Decision Analytics
a SpringerOpen Journal

RESEARCH Open Access

Towards decision analytics in product portfolio management

Sjors Otten[1], Marco Spruit[1*] and Remko Helms[1,2]

* Correspondence: m.r.spruit@uu.nl
[1]Department of Information and Computer Sciences, Utrecht University, Princetonplein 5, Utrecht, The Netherlands
Full list of author information is available at the end of the article

Abstract

An important strategic decision within the food industry is to achieve the optimal product portfolio allowing an organization to support its customers in the best possible way. Numerous models exist to categorize products based on their relative metrics like revenue, volume, margin or constructed scores. In order to make a more profound decision whether to keep or to remove a product from the portfolio a closer look into product interdependencies is desirable. Hence, by exploring existing DM-techniques through literature and evaluating those DM-techniques that seem suited in a PPM-context by applying each to a dataset, we aim to identify those techniques that complement a Product Portfolio Management process in the food industry. Three DM-techniques were selected: Dependency Modeling, Change and Deviation Detection, and Classification. Of these three techniques, two were found to be of complementary value in a PPM-context, Dependency modeling and Classification, respectively. Change and deviation detection was found to be of no complementary value in a PPM-context due to it forecasting future data points based on historical data, which results in future data points never exceeding the maximum historical data points. However, change and deviation detection could be of complementary value in another context. Finally, we propose an algorithm to visualize the data-driven product classifications in a standard portfolio matrix which portfolio managers can intuitively understand.

Keywords: Data mining; Food industry; Product portfolio management; Association rule mining; Dependency modeling; Change and deviation detection; Seasonal time series forecasting; Classification; Business intelligence

Background

An important strategic decision within food related industries is to achieve the optimal product portfolio allowing an organization to support its customers in the best possible way. Product Portfolio Management (PPM) is a "*dynamic decision process, whereby a business's list of active (new) products (and R&D) projects is constantly updated and revised*" (Cooper, Edgett, & Kleinschmidt, 2001). When reflecting such a definition on food related industries the question to be asked by senior management is: "what products do we keep in the product portfolio and which do we remove from the product portfolio?"

Automating PPM based on metrics like revenue, volume and margin, by means of Business Intelligence (BI), is easy and gives senior management a detailed overview, but it is only based on metrics per product. In order to make a more profound decision whether to keep or to remove a product from the portfolio a closer look into product interdependencies is desirable. Hence, the aim of this research is to find patterns by using Data Mining (DM)-techniques (Agrawal, Imieliski, & Swami, 1993; Kotsiantis & Kanellopoulos, 2006). The main research question can be formulated as follows:

> *"To what extent can Business Intelligence technologies such as Data Mining techniques contribute to Product Portfolio Management in the food industry?"*

By selecting DM-techniques based on literature and evaluating those selected DM-techniques by applying them to a dataset we aim to identify those techniques that complement a Product Portfolio Management process within a food manufacturing organization. The remainder of this paper is structured as follows: section 2 provides a detailed overview of related literature regarding PPM and DM. Section 3 comprises the used material and method. Section 4 presents a newly developed conceptual classification algorithm. Section 5 comprises the results and section 6 concludes with a conclusion and discussion.

Related literature

To provide a clear overview and understanding of the two main concepts on which this research is based, PPM and DM respectively, we explore the available scientific literature.

Product portfolio management

Effective portfolio management is vital to successful product innovation and product development. Portfolio management is about making strategic choices - which markets, products and technologies an organization will invest in (Cooper, Edgett, & Kleinschmidt, 2001). PPM is part thereof, and focuses especially on which products a company should include in its portfolio in order to get a maximum return on investment and customer satisfaction. Cooper et al. (2001) define PPM as "*a dynamic decisions process, whereby an organization's list of active (new) products (and R&D) projects are constantly updated and revised*". Bekkers, Weerd, Spruit and Brinkkemper (2010) state that PPM takes a central role in evaluating which active (new) product (project) should be updated/kept or should be declined/removed based on certain assessment criteria/tools.

According to Cooper, Edgett and Kleinschmidt (2002) PPM has four main goals within an organization: (1) Value maximization, (2) Balance, (3) Strategic direction, and (4) right number of projects (products). With *value maximization* the goal is to allocate resources in order to maximize the value of your portfolio. *Balance* is meant to develop a balanced portfolio - to achieve a desired balance of projects (products) in terms of a number of parameters; for example high risk versus low risk and across various markets, and product categories. *Strategic direction* is intended to ensure that, regardless of all other considerations, the final portfolio of projects (products) truly reflects the business strategy. *Right number of projects (products)* ensures that an organization takes into consideration that there are enough resources available to ensure that there is no project queue or pipeline gridlock when it comes to the production/development of new projects (products).

By reviewing the goals of PPM a global message can be derived. The message states that the selection of the right project (product) is of the upmost importance within organizations. Selecting the right projects (products) also implies that there are great negative consequences for an organization when wrong projects (products) are selected (Cooper & Kleinschmidt, 1995). Such consequences are missed opportunities, missed revenues, and loss of competitive advantage.

