Market Basket Analysis using Machine Learning and Data Mining

Bharadwaj Mudumba

COMPSC 445 – Machine Learning in Applied Data Science..

Abstract

Market Basket Analysis is a data mining technique that has evolved into one of the most popular and widely applied retail models. It is used in a variety of industries to comprehend trends, purchasing patterns, and their relationships. This project aims to use a deep learning algorithm and data mining algorithm, to recognize the products in the customer's shopping cart and discover the correlated products between each purchased product in order to provide customer-specific recommendations, which I have achieved using Convolutional Neural Network and FP-growth algorithm.

1 Introduction

Customer service and meeting customer-specific requirements are critical in today's retail world, as businesses strive to address the needs of individual customer. Data mining techniques assess customer purchase records and discover frequent patterns, resulting in recommendations which is extended to a retail in-person business by integrating with the Object detection algorithm to recognize purchased products thereby using data mining to develop recommendations.

In the recent part, there were works on using data mining as a recommender system. In [1] Xavier Amatriain and Josep M. Pujo explained of the main Data Mining techniques used in the context of Recommender Systems in their handout and in [2] Borgelt, Christian. emphasised implementing FP growth as a data mining algorithm . Numerous works on object detection and image classification were published, one such kind is [3] by Zhao,Zhong-Qiu et al. in which they reviewed the concept of Object Detection with deep learning .

However, I Couldn't find an effort attempted to merge the concepts of object detection with data mining. In this project, I have integrated these two ideas and applied them to retail industry scenario.

The existing recommender system analyzes the customer's purchase history, which requires consumer to complete at least one transaction in order to map their purchase information with the frequent patterns to generate recommendations. In this project,rather than waiting for the consumer to complete his purchase , we use object detection to identify the products in the shopping cart and provides recommendations immediately. This accelerates the overall process and provides the customer with personalized recommendations.

1.1 My Contribution

- The FP-Growth algorithm is applied on customer purchase data set to discover the data associations between various products.
- The products in the shopping cart of the customer are recognised using convolutional neural network.
- Identified products are mapped with the associations generated.
- Thus, by incorporating data mining and object recognition, customer specific recommendations are produced.

2 Method

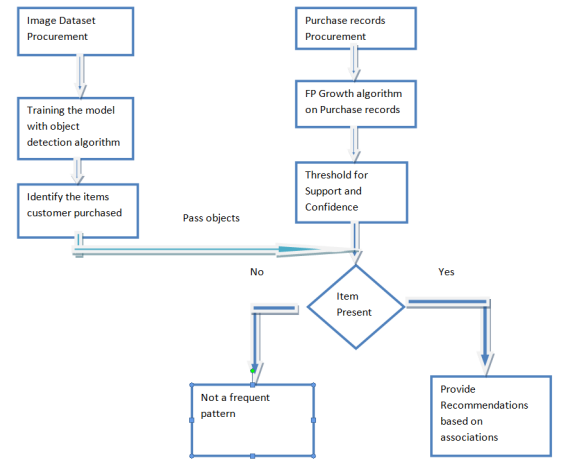

Fig. 1: Block Diagram of Proposed Model

2.1 Dataset Description

I have used the COCO (Common Object Context) dataset[4]. It is specifically designed for the research on image recognition. The dataset is specifically used for object detection. Object detection algorithms are developed considering this as a base dataset. The reason for choosing this dataset is it is being trained to recognise the image patterns i.e, it already has the knowledge of various kernels. It is trained with over 350,000 imgaes of over 80 different categories.

Apart from this, I have also trained the grocery dataset using the for which I used groceries-object-detection-dataset from github [5], which is an enhanced version of the Freiburg Groceries Dataset from the University of Frieberg[6]. The dataset covers a large spectrum of 5000 images covering 25 different categories of groceries, with at least 97 images per category. The reason for choosing this dataset because the enhanced version contains the annotations in PascalVOC format which scaled the images to 224*224*3. The dataset provides the bounding boxes for each image which describe the spatial location of an object. The bounding box is rectangular, which is determined by the X and Y coordinates of the upper-left corner of the rectangle and the such coordinates of the lower-right corner thus representing the (X,Y) axis coordinates of the bounding box center, and the width and height of the box.

