Improved Forecasting and Purchasing of Fashion Products based on the Use of Big Data Techniques

Conference Paper · March 2015

DOI: 10.1007/978-3-658-08809-5_13

2 authors, including:

Diane Ahrens
Deggendorf Institute of Technology
16 PUBLICATIONS **102** CITATIONS
SEE PROFILE

Improved Forecasting and Purchasing of Fashion Products based on the Use of Big Data Techniques

Ali Fallah Tehrani and Diane Ahrens

Abstract

Ordering proper amount of products, taking into account the demand of market, in fashion retail industries is one of the core challenges. Essentially due to the fact that the ordering is typically performed once in the each season, it is absolutely required to carry out precise orders. To make a precise ordering as well as to prevent overstocks and stock-out, there is a need for reliable forecasting methods. A reliable forecasting requires to consider proper predictive models which can consider all deciding factors. Specifically in the case of fashion forecasting since each product is associated with several factors, e.g. price, style, color and even human factors, learn a suitable predictive model is not an easy task. In fact, the challenge here boils down to learn a powerful model, which can cover all these information. To this end, *big data techniques*, namely *data mining* and *machine learning* methods serve the ability to accomplish the challenge. In this paper, we exploit unsupervised learning methods for a goal fitting the data, particularly w.r.t simple models although with higher gain. In essence, our innovative model is able to modify simple regression model, and hence, provide more promising results. In this regard, we apply *big data analyses* and *techniques* specifically in fashion field to analyze and make the sales-prediction.

Keywords: fashion forecasting, big data techniques, machine learning, clustering,

1. Introduction

Business forecasting as a subfield of supply chain management aims to optimize resources or say to reduce losses. Since the resources are limited as well as costly, the business forecasting has attracted increasing attention specifically in recent decades. One of the major task in business forecasting is to optimize the number of sales for various products w.r.t. the market demand. Indeed, the problem can be formulated underlying an optimization problem, where the task is to find optimal parameters regarding the products. However, what makes the optimization problem complicated, are the input factors. Needless to say, to sell a product intrinsically numerous factors are involved, which should be taken into consideration. Thus, the challenge itself comes down to deal with numerous factors underlying products.

More concretely, we are interested in a slightly different problem, namely identifying relationships between ordering and selling in **fashion branch**. Typically, clothing companies suppose a kind of necessity in the market at the beginning every season or even earlier (Mostard et al., 2011; Yu et al., 2011), and on the basis of this requirement they would order the fashion products. Such plans or say judgments are usually made by some experts in related field regarding on the history of sales and the parameters in current situation. Since numerous of parameters are involved, the plan is not optimal, namely the demand in the market is usually less, which is addressed under overstock. The overstock problem occurs when the number of products is more than what is actually needed. Overstock typically leads to discount and consequently to the loss of revenues. On the contrary, understock occurs, when the number of products is less than the market demand. Moreover we would emphasize that sometimes the experts are under bias, which means their opinion would be subjective. For instance, there are some established patterns for women and men in the literature for specifying fashion trend. To make a sound and transparent judgment, it is certainly needed to capture all useful information in history of sale, aggregating with the parameters in current situation. For this purpose, it is recommended to apply **data analysis techniques**, however, depends on the problem of course suitable tools must be taken into account.

Worth mentioning that studying such problems is quite important from several reasons, firstly it is related to manage the raw products. Secondly it tries to model the so-called customer's interest. In this regard, this paper applies big data techniques to analyze the dependency between input factors and sale in the case of fashion retailer which exclusively sells through catalog and e-commerce channels. Accordingly we employ data analysis techniques for the goal loss minimization. More concretely, the ultimate goal is to predict the demand in the market taking into account the feedback of customers as training data. This problem indeed can be tackled by a regression model, however, a simple regression model cannot model such dependency in a sound manner. The simple reason is as follows: high fluctuation in input space causes more deviation in output space. Since typically the optimal solution of linear regression underlying the squared Euclidean norm is "mean", therefore it is not possible to optimize linear regression precisely, which indeed leads to a poor prediction.

The rest of this paper is organized as follows: in Section 2 related researches are presented. In Section 3, we recall the data analysis techniques and discuss on some related approaches. The core idea of algorithm underlying risk minimization is explained in Section 4. The details of our algorithm are described in Section 5. In Section 6, the data is described and first preliminary results are demonstrated. Finally in Section 7, the outlook is given.

2. Related Researches

Frankly speaking, there is no comprehensive research on fashion forecasting underlying big data techniques, however there are some exceptions: Thomassey et al. (2007) employed neural network method, specifically Neural Clustering and Classification (NCC) to tackle the problem. In addition Thomassey et al. (2002) used classification method to cope with mean-term forecasting. Happiette et al. (1996) applied partition technique to cluster similar trend. Then each cluster has a similar characteristic and ultimately the products in each cluster have similar behavior in terms of selling. Apart from forecasting on demand, there are researches focused on trend forecasting for the case of color (Gu et al., 2010; Choi et al., 2012). Frank et al. (2003) deals with the forecasting in a different manner, in fact they proposed online-learning. Basically supposing daily lag and weekly lag, they define on each lag auto correlations functions (ACF). Since they can monitor feedbacks dynamically, it is possible to improve the accuracy of prediction by neural network method. Also to overcome this challenge Bayesian techniques are applied (Green et al., 1973; Yelland et al., 2014). Sun et al. (2008) and Xia et al. (2012) employed also machine learning methods for the goal forecasting.