In order to achieve/assess the progress of the four main goals a variety of methods exists. The first goal *value maximization* relates to financial value maximization. Financial affairs dominate portfolio management. Methods used are, for example, Expected Commercial Value (ECV), Productivity Index (PI) and Return on Investment (ROI) (Cooper, Edgett, & Kleinschmidt, 2001; Dickinson, Thornton, & Graves, 2001). With regards to the second goal, *balance,* a tool which is commonly used is a "scoring model". A scoring model comprises a series of factors and quantitative questions per factor and is popular among decision makers (Archer & Ghasemzadeh, 1999). A viable alternative for determining the product portfolio (product assortment) is the use of a data mining approach, called Association Rule Mining (ARM), which exposes the interrelationships between products by inspecting a transactional dataset (Brijs, Swinnen, Vanhoof, & Wets, 1999; 2004; Fayyad, Piatetsky-Shapiro, & Smyth 1996a, 1996b, 1996c.

With regards to the third goal, *strategic direction,* two main issues are in play when an organization desires to achieve strategic alignment with its product portfolio, strategic fit and spending breakdown respectively. A method for possibly coping with and achieving alignment between strategy and portfolio is called the "strategic bucket method". A strategic bucket can be defined as *"money set aside for new product development aligned with a particular strategy"* (Chao & Kavadias, 2007). For the fourth goal, *the right number of projects,* it is all about allocating the right resources to the right projects. This can be achieved by conducting a resource capacity analysis. . It quantifies the product's demand for resources versus the availability of them (Cooper R. G., 1999; Cooper, Edgett, & Kleinschmidt, 2000). In order to visualize the data obtained by the utilized method(s) per PPM-goal so-called matrices were developed. The first developed, most widely accepted and implemented visualization is the Boston Consultancy Group (BCG) matrix (Hambrick, MacMillan, & Day, 1982).

Data mining

Across a wide variety of fields, including marketing and healthcare, data are increasingly being collected and accumulated at a dramatic pace (e.g. (Pachidi, Spruit, & Weerd 2014); (Spruit, Vroon, & Batenburg, 2014)). Therefore, an urgent need for theories and tools exists to assist in extracting useful information (knowledge) from these rapidly growing volumes of digital data (Fayyad et al. 1996a, 1996b, 1996c). Piatetsky-Shapiro and Frawley (1991) define DM as "*the process of nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.* Vleugel, Spruit, and Daal (2010) notably provide a methodological recipe to help select the appropriate DM techniques given the properties of the analytical task at hand. The various purposes of DM can be broadly divided into six categories: (1) *Classification*; (2) *Regression;* (3) *Clustering;* (4) *Summarization*; (5) *Dependency modeling*; (6) *Change and deviation detection* (Fayyad, Piatetsky-Shapiro, & Smyth 1996a, 1996b, 1996c).

Classification

Classification is a function that maps (classifies) a data item into one of several predefined classes (Weiss & Kulikowski, 1991). It has been studied substantially in statistics, machine learning, neural networks, and expert systems. Classification searches for common properties among a set of objects (data) in a database and classifies them into different *classes*, which is done according to a pre-defined classification model (Chen, Han, & Yu, 1996). To build a classification model, a sample database E is treated as the *training set*, in which each tuple consist of the same set of multiple attributes as the tuples in a large database W, and additionally, each tuple has a known class identity (label) associated with it (Chen, Han, & Yu, 1996). Classification can be very simple with a statement like *"IF $x = 1$ AND $y = 6$ THEN $z = 2$ ELSE $z = 3$"* or one can define a very advanced classification model based on neural networks (Lippmann, 1988), genetic algorithms (Goldberg, 1989), and decision trees (Quinlan, 1993). In terms of ability, decision trees are a rapid and effective method for classifying data set entries, and can provide good decision support capabilities (Aitkenhead, 2008).

Regression

Regression is a function that maps a data item to a real-valued prediction variable (Fayyad, Piatetsky-Shapiro, & Smyth 1996a, 1996b, 1996c). Regression is a statistical method with a predictive purpose (Larose, 2005). The main sub regression methods are: (1) *Linear regression*, (2) *multiple linear regressions*, and (3) *non-linear regression*.

Linear regression involves a response variable Y and a single predictor variable$X$. In order to determine the linear regression from Y on X the formula $Y = b_{xy}X + a_{xy}$ is used to construct the continuous-valued function. Where b_{xy} is the coefficient and a_{xy} is the intercept.

Multiple linear regression involves a response variable Y and more than one predictor variable. It addresses questions about the relationship between a set of independent variables and a dependent variable. The general equation is written as $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p + e$. Each of the X's is an independent variable, each of the b's is the corresponding regression coefficient, and $e$is the error in prediction for each case.

Non-linear regression is based on the same basic idea as linear regression, namely to relate a response Y to a vector of predictor variables $X = (X_1, \ldots, X_k)^{\mathrm{T}}$. Nonlinear regression is characterized by the fact that the prediction equation depends nonlinearly on one or more unknown parameters. Nonlinear regression usually arises when there are physical reasons for believing that the relationship between the response and the predictors follows a particular functional form (Smyth, 2002).

Clustering

Clustering is the division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar to one another and dissimilar to objects of other groups (Berkhin, 2006). Clustering data in a few clusters results in a loss of certain fine details, but achieves simplification. Clustering can be viewed as a density estimation problem, which is the subject of traditional multivariate statistical estimation

(Scott, 1992). This section elaborates on clustering in a data mining context, which is characterized by large datasets with many attributes of different types (Berkhin, 2006) This section elaborates on both the hierarchical- and partitioning-approach. *Hierarchical clustering* builds a cluster hierarchy or a tree of clusters, also known as a *dendrogram*. The dendrogram has a high degree of similarity with so-called decision trees. Every cluster node contains child clusters, which allows exploring data on different levels of granularity (*agglomerative, divisive)* (Jain & Dubes, 1988). Advantages of hierarchical clustering are: flexibility with regards to the level of granularity, ease of handling form of similarity or distance, and its applicability to any attribute type. However, there are some disadvantages such as vagueness of termination criteria and most hierarchical algorithms do not revisit clusters once constructed. The basics of hierarchical clustering include Lance-Williams formula, the idea of conceptual clustering, classic algorithms like SLINK, COBWEB, and newer algorithms such as CURE and CHAMELEON (Sibson, 1973; Fisher, 1987; Guba, Rastogi, & Shim, 1998; Karypis, Han, & Kumar, 1999).

