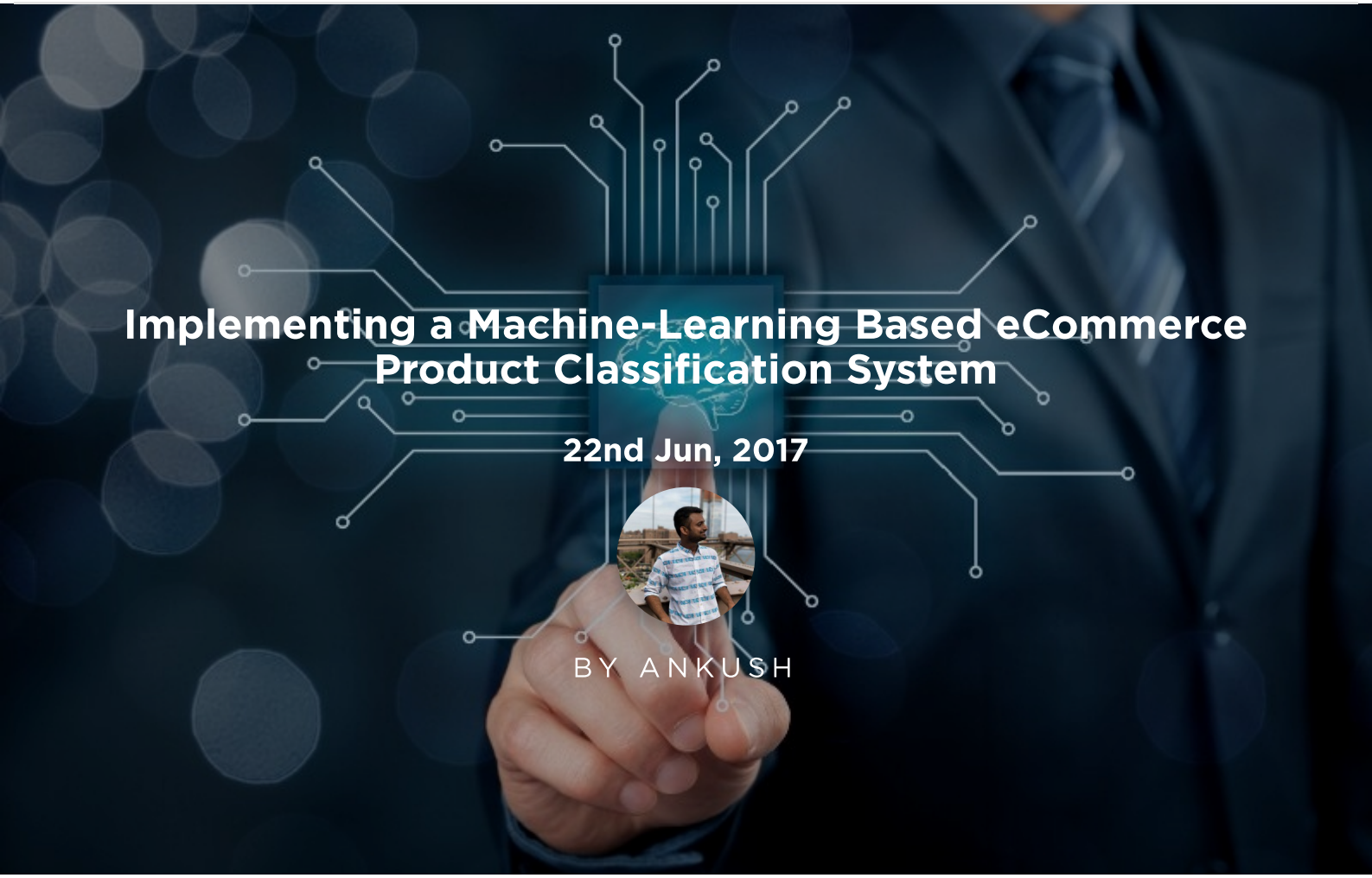


Implementing a Machine-Learning Based eCommerce Product Classification System

22nd Jun, 2017

BY ANKUSH

For online retailers, price competitiveness and a broad assortment of products are key to acquiring new customers, and driving customer retention. To achieve these, they need timely, in-depth information on the pricing and product assortment of competing retailers. However, in the dynamic world of online retail, price changes occur frequently, and products are constantly added, removed, and running out of stock, which impede easy access to harnessing competitive information.

At DataWeave, we address this challenge by providing retailers with competitive pricing and assortment intelligence, i.e. information on their pricing and assortment, in comparison to their competition's.

The Need for Product Classification

On acquiring online product and pricing data across websites using our proprietary data acquisition platform, we are tasked with representing this information in an easily consumable form. For example, retailers need product and pricing information along multiple dimensions, such as — the product categories, types, etc. in which they are the most and least price competitive, or the strengths and weaknesses of their assortment for each category, product type, etc.