2.2 Methodology

2.2.1 Machine Learning

I have used Convolutional Neural Network for object detection. CNN is most widely used for visual image recognition. It requires emphasis on features and locations

The architecture of CNN contains of raw visual image as input, which is sent to stack layer where image preprocessing activities like zooming, shrinking, rotating, contrasting are performed on the input image.

There will be combination of kernel layers which break the preprocessed image into smallest possible kernels

An image is identified based on following,

1. No.of pixels 2. No.of vertical bars in the image 3. No.of loops 4. Aspect ratio 5. No.of horizontal bars 6. No.of corners 7. No.of endpoints 8. No.of crosses 9. No.of junctions I have used yolov3 configurations and weights to train the model.The reason for using a pretrained model is that, it already has a knowledge of what kind of input will be passed, and if input image contains a set of kernels, it tries to relate to the categories it is trained with. Apart from this, using a pre-trained model for many image recognition tasks is beneficial for several reasons. The fact that using a pre-trained model training will be quick. Another positive is the accuracy, using a pre-trained model is significantly more accurate than using a custom-built convolutionary neural network.

2.2.2 Data Mining

Data mining is the process of finding anomalies, patterns and correlations within large data sets. In this project I have used frequent pattern mining also called association rule mining a kind of data mining technique used to find the association and correlation between the data in a dataset. As this project is focused on Market Basket Analysis, the dataset can be related to retail industry. For the purpose of mining, I have created a sample customer purchase history of 60 transactions and used FP growth algorithm to find the association between the item considering its efficiency, scalability and compactness. FP growth algorithm represents the dataset in the form of a tree called a frequent pattern tree It considers support and confidence in identifying the associations between products.

Support represents the popularity of that product of all the product transactions. Support of the product is calculated as the ratio of the number of transactions includes that product and the total number of transactions.

Confidence can be interpreted as the likelihood of purchasing both the products A and B. Confidence is calculated as the number of transactions that include both A and B divided by the number of transactions includes only product A.

The algorith first Scans the dataset once, find frequent single item pattern, Sort frequent items in frequency descending order, eliminates the items that don't meet the threshold. Scan DB again, construct FP-tree ,Construct the conditional FP tree in the sequence of reverse order of F - List - generate frequent item set and so on.

I have passed the objects that are recognized to the object detection algorithm once the FP growth algorithm has finished finding all the associations between the data with respect to minimum support and minimum confidence.

The model first checks to see if it is in the frequent pattern list. If it isn't on the list, the object isn't a frequent pattern because it didn't meet the minimum support. If the object is present in the list, it attempts to find all items associated with it satisfying the minimum level of confidence. The user is then presented with recommendations as a result of this.

2.3 Performance Evaluation

For FP growth algorithm I have used a support threshold of 0.04 and confidence of 0.5. The following is the associations between the products in the given dataset.