Partitioning clustering algorithms obtain a single partition of the data instead of a clustering structure, such as the dendrogram, produced by the hierarchical approach (Jain, Murty, & Flynn, 1999). Partitioning clustering algorithms are subdivided in two categories. The first is called *partitioning relocation clustering* and is further classified into *probabilistic* clustering, *k-medoids* methods, and *k-means* methods (Park, 2009; Larose, 2005). Such methods concentrate on how well points fit into their clusters. The second class of algorithms is called *density-based partitioning* and attempts to discover dense connected components of data, which are flexible. Density-based connectivity is used in the algorithms DBSCAN, OPTICS, DBCLASD (Ester, Kriegel, Sander, & Xu, 1996; Ankerst, Breuning, Kriegel, & Sander, 1999; Xu, Ester, Kriegel, & Sander, 1998).

Summarization

The ability to summarize data provides an important method for getting a grasp on the meaning of the content of a large collection of data. It enables humans to understand their environment in a manner amenable to future useful manipulation (Yager, 1982) In the context of data mining the most appropriate definition of summarization is *"grasp and briefly describe trends and characteristics appearing in a dataset, without doing (explicit) manual 'record-by-record' analysis"* (Niewiadomski, 2008).

Two main approaches exist to summarize a dataset, numerical summarization (NS) and linguistic summarization (LS) (Wu, Mendel, & Joo, 2010). NS was the first summarization technique to be used. NS is concerned with e.g., summarizing statistics, exemplified by the average, median, minimum, maximum, α-percentile (Kacprzyk & Zadronzy, 2005). LS provide easier understandable summarized information than NS. The output it generates is of the form *"about half of the sales in summer is of product group accessories"* (Kacprzyk & Zadronzy, 2005). Over the past years a number of approaches were introduced for summarizing data in a database and can be categorized in four classes: (1) methods based on unary operators, (2) approaches related to multidimensional databases, (3) methods based on statistic and symbolic techniques and (4) approaches based on fuzzy set theory (Triki, Pollet, & Ben Ahmed, 2008).

Yager's approach for linguistic summaries using fuzzy logic with linguistic quantifiers consists of:

- V is a quality (attribute) of interest (i.e. salary in a database of workers)
- $Y = \{y_1, .., y_n\}$ is a set of objects (records) that manifest quality V (e.g. the set of workers; hence $V(yi)$ are values of quality for object y_i
- $D = \{V(yi), \ldots, V(yn)\}$ is a set of data (the "database" on question)

Basically, given a set of data D, a hypothesis can be constructed which states that any appropriate summarizer S and any quantity in agreement Q, and the assumed measure of truth Twill indicate the truth of the statement that Q data items statisfy the statement (summarizer) S. For more information on the actual mathematics and more examples see the work of Yager (1982).

Dependency modeling

Dependency modeling consists of finding a model that describes significant dependencies between variables (Fayyad, Piatetsky-Shapiro, & Smyth 1996a, 1996b, 1996c). In the context of this research it is the aim to find dependencies between transactions in large datasets. Therefore, this section will elaborate on a DM-technique called *Association Rule Mining* (ARM) (Agrawal, Imieliski, & Swami, 1993).

The task of ARM is to determine and derive a set of association (dependency) rules in the format "$X_1\hat{}\ldots\hat{}X_m$ ➔ $Y_1\hat{}\ \ldots\hat{}\ Y_n$ "where X_i (*for i* € {1, …, *m*}) and Y_j (*for j* € {1, …, *n*}) are sets of attribute-values, from the relevant data sets in a database (Kotsiantis & Kanellopoulos, 2006).

An example of ARM is *market basket analysis* and comprises the analysis of customer buying habits by finding associations between the different items that customers place in their "basket" (Han, Kamber, & Pei, 2011). Given the association rule:

$$\textit{Computer} \Rightarrow \textit{financial management software}\ [\textit{support} = 2\%, \textit{confidence} = 60\%] \quad (2.2)$$

Two measures manifested, namely, *support* and *confidence.* They respectively reflect the usefulness and certainty of discovered association rules. Given equation 2.2 with a *support* of 2 % means that 2 % of all the transactions show that computer and financial management software are purchased together. A *confidence* of 60 % means that 60 % of the customers who purchased a computer also bought the financial management software. Association rules are considered interesting if they satisfy both a *minimum support threshold* and a *minimum confidence threshold,* which are determined manually by users or domain experts (Agrawal, Imieliski, & Swami, 1993).

A high value of confidence suggests a strong association rule. However this can be deceptive because if X and/or Y have a high support, we can have a high value for confidence even when they are independent. Hence, the metric *lift* was introduced. *Lift* quantifies the relationship between X and Y (Ordonez & Omiecinski, 1999). In general, a lift value greater than 1 provides strong evidence that X and Y depend on each other. A lift value below 1 states that X depends on the absence of Y or vice versa. A lift value close to 1 indicates X and Y are independent (Ordonez, 2006).