Therefore, there is a need to classify the products in our database in an automated manner. However, this process can be quite complex, since in online retail, every website has its own hierarchy of classifying products. For

example, while “Electronics” may be a top-level category on one website, another may have “Home Electronics”, “Computers and Accessories”, etc. as top-level categories. Some websites may even have overlaps between categories, such as “Kitchen and Furniture” and “Kitchen and Home Appliances”.

Addressing this lack of standardization in online retail categories is one of the fundamental building blocks of delivering information that is easily consumable and actionable.

We, therefore, built a system that can predict a normalized category name for a product, given an unstructured textual representation. For example:

- Input: “Men’s Wool Blend Sweater Charcoal Twist and Navy and Cream Small”
- Output: “Clothing”
- Input: “Nisi 58 mm Ultra Violet UV Filter”
- Output: “Cameras and Accessories”

To classify categories, we first created a set of categories that was inclusive of variations in product titles found across different websites. Then, we moved on to building a classifier based on supervised learning.

What is Supervised Learning?

Supervised learning is a type of machine learning in which we “train” a system by providing it with labelled data. To classify products, we can use product information, along with the associated category as label, to train a machine learning model. This model “learns” how to classify new, but similar products into the categories we train it with.

To understand how product information can be used to train the model, we identified what data points about products we can use, and the challenges associated with using it.

For example, this is what a product’s record looks like in our database:

```
{
“title”: “Apple MacBook Pro Retina Display 13.3” 128 GB SSD 8 GB RAM”,
“website”: “Amazon”,
“meta”: “Electronics > Computer and Accessories > Laptops > Macbooks”,
“price”: “83000”
}
```

Here, “title” is unstructured text for a product. The hierarchical classification of the product on the given website is shown by “meta”.

This product’s “title” can be represented in a structured format as:

```
{
“Brand”: “Apple”,
“Screen Size”: “13.3 inches”,
“Screen Type”: “Retina Display”,
“RAM”: “8 GB”,
“Storage”: “128 GB SSD”
}
```

In this structured object, “Brand”, “Screen Size”, “Screen Type” and so on are referred to as “attributes”. Their associated items are referred to as “values”.

Challenges of Working with Text

Lack of uniformity in product titles across websites –

In the example shown above, the given structured object is only one way of structuring the given unstructured text (title). The product title would likely change for every website it’s represented on. What’s worse, some websites lack any form of structured representation. Also, attributes and values may have different representations on different websites — ‘RAM’ may be referred to as ‘Memory’.

Absence of complete product information –

Not all websites provide complete product information in the title. Even when structured information is provided, the level of detail may vary across websites.

Since these challenges are substantial, we chose to use unstructured titles of products as training inputs for supervised learning.

Pre-processing and Vectorisation of Training Data

Pre-processing of titles can be done as follows:

- Lowercasing
- Removing special characters
- Removing stop words (like ‘and’, ‘by’, ‘for’, etc.)
- Generating unigram and bigram tokens
- We represented the title as a vector using the Bag of Words model, with unigram and bigram tokens.

The Algorithm

We used Support Vector Machine (SVM) and compared the results with Naive Bayes Classifiers, Decision Trees and Random Forest.