Worth mentioning that in (Liu et al., 2013; Choi et al., 2013) a survey on the existing approaches w.r.t. the sales forecasting especially related to artificial intelligence (AI) is given.

The terminology prediction is a widely used term in machine learning, however, in this article refers to the *forecast of demand* in the market.

3. Data Analysis

In the literature analyzing the data can be characterized by descriptive, exploratory and inductive data analysis. While the task in descriptive and exploratory data analysis is to analyze the data by statistical approaches, in inductive data analysis the task is to make an induction given the observations. We describe these approaches in the following great in the details. It is also worth mentioning that beforehand the data should be modified and prepared. For instance, it is not surprising that under a wrong data transformation there would exist a large bias.

3.1 Data Preparation

In order to make sure that the experiments are conducted in a consistent way, before applying any method, it is necessary to prepare the data. Formally data preparation involves data discretization, data cleaning, data integration, data transformation and data reduction.

In our experiments we standardize the data by applying `zscore`, to produce more reliable results underlying clustering. In fact, clustering is very sensitive to the magnitude of attributes. To avoid wrong bias, it is more convenient to use almost the same range for all attributes.

3.2 Descriptive Data Analysis

Extraction sound information from observations in many applications is desirable, however depends on demands, proper techniques should be taken into account. Descriptive data analysis can be seen as a first step in data analysis, which provides rudimentary information about population usually w.r.t. a single variable. Roughly speaking, the duty of descriptive analysis is to present and summarize the population by charts or numerical values. Frequency distributions, central tendency, mode, median, mean, variance and range are most common tools in this field.

3.3 Exploratory Data Analysis

Beyond focusing on a single variable, it is quite desirable to analyze whole population, in which statistical variables are jointly considered. To this end, exploratory data analysis serves approaches to cover these information. Essentially underlying the exploratory data analysis is possible to measure the dependency between statistical variables, which makes the analysis more understandable. The key difference between the descriptive and the exploratory analysis is actually the way to handle the variables. In this regard, correlation analysis, principal component analysis (PCA), box plot, vector quantization (VQ), scatter plot and matrix factorization are relevant tools in this field.

k-means Clustering

Identifying similar objects can provide sound knowledge in terms of studying the structure of data, which can provide better interpretation and data understanding. To this end, the common tool in data mining is addressed as **clustering**. The basic idea underlying clustering is roughly to group similar objects by their properties, however depending on the definition of similarity, the similar objects can be treated in distinctive ways. From a machine learning point of view, clustering belongs to the un-supervised learning. Thus, clustering methods do not consider any response in general, although it is possible to take into account responses as an extra attribute. This enables us to partition the similar objects w.r.t. the output, which is slightly similar to the classification approach.

Particularly, in the case of k-means clustering the goal is to partition instances in k different groups, where k is given in advance. More formally, given n observations in m-dimensional space:

$$\{x_i\}_{i=1}^{n} \subset \mathbb{R}^m,$$

the goal is to identify a set $C = \{C_1, \ldots, C_k\}$ such that the following objective function is minimized:

$$\sum_{i=1}^{k} \sum_{x_j \in C_i} d\left(x_j, \mu_i\right), \tag{1}$$

where μ_i is the mean w.r.t. set C_i and moreover for all $i, j \ (i \neq j)$

$$C_i \cap C_j = \varnothing, \bigcup_{j=1}^{k} C_j = \{x_i\}_{i=1}^{n} \cdot$$

In the literature the distance $d(\bullet,\bullet)$ which indeed measures the dissimilarity between two objects can be substituted by the following distances:

- Squared Euclidean Distance

$$d_E(x, y) = \sum_{i=1}^{m} \left\| x_i - y_i \right\|^2 \cdot \tag{2}$$

- L_1 Distance, City Block Distance, Manhattan Distance

$$d_M(x, y) = L_1(x, y) = \sum_{i=1}^{m} \left| x_i - y_i \right| \cdot \tag{3}$$

k-means clustering underlying Manhattan distance is called k-*medoids* clustering (Hastie et al., 2001) which has ability to handle different types of attributes (numeric, binary, categorical) simultaneously. Hence, there is no need to consider different distances for different types of attributes (Ehmke, 2012).

- Cosine Distance

$$d_C(x,y) = 1 - \frac{\langle x, y\rangle}{\|x\|\|y\|}.$$

- Correlation Distance

$$d_{corr}(x,y) = 1 - corr(x,y),$$

where $x, y \in \mathbb{R}^m$.