A Commonly used algorithm for mining association rules is APRIORI (Agrawal & Srikant, 1994). APRIORI is an algorithm for mining frequent itemsets for Boolean association rules. It employs an iterative approach knowas level-wise search, where *k*-itemsets are used to explore $(k + 1)$-itemsets. A second well known algorithm is the FP-Growth-algorithm for frequent item set mining (Han, Pei, & Yin, 2000). The basic idea of the FP-Growth algorithm can be described as a *recursive elimination* scheme (Borgelt, 2005).

Change and deviation detection

Change and deviation detection focuses on discovering the most significant changes in the data from previously measured or normative values (Fayyad, Piatetsky-Shapiro, & Smyth 1996a, 1996b, 1996c). Change and deviation analysis describes and models regularities or trends for objects whose behavior changes over a period of time. Distinctive features of such analysis are time-series, and sequence or periodically pattern matching (Han, Kamber, & Pei, 2011).

Time-series is in fact a sequence of data points, measured typically at successive time, spaced at uniform time intervals. A sub discipline of time-series forecasting deals with the seasonality component and is called *seasonal time series forecasting* (Zhang & Qi, 2005). *Seasonality* is a periodic and recurrent pattern caused by factors such as weather, holidays, repeating promotions, as well as the behavior of economic agents (Hylleberg, 1992). Accurate forecasting of seasonal and trend time-series is very important for effective decisions in retail, marketing, production, inventory control and other business sectors (Lawrence, Edmundson, & O'Connor, 1985; Makridakis & Wheelwright, 1987). A well-known algorithm for time-series modeling is the autoregressive integrated moving average (ARIMA) and is the most widely used to forecast future events/figures. The major limitation of ARIMA is the pre-assumed linear form of the model (Zhang P. G., 2003). Furthermore, forecasting accuracy is known to be problematic in cases of high demand uncertainty and high variance within demand patterns (Maaß, Spruit, & Waal 2014).

Based on the limitations of the ARIMA models with regards to the lacking capability of capturing nonlinear patterns a need exists for an approach that can cope with the aforementioned limitations. Recently Artificial Neural Networks (ANN) received much attention (Hill, O'Connor, & Remus, 1996; Balkin & Ord, 2000). The major advantage of ANN compared to ARIMA models is their flexible nonlinear modeling capability (Zhang P. G., 2003). Additionally, no specific assumptions need to be made about the model and the underlying relationship is determined through data mining (Zhang & Qi, 2005).

Pattern matching in a data mining context, is the act of checking some sequence of tokens for the presence of constituents of same pattern. In contrast to pattern recognition, the match usually has to be exact. The patterns generally have the form of sequences or tree structures (Zhu & Takaoka, 1989). Several pattern matching algorithms where developed over time such as the KPM algorithm, BM algorithm and RK algorithm (Zhu & Takaoka, 1989).

Selected data mining techniques

Three data mining techniques appear to be the most suited, classification, dependency modeling, and change and deviation detection respectively. Reflecting on PPM a useful tool is to classify products in quadrants (i.e. BCG-matrix, DPM-matrix) which relates to the idea of classification.

The second option, dependency modeling, and in particular association rule mining, is a viable option for discovering associations between products in large transactional datasets and has been proven useful for retailers when deciding on promotions and shelve arrangements. A third possible category is change and deviation detection, in particular seasonal time series forecasting.

Material and methods

Material

The material used for evaluating the selected data mining techniques is a transactional dataset comprising 12 months of sales statistics. The transactional dataset was provided by a case study organization which produces poultry for retailers in the Benelux.

The dataset (fact table) used for determining the applicability of the selected data mining techniques comprises 7186349 records of transactional data generated between 2011-01-01 and 2011-12-31. The dataset is stored in a MSSQL-database. Table 1 provides an overview of the most significant attributes within the dataset. The column "type" can obtain the values "*PK*", "*FK*", "*M*" which stands for *PrimaryKey*, *ForeignKey*, and *Metric* respectively.

The attribute with type "FK" are foreign keys, which are related to primary keys in each dimension table. The dimension tables are at our disposal and can be used for possible data understanding and data preparation. The dimension tables are "Customer_Base", "Article_Base", "Article_ProductGroup", "Article_MainProductGroup".

The data quality of the initial dataset is found to be satisfactory with regards to data types and completeness. No inconsistencies, errors, missing values or noise was found in the dataset. However, three issues require attention:

1. The presence of articles which have multiple article numbers (i.e. 303125 and 15303125).

Table 1 Dataset – attributes

Attribute	Data type	Description	Type (PK/FK/M)
FA076_LS_DATUM	Datetime (yyyy-mm-dd)	Date of actual transaction	PK
FA076_BS_NR	Integer	Transaction number	PK
FA076_ADR_NR	Integer	Customer number	FK
FA076_ART_NR	Integer	Article number	PK/FK
FA076_WG	Integer	Article product group number	FK
FA076_OWG	Integer	Article main product group number	FK
FA076_KG_MENGE	Decimal	Sales volume in KG	M
FA076_BS_MENGE	INT	Ordered items	M
FA076_PE_MENGE	INT	Sales volume in items	M
FA076_LI_MENGE	INT	Delivered items	M
FA076_UPD_DATUM	Datetime (yyyy-mm-dd)	The actual date when a record has been updated by the system	M
FA076_REC_NR	INT	Auto-increment number per record per order	PK

2. The presence of main product groups which are not of concern to the product portfolio such as "crates", "packaging", "labels", "diverse" and "clothing".
3. The presence of product groups with large fluctuations in sales volume (unit or KG) per periodic interval.