a	b	support	confidence
['skateboard', 'sandwich']	['tv']	0.05263	1.00000
['cell phone', 'donut']	['frisbee']	0.05263	1.00000
['tv', 'sandwich']	['skateboard']	0.05263	1.00000
['tv', 'skateboard']	['sandwich']	0.05263	1.00000
['tennis racket', 'refrigerator']	['apple']	0.05263	1.00000
['apple', 'tennis racket']	['refrigerator']	0.05263	1.00000
['apple', 'refrigerator']	['tennis racket']	0.05263	1.00000
['hair drier', 'baseball bat']	['donut']	0.05263	0.75000
['donut', 'hair drier']	['baseball bat']	0.05263	0.75000
['donut', 'frisbee']	['cell phone']	0.05263	0.75000
['donut', 'frisbee']	['sports ball']	0.05263	0.75000
['sports ball', 'frisbee']	['donut']	0.05263	0.75000
['cell phone', 'frisbee']	['donut']	0.05263	0.75000
['banana']	['donut']	0.07018	0.66667
['donut', 'baseball bat']	['hair drier']	0.05263	0.60000
['donut', 'sports ball']	['frisbee']	0.05263	0.60000
['orange']	['donut']	0.07018	0.57143
['vase']	['donut']	0.08772	0.55556
['baseball bat']	['donut']	0.08772	0.55556
['skateboard']	['donut']	0.08772	0.55556
['sandwich']	['donut']	0.08772	0.55556
['sports ball']	['donut']	0.08772	0.55556
['toothbrush']	['wine glass']	0.07018	0.50000
['wine glass']	['donut']	0.08772	0.50000
['spoon']	['donut']	0.07018	0.50000
['tie']	['donut']	0.07018	0.50000
['tie']	['sports ball']	0.07018	0.50000
['toothbrush']	['donut']	0.07018	0.50000
['mouse']	['donut']	0.07018	0.50000

Fig. 2: Suppor and Confidence

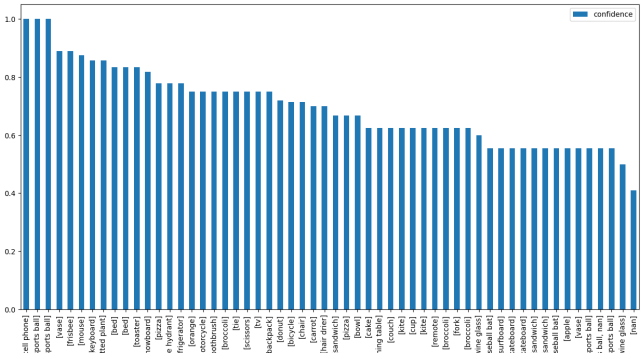

Fig. 3: Confidence

As the aim of the project is to combine Object detection and FP- 'growth, I could accomplish the task using coco dataset.I have additionally trained the grocery dataset. The following is the categories and its distribution of images