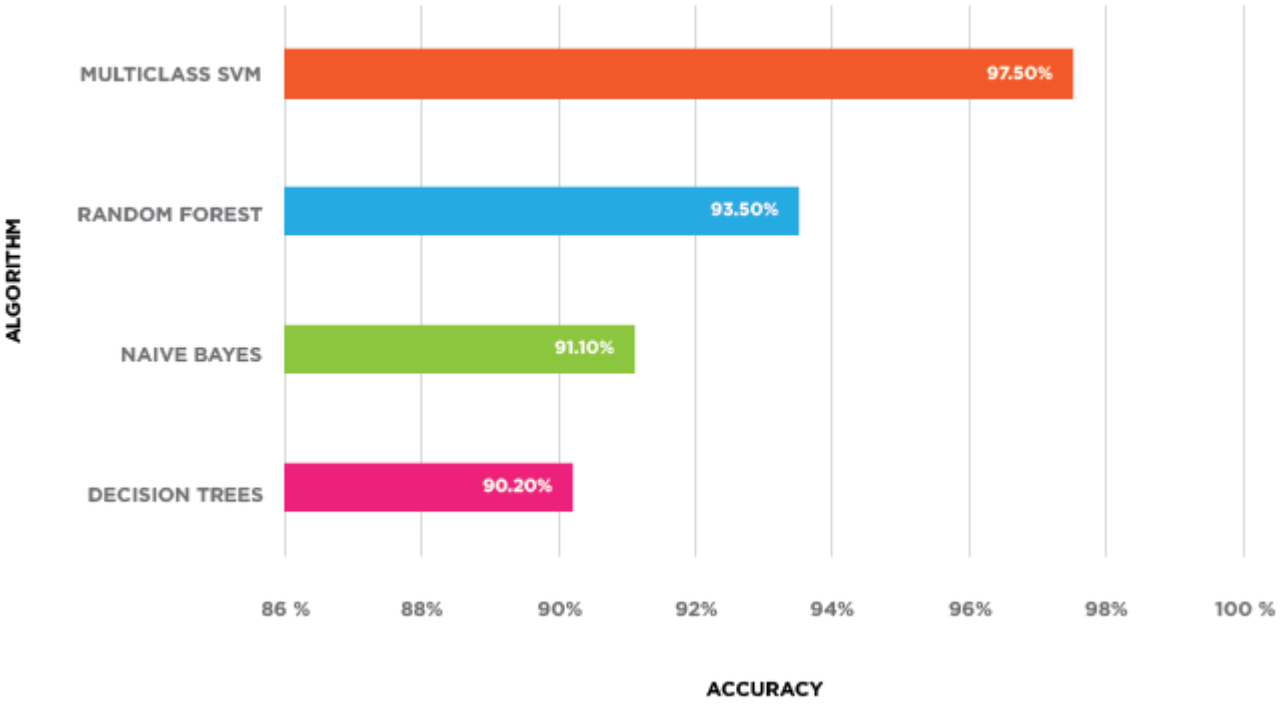
Training Data Generation

The total number of product data we've acquired runs into the hundreds of millions, and every category has a different number of products. For example, we may have 40 million products in "Clothing" category but only 2 million products in the "Sports and Fitness" category. We used a stratified sampling technique to ensure that we got a subset of the data that captures the maximum variation in the entire data.

For each category, we included data from most websites that contained products of that category. Within each website, we included data from all subcategories and product types. The size of the data-set we used is about 10 million, sourced from 40 websites. We then divided our labelled data-set into two parts: training data-set and testing data-set.

Evaluating the Model

After training with the training dataset, we tested this system using the testing dataset to find the accuracy of the model.



Clearly, SVM generated the best accuracy compared to the other classifiers.

Performance Statistics

- System Specifications: 8-Core system (Intel(R) Xeon(R) CPU E3-1231 v3 @ 3.40GHz) with 32 GB RAM
- Training Time: 90 minutes (approximately)
- Prediction Time: Approximately 6 minutes to classify 1 million product titles. This is equivalent to about 3000 titles per second.

Example Inputs and Outputs from the SVM Model (with Decision Values)

→ Input: “Washing Machine Top Load”

Output: {“Home Appliances”: 1.45, “Home and Living”: 0.60, “Tools and Hardware”: 0.54}

→ Input: “Nisi 58 mm Ultra Violet UV Filter”

Output: {“Cameras and Accessories”: 1.46, “Eyewear”: 1.14, “Home and Living”: 1.12}

→ Input: “NETGEAR AirCard AC778AT Around Town Mobile Internet — Mobile hot”

Output: {“Computers and Accessories”: 0.82, “Books”: 0.61, “Toys”: 0.27}

→ Input: “Nike Sports Tee”

Output: {“Sports and Fitness”: 1.63, “Footwear”: 0.63, “Toys and Baby Products”: 0.59}

Largely, most of the outputs were accurate, which is no mean feat. Some incorrect outputs were those of fairly similar categories. For example, “Home and Living” was predicted for products that should have ideally been part of “Home Appliances”. Other incorrect predictions occurred when the input was ambiguous.

There were also scenarios where the output decision values of the top two categories were quite close (as shown in the third example above), especially when the input was vague. In the last example above, the product should have been classified as “Clothing”, but got classified as “Sports and Fitness” instead, which is not entirely incorrect.

Delivering Value with Competitive Intelligence

The category classifier elucidated in this article is only the first element of a universal product organization system that we’ve built at DataWeave. The output of our category classification system is used by other in-house machine-learning and heuristic-based systems to generate more detailed product categories, types, subcategories, attributes, and the like.

Our universal product organization system is the backbone of the Competitive Pricing and Assortment Intelligence solutions we provide to online retailers, which enable them to evaluate their pricing and assortment against competitors along multiple dimensions, helping them compete effectively in the cutthroat eCommerce space.

Click here to find out more about DataWeave’s solutions and how modern retailers harness the power of data to drive revenue and margins.

- *Ankush Bhalotia*

Data Scientist at DataWeave, 22nd Jun, 2017

DATA ENGINEERING E COMMERCE