Note that, taking into account different types of attributes (numeric, binary, categorical) indeed it is needed to suppose different similarity measures. As mentioned, a typical solution is k-medoids; different class of attributes can be tackled under L_1-norm. However, theoretically it is possible to consider several similarity measures, suitable for each type of attribute and finally aggregate the measure (Ahmad et al., 2007). More formally, in our case, when there are numeric and binary attributes, one can assume minimizing the following objective function:

$$\underset{C_1,\cdots,C_k}{\arg\min} \leftarrow \sum_{i=1}^{k} \left\{ \sum_{x_j \in C_i} d\left(x_j^*, \mu_i^*\right) + \Psi \sum_{x_j \in C_i} \delta\left(x_j^{**}, \mu_i^{**}\right) \right\}. \tag{4}$$

Here Ψ is a kind of trade-off parameter, which must be determined in advanced, and δ is a similarity measure underlying binary attributes. In above setting x_j^* and x_j^{**} refer to numeric part and binary part of object x_j respectively. We call this clustering method "k-means mixed". In fact, the difficulty here is to identify a proper Ψ parameter. Unfortunately, there is no way to figure out, whether the parameter Ψ is chosen correctly or not. Note that, choosing different values for Ψ parameter leads to different clusters, which from a consistency point of view is not desirable.

3.4. Inductive Data Analysis

On the contrary, inductive data analysis aims to generalize observations, i.e., given observations the goal is to induce a suitable model. More concretely, the task is to identify optimal hypothesis in hypotheses space (the space of all existing hypothesis is called hypotheses space) given some observations. Particularly in our case, the observations are thought of as a subset of m-dimensional space ($\mathbb{R}^m$). The induction-process is called *learning underlying the data*, and usually is referred to *machine learning* as a main framework. The most popular learning-problems are regression (Hastie et al., 2001), classification (Bishop, 2006), multi-class classification (Hastie et al., 2001), multi-label classification (Tsoumakas et al., 2007) and learning the preferences (Frünkranz et al., 2011).

Regression

Particularly, in our case the observations are supposed as follows:

$$\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n} \subset \mathbb{R}^m \times \mathbb{R},$$

where x_i is representing an observation; which in the literature is called a regressor w.r.t. response y_i. Moreover, we assume the data points are i.i.d. (identically independently distributed). Roughly speaking, here the goal is to model a dependency between regressors and responses. Typically, we are interested in a vector ω^* which minimizes following objective function:

$$\omega^* \leftarrow \arg\min_{\omega} \sum_{i=1}^{n} d\left(F\left(\omega, x_i\right), y_i\right), \tag{5}$$

where $d(\bullet,\bullet)$ is a distance function and $F(\bullet,\bullet)$ describes the dependency between regressors and the vector ω. In the experiments the function $F\left(\omega, x_i\right)$ is considered as $\left\langle \omega, x_i \right\rangle$, which in the literature is addressed as linear regression. In addition, the distance function $d(\bullet,\bullet)$ is supposed as the Euclidean distance.

4. Dealing with Outliers underlying Linear Regression

Accounting for the fact that trend products are highly sold, there is always a gap between common products and trend products. From a descriptive data analysis point of view, trend products underlying their number of sales can be assumed as outliers. The ordinary linear regression is poor to model outliers, as it cannot cope with deviation from linearity. This indeed is a disadvantage, and causes a low performance. To cope with this weakness, we propose a method on the basis of clustering.

Before introducing our predictive model, we shall recall some preliminaries regarding linear regression. From a linear regression model point of view underlying L_2-norm, the optimal solution is obtained by mean:

Assume Y is a random variable underlying density function $f_Y(\bullet)$, and the goal is to estimate a parameter $\overset{\wedge}{\theta}$ which minimize the following risk function:

$$E\left[\left(Y-\overset{\wedge}{\theta}\right)^2\right], \tag{6}$$

where the function $E(\bullet)$ stands for the expectation operator. Following the classical approaches in the probability theory, we obtain:

$$\begin{aligned} E\left[\left(Y-\overset{\wedge}{\theta}\right)^2\right] &= \int_{-\infty}^{+\infty}\left(y-\overset{\wedge}{\theta}\right)^2 f_Y(y)\,dy \\ &= \int_{-\infty}^{+\infty} y^2 f_Y(y)\,dy + \int_{-\infty}^{+\infty} \overset{\wedge}{\theta}^{2}\ f_Y(y)\,dy - 2\int_{-\infty}^{+\infty} y\overset{\wedge}{\theta}\, f_Y(y)\,dy \cdot \end{aligned} \tag{7}$$

Taking a derivative from Equation (7) with respect to the $\overset{\wedge}{\theta}$ and seeking the zeros, leads to the following equality:

$$0 = \frac{\partial E\left[\left(Y-\overset{\wedge}{\theta}\right)^2\right]}{\partial \overset{\wedge}{\theta}} = 0 + 2\overset{\wedge}{\theta}\int_{-\infty}^{+\infty} f_Y(y)\,dy - 2\int_{-\infty}^{+\infty} f_Y(y)\,dy \cdot \tag{8}$$

Therefore

$$\overset{\wedge}{\theta} = \frac{\int_{-\infty}^{+\infty} y f_Y(y)\,dy}{\int_{-\infty}^{+\infty} f_y(y)\,dy} = \frac{E[Y]}{1} = \mu \cdot$$

This means that the risk function is minimized at mean.