Table 2 provides the scripts used for resolving the issues mentioned above.

Method

For evaluation of the selected data mining techniques with regards to their applicability in a PPM-context the CRISP-DM framework is used (Chapman, et al., 2000). It comprises six stages which are described in Table 3.

Per data mining technique the stages data preparation, modeling, and evaluation are elaborated on. The stage Deployment is elaborated on in the section "Deployment". Note that we omit the CRISP-DM-stage "data understanding" per data mining technique due to the fact that we described and explored the initial dataset in the previous section.

Calculation

Based on the literature survey, it was found that a common visualization tool used in a PPM-context is a portfolio matrix. A possibility for automatically placing products in the product portfolio in a certain quadrant of a matrix is classification. After exploring various classification techniques (i.e. classification rules and decision trees) it was found that none of the algorithms were able to plot and visualize the classified products on a portfolio matrix. In addition, none of the algorithms are able to combine outcomes of data mining techniques (ARM) and explicit available metrics. Hence, we propose the following algorithm for classifying, plotting and visualizing products in the product portfolio:

```
$ForEach\ Article\ (ItemX) in\ (X \rightarrow Y)$
        $Derive\ top10\ AssocRule\ based\ on\ metric\ LIFT$
          $Derive\ MAX(LIFT), MIN(LIFT), AVG(LIFT)$
$ForEach\ AssocRule\ derive\ ItemY\ in\ (X \rightarrow Y)$
        $ForEach\ derived\ ItemY$
          $Derive\ MAX(SalesVolumeInKG), MIN(SalesVolumeInKG), AVG(SalesVolumeInKG)$
$Construct\ ProductMatrix\ WHERE$
          $X - axis = MAX, AVG, MIN(LIFT)$
          $Y - axis = MAX, AVG, MIN(SalesVolumeInKG)$
$ForEach\ ItemY \rightarrow PLOT(ItemY)\ on\ ProductMatrix(x, y)\ WHERE\ (x, y) = (LIFT, SalesVolumeInKG)$
$COUNT(PLOT(ItemY)) per\ Quadrant$
$SELECT\ MAX\big(COUNT((PLOT(ItemY))\big) GROUP\ BY\ Quadrant$
$PLOT(ItemX)\ WHERE$
$COUNT(PLOT(ItemY)) GROUP\ BY\ Quadrant = MAX$
```

The proposed algorithm is used for classifying, plotting, and visualizing the strength of the association between X and$Y$; based on Lift, and the SalesVolumeInKG of Y. This aid in determining whether or not X should be kept in the product portfolio or be removed from it. In Table 4 the parameters for the algorithm are presented.

Results

This section presents the results of the data mining technique evaluation of dependency modeling, change and deviation detection, and classification respectively.

Table 2 SQL-functions - scripts

Issue	Script

Issue 1	select CONVERT(int,RIGHT((CONVERT(varchar,<article_nr>))),6))
	FROM < table_name>
	WHERE len(<article_nr>) = 6 and
	FA076_OWG = 1

Issue 2	SELECT * from < table_name> WHERE FA076_OWG <> < mainproductgroup_nr > and FA076_OWG <> < mainproductgroup_nr > and .... FA076_OWG <> < mainproductgroup_nr>

Issue 3	SELECT < productgroup_nr>, DATEPART(YEAR, FA076_LS_DATUM),
	DATEPART(MONTH,FA076_LS_DATUM), SUM(FA076_KG_MENGE)
	FROM < table_name>
	WHERE FA076_WG = <productgroup_nr>
	GROUP BY productgroup_nr, DATEPART(YEAR, FA076_LS_DATUM),
	DATEPART(MONTH, FA076_LS_DATUM)
	ORDER BY productgroup_nr, DATEPART(YEAR, FA076_LS_DATUM),
	DATEPART(MONTH, FA076_LS_DATUM)

Dependency modeling

With regards to dependency modeling, a sub technique in the context of this research is selected. This sub technique is called Association Rule Mining (ARM) and seeks to identify associations between articles among transactions.

Data preparation

In order to perform ARM on a dataset it is inevitable that data in the initial dataset is selected, cleansed, restructured and formatted. ARM requires a dataset with the structure and format such as presented in Table 5. Where "*TransactionID*" represents the unique transaction in the dataset and "*In*" presents a unique article in the dataset which can obtain the values 1 and 0. Where 1 indicates that the article's *presence* is true; whereas Zero indicates that the article's *presence* is false for a specific transaction.

Table 3 CRISP-DM framework - stages

Stage	Description
Business understanding phase	Focuses on understanding the project objectives and requirements from a business perspective, converting this knowledge into a data mining problem definition and a preliminary plan is designed to achieve the objectives (Chapman, et al., 2000)
Data understanding phase	The initial data collection and proceeds with activities to get familiar with the data, discover quality problems and insight (Chapman, et al., 2000)
Data preparation phase	Covers all activities needed to construct the final data set (Chapman, et al., 2000).
Modeling phase	The selection and application of modeling techniques. Typically, several techniques are available for the same data mining problem type. Some techniques have specific requirements regarding the structure of the data. Therefore, going back to the data preparation phase is often necessary (Chapman, et al., 2000).
Evaluation phase	The evaluation of the created data mining models, before actually deploying them. By thoroughly evaluating and reviewing the steps used to create the model it is determined if the model achieves the business objectives. If not, a new iteration of the modeling phase is started (Chapman, et al., 2000).
Deployment phase	The implementation of the results found by the data mining model within the supplied data set via tools such as reports or dashboards (Chapman, et al., 2000)

