Deep recommender engine based on efficient product embeddings neural pipeline

Andrei Ionut DAMIAN
University POLITEHNICA of Bucharest
Bucharest, Romania
damian@cloudifier.net

Laurentiu Gheorghe PICIU
University POLITEHNICA of Bucharest
Bucharest, Romania
laurentiupiciu@gmail.com

Nicolae TAPUS
University POLITEHNICA of Bucharest
Bucharest, Romania
ntapus@cs.pub.ro

Bogdan DUMITRESCU
Hight-Tech System&Software
Bucharest, Romania
bogdan.dumitrescu@htss.ro

Abstract — **Predictive analytics systems are currently one of the most important areas of research and development within the Artificial Intelligence domain and particularly in Machine Learning. One of the "holy grails" of predictive analytics is the research and development of the "perfect" recommendation system. In our paper we propose an advanced pipeline model for the multi-task objective of determining complementarity, similarity and sales prediction using deep neural models applied to big-data sequential transaction systems. Our highly parallelized hybrid pipeline consists of both unsupervised and supervised models, used for the objectives of generating semantic product embeddings and predicting sales, respectively. This paper will argue that our system surpasses known and classical approaches to this particular problem. Our experimentation and benchmarking has been done using very large pharma-industry retailer Big Data stream.**

Keywords — **recommender systems; efficient embeddings; machine learning; deep learning; big-data; high-performance computing, GPU computing.**

I. INTRODUCTION

Recommender systems are by far one of the most important areas where machine learning in conjunction with Big Data are applied with proven success. The goal is to find what is likely to be of interest to a certain customer or group of customers and to provide personalized services to them. The interest in this area is very high because finding the "perfect" recommendation system is crucial and will allow retailers to structure their offer including sales strategy and marketing campaigns very well in order to optimize consumers choices.

This paper proposes a state-of-the-art deep neural model for the multi-task objective of determining product complementarity/similarity, which leads to diverse market baskets, and sales prediction. The pharmaceutical industry has been proposed as an experimental environment in our research. As a result, we used the advantage of a Big-Data sequential transaction system provided by a successful pharma retailer.

Our work presents a way to extend capabilities of recommender systems by learning low-dimensional vector space representation of products – or product embeddings - used in the following stages stage for sales prediction. We will compare our system with classical approaches to this particular problem (Collaborative Filtering*, Non-Negative Matrix Factorization*) and we will benchmark the results in order to demonstrate that our pipeline is better.

Semantically, our pipeline can be divided in two separate general models with end-to-end learning capabilities: the early stage of product semantic analysis and later stage of product sales regression. The key point within the initial stages of our pipeline system is learning product feature vector embeddings for each customer. Our work is analogous to Word2Vec* and GloVe* approaches which are used to learn linguistic regularities and semantic information for natural language processing. The general approach within the initial stage is based on analysing each sequence from the transactions database, choosing c products around a target product p_i in the same way a NLP system would analyse text semantics. The result will reveal that the embeddings generate clusters of complementary and similar products. As a result, through this approach we can identify a set of k items that will be of interest to a certain customer. Moreover, besides the product similarity, the resulting embeddings will bring to light an important aspect: if a product is not available anymore, it can be replaced with other two products whose embeddings sum will be very close to the initial product feature vector. Simultaneously, our system is able to generate product similarities and complementarities based on discrete time-series such as seasons.

In order to finish the pipeline, our work includes also a deep neural model which uses the feature vectors and predicts the sales. As we said, this is a crucial aspect for all companies if they want to set up an efficient marketing campaign.

Our pipeline was trained and tested using a Big-Data sequencial transaction system comprising more than 200 million purchases made by 1.7 million users, involving about 27,000 unique pharmaceutic products.

To the best of our knowledge, this work represents the first study that offers an end-to-end recommendation solution in the pharmaceutical field.

To summarize, our work will include:

- Analysis and benchmarking against *Collaborative Filtering* algorithm implemented in high-performance computing environment

- Analysis and benchmarking against *Non-Negative Matrix Factorization* algorithm implemented in high-performance computing environment
- Our proposed P2E (Product-to-Embeddings) and ProVe (Product Vectors) models
- Our proposed deep neural model for final stages of sales prediction

II. Related Work

Our work relates with several approaches either derived from Natural Language Processing or from classic methods that address the problem of recommendation systems

Existing methods for recommender systems can easily be categorized into **collaborative filtering** methods* and content-based methods*, which make use of the user or product content profiles. Collaborative filtering is based on user-item interactions and predict which products a user will most likely be interested in by exploiting purchase behaviour of users with similar interests or by using user's interaction with other products. In practice, CF methods are very popular because they can discover interesting associations between products and do not require the heavy knowledge collection needed by content-based methods. To mitigate the cold-start problem, which CF methods suffer from, matrix factorization based models have been developed and now they are very popular after their success in the Netflix competition.