It is also not difficult to show that for any statistical estimator A defined on random variables $X_1 \times \cdots \times X_m$,

$$E\left[\left(Y - E[Y]\right)^2\right] \leq E\left[\left(Y - A[\mathrm{X}_1, \cdots, \mathrm{X}_m]\right)^2\right]\cdot \tag{9}$$

Specifically the above inequality is valid for the linear regression. More concretely, let us assume n data points $\left\{(x^1, \mathrm{y}^1), \cdots, (x^n, y^n)\right\} \subset \mathbb{R}^m \times \mathbb{R}$ are given, and moreover envisage $\hat{\theta}_i = \sum_{j=1}^{m} \omega_j x_j^i$, where $x^i = \left(x_1^i, \cdots, x_m^i\right) \in \mathbb{R}^m$. Assuming $\omega = \left(\omega_1, \cdots, \omega_m\right)$ as a free parameter, the target is to find an optimal regression vector, namely the vector ω which minimizes the following function:

$$\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{m} \omega_j x_j^i - y^i\right)^2\right]\cdot \tag{10}$$

Roughly speaking, the problem comes down to solving the following system of equations:

$$\begin{cases} \sum_{j=1}^{m} \omega_j x_j^1 = \mu_D \\ \cdot \\ \cdot \\ \cdot \\ \sum_{j=1}^{m} \omega_j x_j^n = \mu_D \end{cases},$$

where μ_D is the mean of $\left\{y^i\right\}_{i=1}^{n}$.

In general there is no guarantee that the above system of equations has an analytical solution. In fact, since usually the number of data points and the number of attributes are not equal, the solution cannot be determined by solving the system of equations analytically. In fact, the optimal solution can be determined by taking the derivative directly from the objective function. Note that, here we are not only interested in the optimal solution, but also interested in the way to determine it. Equivalently the optimal solution of the system of equations can be estimated by minimizing the following objective function:

$$\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{m}\omega_j x_j^i-\mu_D\right)^2\right]\cdot \qquad (11)$$

This is quite clear, that it is not possible to find a vector ω, which makes the function in (11) equal to zero. The simple reason is that, the data points are fluctuating. In extreme case, when all data points are equal, the "ideal-optimal solution" can be obtained by solving $\left\langle \omega, x^i \right\rangle = \mu_D$, although here there is a numerical instability problem; means the ideal-optimal solution is not unique.

To establish a more consistent solution our idea refers to shrink the space of solution by shrinking the input space. More concretely, let us assume there is an

m-dimensional sphere[1] called S, with center $\left(o_1,\ldots,o_m\right)$ and radius $\varepsilon\cdot$ Without loss of generality, assume that $\omega_j \geq 0$[2] and $\sum_{j=1}^{m}\omega_j o_j = \mu_D\cdot$ Hence, for each x in sphere S,

$$\sum_{j=1}^{m}\omega_j\left(o_j-\varepsilon\right)<\sum_{j=1}^{m}\omega_j x_j<\sum_{j=1}^{m}\omega_j\left(o_j+\varepsilon\right)\cdot$$

By extending the above inequality, the following inequality is obtained:

$$\sum_{j=1}^{m}\omega_j o_j-\varepsilon\sum_{j=1}^{m}\omega_j<\sum_{j=1}^{m}\omega_j x_j<\sum_{j=1}^{m}\omega_j o_j+\varepsilon\sum_{j=1}^{m}\omega_j\cdot \qquad (12)$$

Therefore

$$\mu_D-\varepsilon w<\sum_{j=1}^{m}\omega_j x_j<\mu_D+\varepsilon w,$$

where $w=\sum_{j=1}^{m}\omega_j\cdot$

The last inequality actually assures that by choosing the data points from sphere S, the output of linear regression is bounded in $\left(\mu_D-\varepsilon w,\mu_D+\varepsilon w\right)$. Here there is a

[1] Note that the claim is valid for L_1-norm and L_2-norm.

[2] This can be done by multiplying corresponding attributes to -1.

meaningful intuition between the radius of sphere and the optimal solution. Let suppose $\varepsilon \to \infty$, in other words, we take into account all data points. Therefore, the output of the model is unbounded, and the optimal solution converges to the solution of common linear regression. In this case, the solution achieved by classical methods are affected by fluctuations in the data. Therefore, it exhibits a poor performance.

In another extreme case when $\varepsilon \to 0$, each sphere contains only its center point. In this case there are infinite strictly minimum solutions, however, because of lacking of data they are not representative. Note that, in this case lack of data causes over-fitting problem and leads again to a poor performance.

Loosely speaking, the optimal solution is obtained by choosing a proper trade-off between number of data in each sphere and the radius of sphere; proper number of clusters. For instance, this can be done by applying a nested cross validation.

5. Modifying Prediction in the Light of Clustering

As mentioned earlier, the ultimate goal of this article is to make a prediction, namely, adapt a model on the basis of seen observations to estimating the output for an unseen observation. In fact, the difficulty here is to choose a proper model which can cover all properties. We already mentioned that the data contains various products, in which each product has its specific characteristic. In general, one can assume a powerful and flexible model which can capture all these properties. However, fitting advanced models usually costs time (running time), causes over-fitting (especially due to the lack of data) and in addition they are not easy to interpret. Note that, in this case we are not only interested in the so-called gain but also the interpretability is a crucial point.