Table 4 Classification data mining - parameters

Parameter	Description
MAX(LIFT)	The maximum derived LIFT value from top 10 AssocRule dataset
MIN(LIFT)	The minimum derived LIFT value from top 10 AssocRule dataset
MAX(SalesVolumeInKG)	The maximum derived SalesVolumeKG value from the Y-items in the top 10 AssocRule dataset
MIN(SalesVolumeInKG)	The minimum derived SalesVolumeKG value from the Y-items in the top 10 AssocRule dataset

Modeling

With regards to ARM we have chosen the APRIORI-algorithm for deriving association rules by means of the arules-package (Hahsler, Grun, & Hornik, 2005). The model depends on two main parameters as presented in Table 6.

For the analysis we set the parameters *Min_support* = 0.1 (10 %) and a *Min_confidence* = 0.1 (10 %). In order to derive any meaning with regards to the usefulness of an association rule, the metrics *support* and *confidence* do not suffice. Hence, the metric *lift* is used to determine the usefulness of an association rule.

Evaluation

In order to evaluate the correctness, usefulness and best approach of the selected data mining technique we apply the constructed model on 3 variations in the dataset:

- Analysis #1; we analyze the whole target dataset (containing all 12 product groups belonging to the main product group 1);
- Analysis #2; we analyze each product group over a whole year;
- Analysis #3; we analyze a product which is marked for analysis in the case study due to it performing under 250 kg of sales volume per week.

Analysis #1 – whole dataset

The APRIORI-algorithm (m*in_support* = 0.1; *min_confidence* = 0.1) derived 2526 association rules from the target dataset, presented in Table 7. We found that the majority of the rules is plotted between a support of 0.1 – 0.2, and a confidence of 0.27 – 0.98. Additionally, we found, the larger the support, the lower the confidence for an association rule.

After cross-referencing the items presented in the association rules with the data in the dimension-table "Article_Base", it was found that all items belonged to product group 1 and 2. The cause hereof is the fact that the majority of the records are distributed among product group 1 and a relative large proportion of the records in product

Table 5 ARM - dataset structure

TransactionID	I1	I2	I..	I*n*
1	1	0	…	1
….	…	…	…	…
N	1	1	…	0

Table 6 Association rule mining - parameters

Parameter	Data type	Description
Min_support	decimal (0,2)	*Min_support* means that x% of all the transactions under analysis show that X and Y are purchased together
Min_confidence	decimal (0,2)	*Min_confidence* means that x% of the customers who purchased X also purchased Y.

group 2. Figure 1 depicts the visualization of the top 10 association rules, where the size of each vertex indicates the support value and the color indicates the value of the metric *lift*.

Analysis #2 – dataset(s) per product group

For analysis 2 we generated separate datasets per product group. Each separate dataset comprises 12 months of transactional data. Per dataset (12 in total) we applied the APRIORI-algorithm (min_support = 0.1; min_confidence = 0.1). An excerpt of the analysis output is presented in Table 8.

Analysis #3 – dataset per underperforming article (item)

For the third analysis, we use the following approach:

1. Identify an article (item) which performs under the so-called "threshold-value"
2. Generate a dataset comprising all transactional data in which the identified article is included
3. Apply the APRIORI-algorithm on the generated dataset for deriving association rules
4. Analyze the derived association rules with regards to the identified article.

Step 1 The following underperforming articles were identified (Table 9):

683 articles were identified, which do not meet the threshold value in a given week. However, just merely selecting one of these articles could also cause unreliable results. Therefore, we selected articles which chronologically (periodical interval) over time, did not meet a pre-defined threshold value.

Step 2 Per article (405245, 701024), a dataset is generated comprising all the transactional data of which the specific article is part of. Each dataset is stored in a *. CSV conforming to the structure 5.

Table 7 Dataset results

Dataset	Value
Size	7186349 records
Product groups	1,2,3,4,5,6,7,8,9,10,11,12
Months	1,2,3,4,5,6,7,8,9,10,11,12
Transactions	385100
Min_support	0.1
Min_confidence	0.1