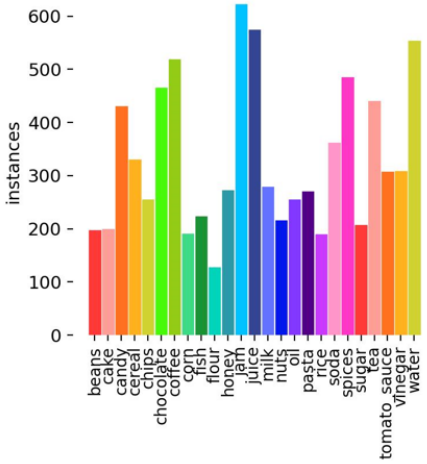

Fig. 4: Categorical Distribution

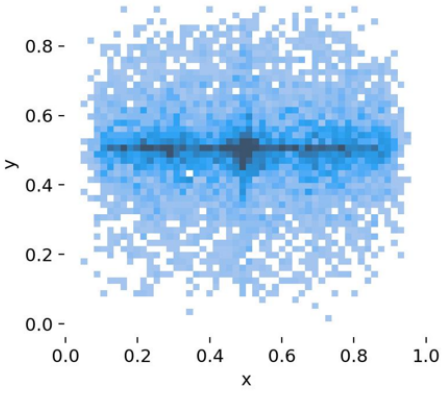

Fig. 5: Image Coordinates

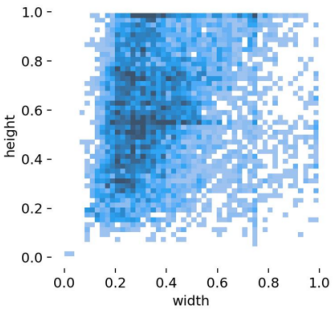

Fig. 6: Image Dimensions

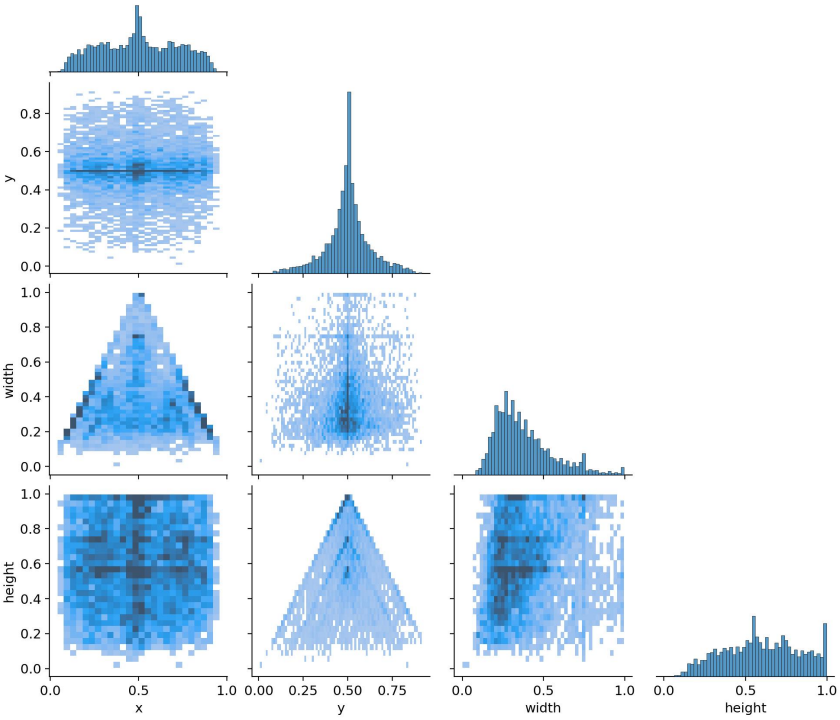

Fig. 7: Labels Correlogram

Correlogram is a group of 2d histograms showing each axis of your data against each other axis. The labels in your image are in x-y-w-h space.
This is the CNN model summary to represent that dataset.

```
Model summary: 283 layers, 7128270 parameters, 7128270 gradients, 16.6 GFLOPs

Transferred 355/361 items from yolov5s.pt
Scaled weight_decay = 0.0005
optimizer: SGD with parameter groups 59 weight (no decay), 62 weight, 62 bias
```

Fig. 8: CNN model summary

Usually a minimum of 100 epochs with a batch size of 32 gives a better accuracy to train a model of this size, . However, its requires a greater computational resources. When I tried I got the following error.
The best possible training I could do is to configure with a batch size of 16

```
    return F.conv2d(input, weight, bias, self.stride,
RuntimeError: [enforce fail at C:\actions-runner\_work\pytorch\pytorch\builder\windows\pytorch\c10\core\impl\alloc_cpu.cpp:81] data. DefaultCPUAllocator: not enough memory: you tried to
 allocate 26214400 bytes.
```