As mentioned in previous section, to establish a high performance predictive model underlying simple regression, the so-called outliers should be handled in a convenient way. To this end, our idea refers to simplify the original learning problem by saying that similar objects from a numerical point of view can be tackled by a same learner. Note that, here for example a trouser and a shirt might be assumed as similar objects, since they might have similar numerical characteristics, whereas from a customer point of view, they seem differently.

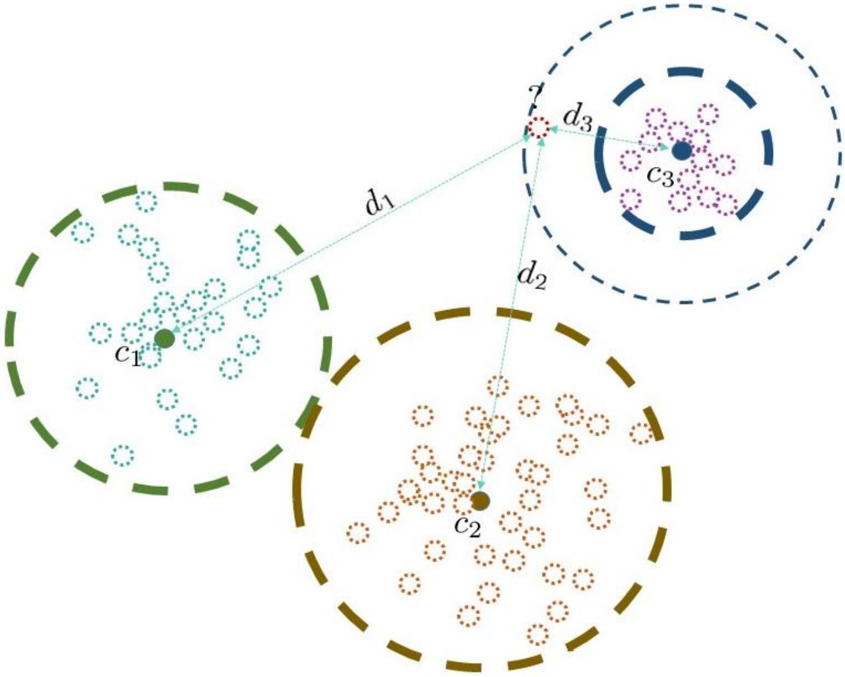
This makes possible to utilize several basic learners, and in the end the adapted learners are aggregated. In the following we describe the algorithm in greater details:

Training phase:

Loosely speaking, in the training phase the algorithm contains of two major steps: the first step is referred to identifying similar objects, which can be achieved underlying clustering approach. The next step actually is to fit different learners for different clusters. In fact, for various cluster, various type of learner can be taken into account, however, for our experiments we apply solely linear regression. Basically we apply `lsqcurvefit` function in MATLAB. The function allows us to define arbitrary objective function for the goal of fitting. More formally, for each cluster C_k we take the following objective function into consideration:

$$\omega^*_k \leftarrow \arg\min_{\omega} \left\{ \sum_{x^j \in C_k} \left(\omega_0 + \sum_{i=1}^{m} \omega_i x_i^j - y_i^j \right)^2 \right\}, \qquad (13)$$

***Figure 1:** Illustration of k-means clustering approach facing a new instance*



where $x^j = \left(x_1^j, \ldots, x_m^j\right) \in \mathbb{R}^m, y^j \in \mathbb{R}$ and ω_0 is bias term. Supposing learnt optimal weights, namely $\left\{\omega_i^*\right\}_{i=1}^K$ are given, a non-trivial question is, how an unseen object can be evaluated?

Evaluating phase:

To this end, the algorithm finds the cluster, which has a minimum distance to the unseen object. To simplify the notation, we use C_k as $k-th$ cluster an c_k as the center of cluster C_k .

Thus, it is possible to compute the distance between the object and the center of each cluster. More concretely, it finds the nearest cluster as follows:

$$l \leftarrow \arg\min_{1 \le i \le K} d\left(c_i, x\right) \cdot \tag{14}$$

Then a prediction is carried out on the basis of learnt weights w.r.t. the cluster C_l , e.g. ω_l^* . The illustration of clusters and center points are demonstrated in Figure 1. As it can be seen, new instance is closer to the center point c_3 , hence it belongs to the cluster C_3 . Predicting a response of the new instance is accomplished by applying learned weights underlying cluster C_3 namely $F\left(\omega_3^*, x\right)$.

6. Case Study

We performed a study in the fashion field, where it is possible to monitor the feedback of customers over time. The data is taken from an apparel retailer, which produces varies clothes and accessories and known as a large producer in this field. The task here is to model dependencies between products and demand. Moreover, the second goal is to predict a demand of the market for future.

Accounting to the fact that the demand of the market varies in a dynamical way, it is certainly required to make a prediction on these dynamical changes. In fact, fashion products are quite sensitive to the customer's feedback, which makes it very dynamic. Seen from this view, it is certainly required to make a prediction for future. Such forecasting is not only in the use of company, but also is in the use of customer as well resources. Hence, making a precise forecast assuming the history of sales and actual parameters is a crucial point. We would like again to emphasize that since the information are very large and might be contradicting, it is not possible to handle the problem analytically.