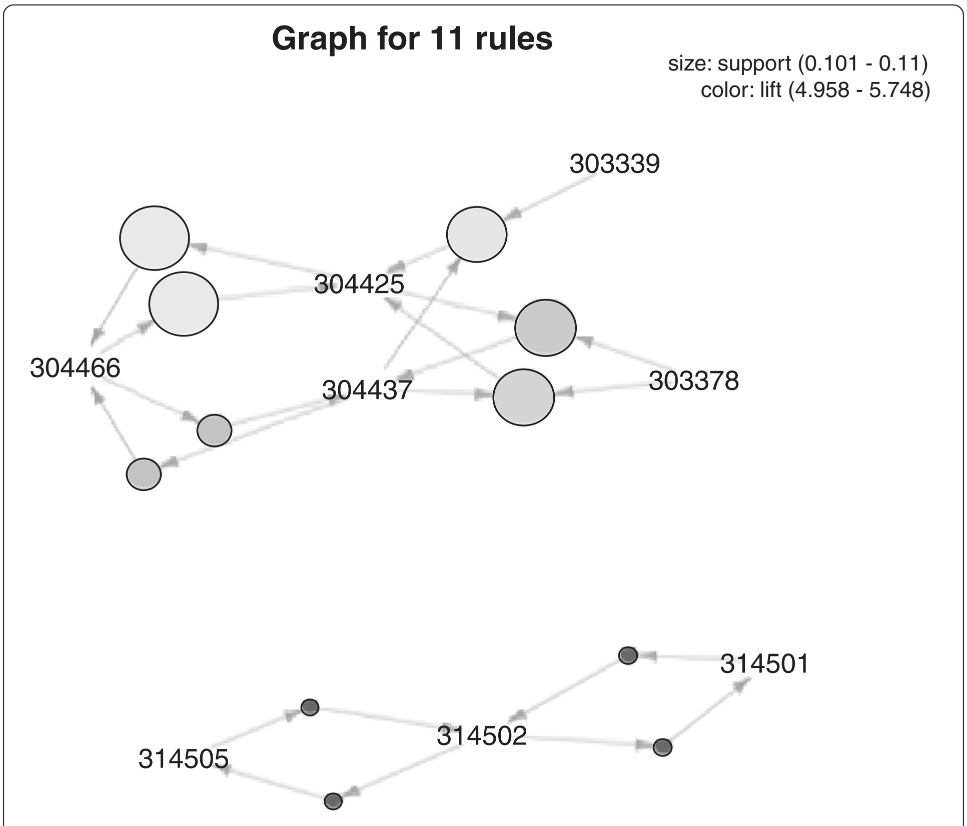


Fig. 1 Visualization - association rules

Step 3 Per article-dataset the APRIORI-algorithm is applied. In Table 10, per article-dataset, the number of association rules is presented, together with the parameter-settings for *min_support* and *min_confidence*. The parameter-settings were left to their default values.

Step 4 We visualized the derived association rules per article dataset by means of the metric *lift*. In order to do so we created a subset of the association rules, which only contained the article under investigation and the highest possible *lift-scores* for 10 rules.

Visualization

Dataset – 405245

The subset is created by selecting all left-hand-side item sets (X) which contain article 405245 and a lift higher than 1. This resulted in a subset of 911 association rules with *support* ranging from 0.1003 – 0.5206, *confidence* ranging from 0.8 – 0.9827, and *lift* ranging from 1.193 – 4.844. In order to reduce the association rules we selected only those association rules with a minimal lift value of 4.062, which is the value of the lift metric at rule 10 when sorting them in ascending order. Figure 2 depicts the visualization of the top 10 association rules. The size of each vertex resembles the size of *support* and the color represents the *lift*.

Dataset – 701024

The subset is created by selecting all left-hand-side item sets (X) which contain article 701024 and a lift higher than 1. This resulted in a subset of 1103 association rules with

Table 8 Analysis #2 - dataset

Dataset – product group 1	Value
Size	1731411
Transactions	361832
#Association rules	92
Support	0.1029 – 0.2760
Confidence	0.2972 – 0.9761
Lift	1.836 – 2.893
Dataset – product group 12	
Size	3158
Transactions	1110
#Association rules	50
Support	0.3306 – 0.5414
Confidence	0.6571 – 0.9980
Lift	1.701 – 2.339

support ranging from 0.1008 – 0.8853, *confidence* ranging from 0.8 – 0.9737, and *lift* ranging from 0.9599 – 4.3651. In order to reduce the association rules we selected only those association rules with a minimal lift value of 3.68, which is the value of the lift metric bat rule 10 sorting them in ascending order. Figure 3 depicts the visualization of the top 10 association rules. The size of each vertex resembles the size of *support* and the color represents the *lift*.

Change and deviation detection

With regards to change and deviation detection, a sub technique in the context of this research is selected. As mentioned in section 2.3.5, this sub technique is called time series analysis and seeks to predict future events based on data points recorded over equally spaced points in time.

Data preparation

In order to perform Time Series Analysis (TSA) on a dataset it is inevitable that data in the initial dataset is selected, cleansed, restructured and formatted. TSA requires a dataset with the structure and format such as presented in Table 11. Where *"TimeStamp"* is the point in time at which the measurement was observed and "*ObservedValue"* represent the observed value at that specific point in time.

Modeling

We have chosen to use the forecast-package, which is freely available (Hyndman & Khandakar, 2007). The forecast package comprises approaches for automatic forecasting such as exponential smoothing and ARIMA-modeling. For our model creation we use the

Table 9 Excerpt of underperforming articles

Underperforming articles			
Year	Week	ItemID	SalesVolumeKG
2010	1	405245	14,690
2010	8	403953	96,410
2010	52	305944	53,565

Table 10 Association rules - Article dataset

Dataset	Value
Dataset – 405245	
Min_support	0.1
Min_confidence	0.8
#ofAssociationRules	21241
Dataset – 403953	
Min_support	0.1
Min_confidence	0.8
#ofAssociationRules	3882
Dataset – 701024	
Min_support	0.1
Min_confidence	0.8
#ofAssociationRules	48248

exponential smoothing approach. The model depends on the parameters presented in Table 12 for forecasting purposes.

The forecast package deals with the computation of the aforementioned parameters. Hence, no further elaboration on the values given to the parameters is provided. The parameters are used for smoothing out the time series dataset properly.