Matrix factorization for collaborative filtering models can approximate a sparse user-item interaction matrix by learning latent representation of users and items using SVD or stochastic gradient descent which give the optimal factorization that globally minimizes the mean squared prediction error over all user-item pairs.

In a number of Natural Language Processing (NLP) tasks, such as computing similarity between two documents, learning linguistic regularities and semantic information are essential. Therefore, a mathematical model has been developed by Mikolov et. al* (**Word2Vec**), which can be used for learning high-quality low-dimensional word embeddings from huge datasets and huge vocabularies, using two architectures of neural networks: continous bag-of-words (CBOW) and skip-gram (SG).

This powerful and efficient model takes advantage of the word order in the text documents, explicitly modeling the assumption that closer words in a context window are statistically more dependent. In the SG architecture, the objective is to predict a word's context given the word itself, whereas the objective in the CBOW architecture is to predict a word given its context.

GloVe* is a novel model for learning low-dimensional vector representations of words by combining the advantages of two major model families in the NLP literature: global matrix factorization and local context window methods (Word2Vec).

They consider the primary source of information available about a corpus of words being the word-word co-occurrence counts which is used to train the fine-grained word embeddings. Explicitly, the ratio of the co-occurrence probabilities of two words (rather than their co-occurrence probabilities themselves) is what contains the information encoded as word embeddings.

Traditionally, in NLP applications, each word is represented as a feature vector using a one-hot representation where a word vector has the same length as the size of the vocabulary. Our first approach was to create a corpus of words from all pharmaceutical prospectuses and to encode each product, using **hand engineered features,** where feature *i* is 1 if the word *i* appears in the prospectus of a product and 0 otherwise. Then, we created a model based on XGBosst Regression Trees* which predicted sales with a 91% recall. However, this approach suffers from high dimensionality and data sparsity and does not meet our first scope - predicting market baskets.

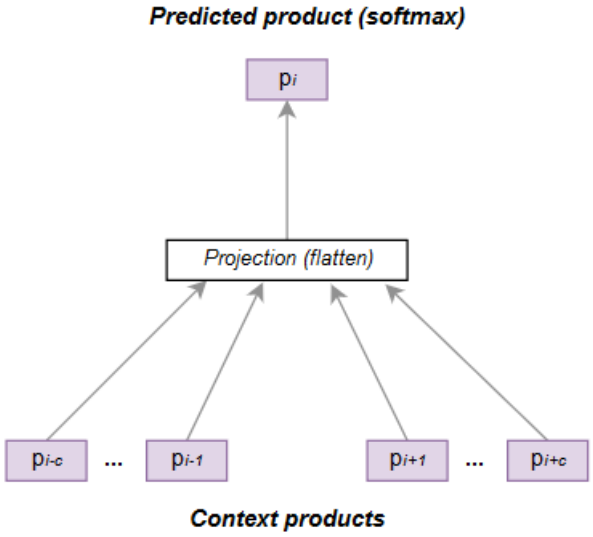

Figure 1 – P2E CBOW Architecture

III. Approach

A. *From NLP to Reccommender Systems*

Considering the big improvement word embeddings brought to NLP domain, we were confident that from language models to product business analytics is a very slight difference.

To address the task of finding complementarity/similarity between products and diversified market baskets for a certain customer, we proposed to learn representations of products in low-dimensional space, using the Big-Data sequencial transaction system provided by the pharma retailer, which was used also for computing the word-word co-occurrence counts.

More specifically, we developed two models (**P2E** and **ProVe**) inspired by the NLP models **Word2Vec** and **GloVe**.