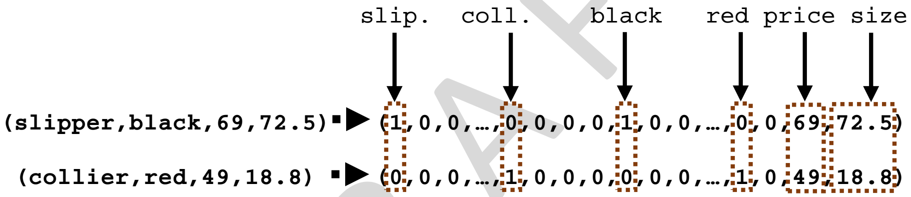
6.1 Data Description

Each product is characterized by 5 attributes as follows:

`type of product, color, price, size of product, size of photo`

The output considers number of sales w.r.t. the products. In our case, the data is cumulated under size, meaning if for different products all attributes except size of product are equal, then one product with cumulated number of sales remains. Finally the data set contains 1421 different cumulated products.

In order to convert each qualitative feature to a nominal value the feature space is expanded by considering dichotomous variables. Basically, for each color and for each type of prodcut, a dichotomous variable is added respectively. This procedure allows us to deal with qualitative attributes in a nominal manner, which certainly can be seen as an advantage. In this case, the feature space has in total 250 attributes. An alternative could be to convert the colors by their RGB properties to basic colors, although it is not possible to do the same procedure for the different type of products. In the following, by way of example the converting procedure is explained:



Accordingly, in the presence of each qualitative attribute (type of product, color) the corresponding attribute in extended feature space, takes nominal value 1, whereas in the absence of the qualitative attribute it takes value 0.

Table1: *basic statistical information regarding data*

Attributes	Numeric			Binary	
No. of Attributes	3			248	
Size of dataset	1421			1421	
Name of Attributes	*Price (€)*	*Size (cm^2)*	*Sales*	*Colors*	*Type of product*
Mean	145.69	26.98	107.29		
Median	119.95	13.62	44		
Std.	182.50	41.11	188.53		

An overview of descriptive statistics is given in Table 1.

6.2 Data Analysis

Typically data related to business forecasting are tremendous, conflicting and contain noises, therefore extracting sound patterns from them is not a trivial task. To accomplish the challenge, classically big data techniques can be occupied, which allow us to deal with enormous information elegantly.

Now we employ the exploratory and inductive data analysis methods and demonstrate the corresponding results. In the first experiment we conducted an experiment by clustering approach to recognize similar products. Then forecasting results, which its learning process is conducted underlying clustering are presented.

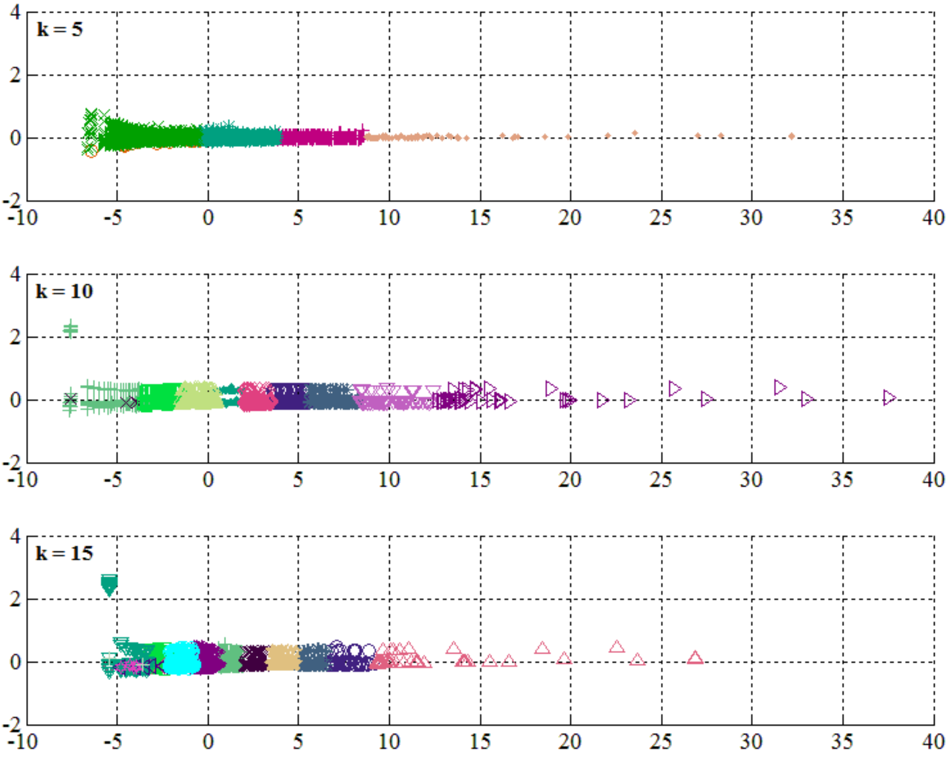
6.3 Clustering Analysis

Recognizing similar objects allows to establish homogenous groups, which enables us to facilitate the original problem. In fact, the goal in this case is to detect similar objects by means of clustering; needless to say similar objects have similar characteristic. Thus, it is possible to handle the similar objects by a same approach, e.g. linear regression or kernelized logistic regression (Zhu et al. 2001). In essence, taking into account whole data begs to consider more advanced learners, which from a computational point of view, usually makes the problem indeed convoluted. However, by clustering the data it is possible to apply simpler learners, which from a complexity point of view clearly is an advantage.