Fig. 9: Not Enough Memory

images and 5 epcohs. Each epoch took around 5 hrs to complete , and the 5the epoch ran for 2 hrs and ran out of computational resources and training terminated. Taking a total 22 hrs.

```
Epoch   gpu_mem       box       obj       cls    labels  img_size
  0/4        0G   0.06533   0.04014   0.07994        30       640: 100%|██████████| 217/217 [4:58:47<00:00, 82.62s/it]
           Class     Images     Labels          P          R     mAP@.5 mAP@.5:.95: 100%|██████████| 33/33 [11:52<00:00, 21.59s/it]
             all       1039       2362     0.0501      0.339     0.0528     0.0309

Epoch   gpu_mem       box       obj       cls    labels  img_size
  1/4        0G   0.03909   0.03088   0.07445        39       640: 100%|██████████| 217/217 [4:45:18<00:00, 78.88s/it]
           Class     Images     Labels          P          R     mAP@.5 mAP@.5:.95: 100%|██████████| 33/33 [10:27<00:00, 19.00s/it]
             all       1039       2362      0.122       0.29     0.0748     0.0435

Epoch   gpu_mem       box       obj       cls    labels  img_size
  2/4        0G    0.0356   0.02853   0.07227        29       640: 100%|██████████| 217/217 [4:50:48<00:00, 80.41s/it]
           Class     Images     Labels          P          R     mAP@.5 mAP@.5:.95: 100%|██████████| 33/33 [10:30<00:00, 19.09s/it]
             all       1039       2362     0.0744      0.433     0.0887     0.0505

Epoch   gpu_mem       box       obj       cls    labels  img_size
  3/4        0G   0.03044   0.02751   0.06835        72       640:  87%|████████▋ | 189/217 [4:22:32<38:53, 83.35s/it]
```

Fig. 10: Image Dimensions

3 Experiments and Test Results

I have passed an image of "donut" as test object and the object detection algorithm has identified it as donut. This result is used as input to frequent pattern mining algorithm to find the associations. If the object is not a frequent pattern or does not meet the minimum support or associations does not meet tyhe minium confidence then the following message will be displayed

However, In the given data souce, donut is a frequent pattern. The recommendations for donut with support threshold 0.05 and confidence threshold

Fig. 11: Test object is a Frequent Pattern

```
['donut']
It is not a frequent pattern or Try adjusting the the threshold for Support and Confidence
```

Fig. 12: Test object Not a Frequesnt Pattern

0.4. As donut has highest support in the dataset, it is well associated with other products

```
['donut']
You may also want ['frisbee']
You may also want ['baseball bat']
You may also want ['cell phone']
You may also want ['sports ball']
You may also want ['hair drier']
You may also want ['frisbee']
You may also want ['baseball bat', 'hair drier']
You may also want ['sports ball', 'frisbee']
You may also want ['frisbee', 'cell phone']
You may also want ['banana']
You may also want ['orange']
You may also want ['sports ball']
You may also want ['vase']
You may also want ['skateboard']
You may also want ['baseball bat']
You may also want ['sandwich']
You may also want ['cup']
You may also want ['toothbrush']
You may also want ['wine glass']
You may also want ['tie']
You may also want ['mouse']
You may also want ['spoon']
You may also want ['frisbee']
You may also want ['fire hydrant']
You may also want ['potted plant']
You may also want ['book']
You may also want ['hair drier']
```

Fig. 13: Test object recognised

4 Conclusion

I have successfully, combined the object detection algorithm with data mining to produce customer specific recommendations. This project can be used with live data considering the actual customer purchase data and providing the required support and confidence threshold. Even though this project is focused on the market basket analysis, this model can be further extended to other data mining techniques as well..

References

1. Amatriain, X., Pujol, J.M. (2015). Data Mining Methods for Recommender Systems. In: Ricci, F., Rokach, L., Shapira, B. (eds) Recommender Systems Handbook. Springer, Boston, MA.

2. Borgelt, Christian. (2010). An Implementation of the FP-growth Algorithm. Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations. 10.1145/1133905.1133907.

3. Zhao, Zhong-Qiu Zheng, Peng Xu, Shou-Tao Wu, Xindong. (2019). Object Detection With Deep Learning: A Review. IEEE Transactions on Neural Networks and Learning Systems. PP. 1-21. 10.1109/TNNLS.2018.2876865.

4. https://cocodataset.org/

5.https://github.com/aleksandar-aleksandrov/groceries-object-detection-dataset

6. @articlejund16groceries, author = Philipp Jund and Nichola Abdo and Andreas Eitel and Wolfram Burgard, title = The Freiburg Groceries Dataset, booktitle = CoRR, volume = abs/1611.05799, year = 2016, url = https://arxiv.org/abs/1611.05799