pairs pattern and also for users	0.681	0.4944	0.681	0.669

6. Discussion of results, conclusion and future work

We can see that we can achieve 68% percent accuracy with a decent validity. This means that we are able to predict future costumer behavior without any actual features using only the similarities between books

and users to certain degree only 15,000 sample to train the model. This means that under different circumstances with more data and features these numbers could much more accurate.

Also, we can see in the results that SVM fails to work without the standard scaler. This is mainly due to the fact that SVM tries to maximize the distance between separating plane and support vectors and if one feature is consist of very large values, which in our case is the our first 4 features, those features will dominate the data set when calculating distances and consequently bias the model. On the other hand we could see that standard scaling barely effects the logistical regression results. The main reason for this phenomenon is that dependent variables in logistical regression are not measure in an interval or ratio scale, meaning the model would not be biased towards larger value features.

Another observation that we can make is that our ranking features (features 3 and 4) are hardly influencing the performance of the model. The reason is probably that we already have those features in a different format as our book popularity and user book savviness in our features 1 and 2.

All in all, it can be seen that using similarity classification methods can be quite useful in predicting future consumer behavior. Our artificial test sample created here consisting of the 0 value labels may not be the most accurate approach given that those users might in reality buys those book in the future or buy it from other sellers. However, this project gives us a glimpse of the possibilities in consumer behavior predations and pattern recognition by utilizing similarity based classification.

For future work, we could add more insightful features to our data, to combine that with our similarity based model. This features could include book genre, rating, author, sales, size, price, etc. We could also increase our training sample to increase our accuracy. Big tech company use millions of data points for this sort of estimations. Also, taking into account some simulation models, marketing spending, sales and profit could be interesting to indicate the actual fiscal contributions of this type model.

7. References

Auch, M. e. (2020). Similarity-based analyses on software applications: A systematic literature review." Journal of Systems and Software. *Journal of Systems and Software*.

Gao, Z. e. (2019). A robust distance measure for similarity-based classification on the SPD manifold. IEEE Transactions on Neural Networks and Learning Systems .

Kaggle. (n.d.). Retrieved from https://inclass.kaggle.com/c/cse158258-fa19-read-prediction/data

Mosavi, A. a. (A review." (2019)). Energy consumption prediction using machine learning:.

Rastrollo-Guerrero, J. L.-P.-D. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Applied Sciences*.

Sandiego, U. (n.d.). *UC San Diego*. Retrieved from https://ucsd.edu/

Yihua Chen, E. K. (2009). Similarity-based Classification: Concepts and Algorithms. *Journal of Machine Learning Research 10*.