The illustration of different levels of clustering is showed in Figure 2, which as distance function the L_1-norm is taken into account. Indeed, the idea is to group the products in homogenous groups underlying Manhattan distance. Needless to say, all distances, which already introduced in Section 3.3 can be applied, however, can provide different clusters. In fact, k-means clustering is extremely sensitive to the distance function. In our case, since we are dealing with different class of attributes, the L_1-norm is considered.

In the following, we use k-medoids clustering technique to cluster the data-points, which is shown in Figure 2.

***Figure 2:** Kernel PCA scatter plots of data respect to k-means clustering*



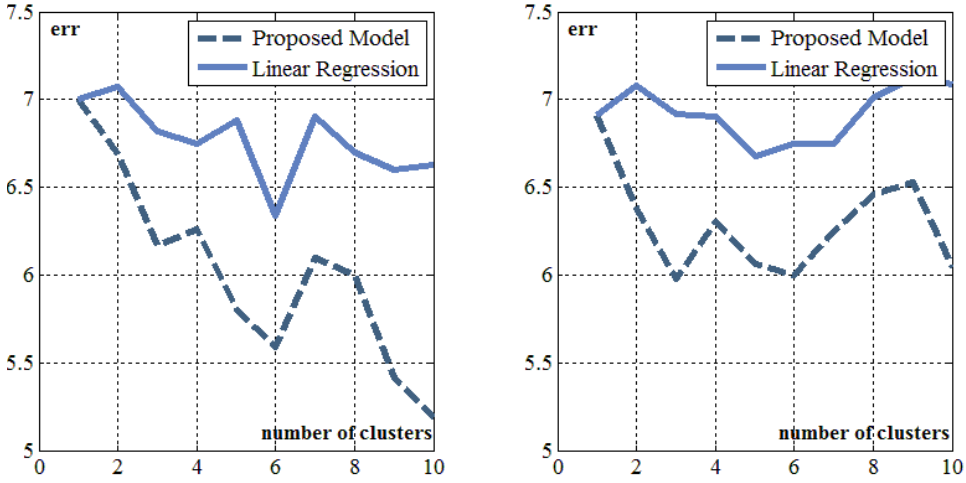
The illustration of kernel PCA (Schölkopf et al., 1999) scatter plots shows the different clusters by choosing different level of k-medoids (k=5, 10 and 15).

6.4 Make a Forecast

In this part, we estimate the optimal parameters by data, namely the model is fitted by given data. In this regard, we compare common linear regression with our proposed algorithm. As input factors, article name, color name, price and size of photo were taken under converting procedure mentioned in Section 6.1. To fit the both methods, common linear regression under L_2-norm considering only main effects (without any interaction term) is considered. We claim that by fitting the data considering clustering techniques there is chance to improve the accuracy of the forecast.

The first preliminary results were performed by 5-fold cross validation and the error is measured by mean absolute percentage error (MAPE). In Figure 3 the corresponding results are demonstrated. Note that, the experiments were performed by using the same training and testing data for both methods, therefore the results of common linear regression vary. To cluster the data two different approaches, namely, "k-medoids" and "k-means mixed" are taken into account. For setting in "k-means mixed" the trade-off parameter (Ψ) is chosen equal to 10. In fact, by choosing proper number of clustering there exists a significant improvement. Thus, for this case splitting the data by adequate clustering technique can improve the accuracy of prediction, however, the future challenge is to identify the proper number of clusters given data. For instance, this can

Figure 3: comparison between proposed algorithm and linear regression



The results for approach in Section 5 are showed by dash, whereas the results for simple linear regression are presented by line. On the X-axis stands the number of clusters. On the left side the results regarding k-medoids are presented, while on the right side the results of k-means mixed.

be done under nested cross validation, i.e., the number of cluster can be chosen by supervised learning.

Assuming null hypothesis (H_0) as "there is no difference between proposed algorithm and common linear regression" and alternative hypothesis (H_1) when "proposed algorithm is better than linear regression", significant test underlying Wilcoxon test (paired rank test for MAPE) in the case of "k-means mixed" is at significant level of 0.05 for alternative hypothesis (H_1). Similarly significant test underlying Wilcoxon test in the case of "k-medoids" is at significant level of 0.001 for the alternative hypothesis. Both significant tests are valid for the number of clusters bigger than 1.

7 Conclusion and Future Research

In this paper, we applied data analysis techniques in the use of fashion products. We advocate the usefulness of big data techniques for the goal forecasting. Indeed, such techniques can improve the accuracy of prediction drastically. The first preliminary results were presented and confirmed a significant gain. In this article, particularly the experiments have been conducted under ordinary linear regression, however it is quite interesting to investigate deeply in the approach by applying more advanced models, e.g. support vector regression machines equipped by polynomial kernel or RBF kernel. This extension allows us to consider non-linear dependency between attributes, which in the fashion field typically exists, e.g. the interaction between trend colors and size of photo. In addition, it is quite desirable to estimate the number of clusters in advance. Needless to say, a wrong splitting the data simply leads to a very poor performance.