Our first model uses the sequentiality of transactions. The transaction system can be represented as a set S, where $S = (t_1, t_2, t_3, \dots, t_M)$, M = total number of transactions and $t_i = (p_i,\ timestamp_i)$ (a product bought at a certain time). Our objective is to find D-dimensional real-valued representations $w_{p_i} \in \mathbb{R}^D$ of each product p_i such that they lie in a latent vector space. The general approach within this initial stage is based on analysing each sequence from the transactions database, choosing c products around a target product pi in the same way a NLP system would analyse text semantics (Figure 1). Therefore, the objective function of CBOW architecture is defined as follows

$$J = \frac{1}{M}\sum_{i=1}^{M} \log \mathbb{P}(p_i \mid p_{i-c}, \dots, p_{i-1}, p_{i+1}, \dots, p_{i+c})$$

where probability $\mathbb{P}(p_i \mid p_{i-c}, \dots, p_{i-1}, p_{i+1}, \dots, p_{i+c})$ of observerving the current product p_i, given a product context is defined using the softmax function

$$\mathbb{P}(p_i \mid p_j) = \frac{e^{w_{p_j}^T w_{p_i}}}{\sum_{k=1}^{P} e^{w_{p_j}^T w_{p_k}}}$$

The original **Word2Vec** model proposes a sampled-softmax loss which is a computationally efficient approximation of the full softmax. Given the fact that our models are trained in a high-performance computing environment, we used the non-approximated version which resulted in better results. Simultaneously, the projection layer computes a concatenation of all $c * D$ embeddings, instead of sumation. This aspect leads to more computation for the softmax layer, but also to a better accuracy, as we will show in Section IV.

The second model (**ProVe**) seeks to learn low-dimesional product embeddings using the ratio of co-occurrence probabilities of two products. Therefore, the model generates two sets of product vectors W and $\widetilde{W}$ and it should minimize the weighted least squares objective J defined as follows:

$$J = \sum_{i,j=1}^{P} f(X_{ij})\,(w_{p_i}^T \tilde{w}_{p_j} + b_i + \tilde{b}_j - logX_{ij})^2$$

To achieve this, we need to define our transaction system as a set

- $S = (b_1, b_2, b_3, \dots, b_N)$, N = total number of market baskets (receipts);
- $b_i = \left(p_{i_1}, p_{i_2}, \dots, p_{i_{B_i}}\right), B_i$ = total number of products on i -th receipt.

and to define our word-word co-occurrence score as following: $X_{ij} = X_{ij} + \frac{1}{d(p_i,p_j)}$, where p_j represents a product that is in the context of the product p_i for each receipt and d represents the distance between these products in a context window.

For both **P2E** and **ProVe**, we applied K-Means algorithm on the resulting embeddings which is a naïve approach for our objective of determining the **concept vectors** which should capture the needs of a certain pharmaceutic product. This will be the point where we will can obtain 100% complementarity in a market basket. The distance between two feature vectors is computed using **cosine similarity**.

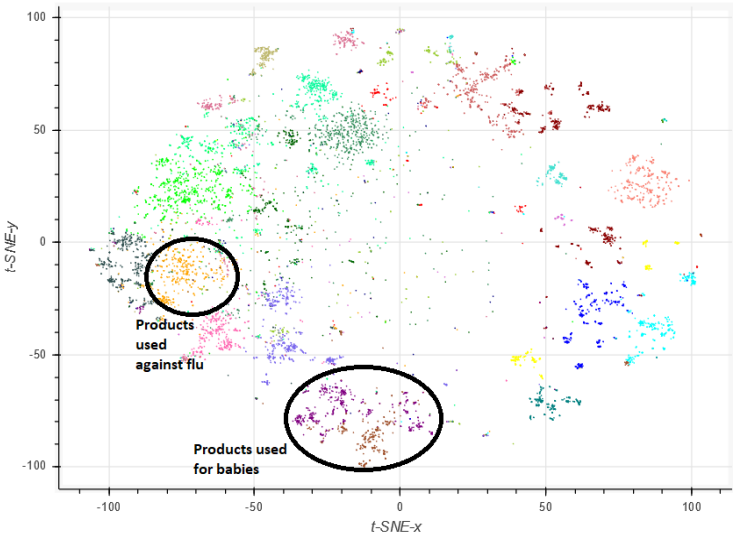

Figure 2 – t-SNE representation of the P2E resulting embeddings

B. Latent Product Space

The resulting embeddings show encouraging results. The products lay in a latent space where they are grouped based on their main need. Our following research will include mainly finding the concept vectors that lead to products separations.

Figure 2 shows a graphic representation of all 64-D embeddings resulted from P2E model for the first 10,000 most transactioned products. The 2-D coordinates of the products were computed using t-SNE method*.

C. Neural Product Recommender Regression

TODO