Apart from forecasting, interpretability of predicted model is a crucial point, which is addressed in forthcoming researches.

References

Ahmad, A.; Dey, L. (2007): A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, 63(2):503 – 527.

Bishop, C. M. (2006): Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Choi, T. M.; Hui, C. L.; Ng, S. F.; Yu, Y. (2012): Color trend forecasting of fashionable products with very few historical data, in: Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 42(6):1003–1010.

Choi, T. M.; Hui, C. L.; Yu, Y. (2013): Intelligent Fashion Forecasting Systems: Models and Applications. Springer Publishing Company, Incorporated.

Ehmke, J. F. (2012): Integration of Information and Optimization Models for Routing in City Logistics. Springer Publishing Company, Incorporated.

Frank, C.; Garg, A.; Sztandera, L.; Raheja, A. (2003): Forecasting women's apparel sales using mathematical modeling. International Journal of Clothing Science and Technology, 15(2):107–125.

Frünkranz, J.; Hüllermeier, E. (2011): Preference learning. Künstliche Intelligenz, Springer.

Green, M.; Harrison, P. J. (1973): Fashion forecasting for a mail order company using a bayesian approach. Journal of the Operational Research Society, 24(2):193–205.

Gu, W.; Liu, X. (2010): Computer-assisted color database for trend forecasting. In Computational Intelligence and Software Engineering (CiSE), pages 1–4.

Happiette, M.; Rabenasolo, B.; Boussu, F. (1996): Sales partition for forecasting into textile distribution network. In Systems, Man, and Cybernetics, 1996., IEEE International Conference on, volume 4, pages 2868–2873 vol.4.

Hastie, T.; Tibshirani, R.; Friedman, J. (2001): The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Liu, N.; Ren, S.; Choi, T. M., Hui, C. L.; Ng S. F. (2013): Sales Forecasting for Fashion Retailing Service Industry: A Review, Mathematical Problems in Engineering, vol. 2013, Article ID 738675, 9 pages, 2013. doi:10.1155/2013/738675

Mostard, J.; Teunter, R.; Koster, R. (2011): Forecasting demand for single-period products: A case study in the apparel industry. European Journal of Operational Research, 211(1):139 – 147.

Nenni, M. E.; Giustiniano, L.; Pirolo, L. (2013): Demand forecasting in the fashion industry: a review, International Journal of Engineering Business Management.

Schölkopf, B.; Smola, A.; Müller, K. R. (1999): Kernel principal component analysis. In Advances in kernel methods: Support vector learning, pages 327–352. MIT Press.

Sun, Z. L.; Choi, T. M.; Au, K. F.; Yu. Y. (2008): Sales forecasting using extreme learning machine with applications in fashion retailing. Decision Support Systems, 46(1):411 – 419.

Thomassey, S.; Castelain, J. M. (2002): An automatic textile sales forecast using fuzzy treatment of explanatory variables. Journal of Textile and Apparel, Technology and Management.

Thomassey, S.; Happiette, M.; Castelain, J. M. (2002): A short term forecasting system adapted to textile distribution. pages 1889–1893. IPMU 2002, Annecy, France.

Thomassey, S.; Happiette, M.; Castelain J. M. (2002): Textile items classification for sales forecasting. Proceeding 14th European Simulation Symposium (ESS).

Thomassey, S.; Happiette, M. (2007): A neural clustering and classification system for sales forecasting of new apparel items. Applied Soft Computing, 7(4):1177 – 1187, 2007. Soft Computing for Time Series Prediction.

Tsoumakas G.; Katakis I. (2007): Multi-label classification: An overview. Int J Data Warehousing and Mining, 2007:1–13.

Vroman, P.; Happiette, M.; Rabenasolo, B. (1998): Fuzzy adaptation of the holt-winter model for textile sales-forecasting. Journal of the Textile Institute, 89(1):78–89.

Xia, M.; Zhang, Y.; Weng, L.; Ye, X. (2012): Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs. Knowledge-Based Systems, 36(0):253 – 259.

Yelland, P. M.; Dong X. (2014): Forecasting demand for fashion goods: a hierarchical bayesian approach. In Tsan-Ming Choi, Chi-Leung Hui, and Yong Yu, editors, Intelligent Fashion Forecasting Systems: Models and Applications, pages 71–94. Springer Berlin Heidelberg.

Yesil, E.; Kaya, M.; Siradag, S. (2012): Fuzzy forecast combiner design for fast fashion demand forecasting. In Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on, pages 1–5.

Yu, E.; Choi, T. M.; Hui C. L. (2011): An intelligent fast sales forecasting model for fashion products. Expert Systems with Applications, 38(6):7373 – 7379.

Zhu, J.; Hastie, T. (2001): Kernel logistic regression and the import vector machine. In Journal of Computational and Graphical Statistics, pages 1081–1088. MIT Press.