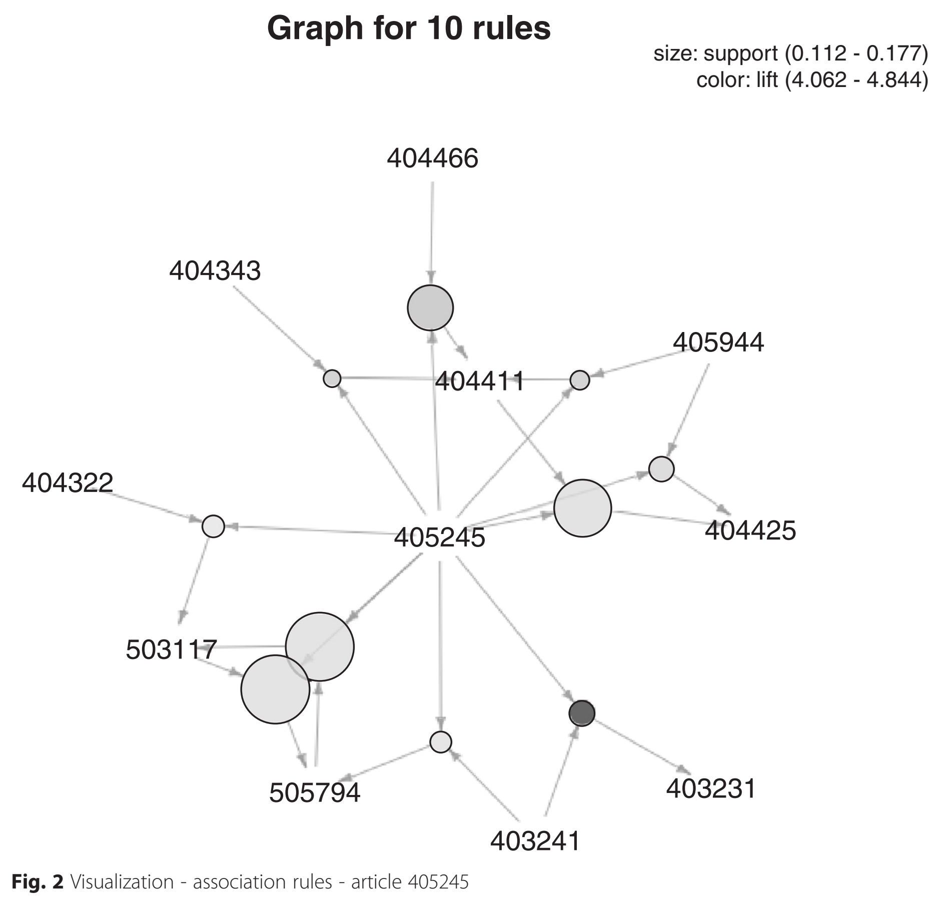


Fig. 2 Visualization - association rules - article 405245

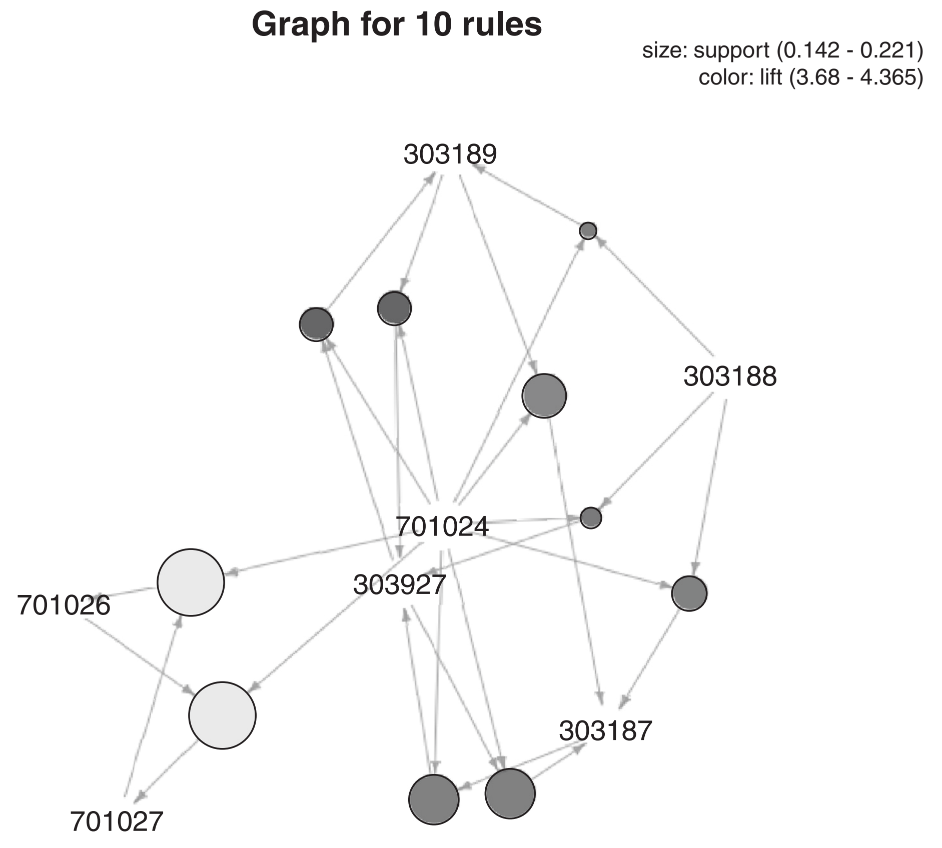


Fig. 3 Visualization - association rules - article 701024

Evaluation

In order to evaluate the correctness, and usefulness of the selected data mining technique the exponential smoothing model is applied on the most detailed level; the article level. By forecasting the SalesVolumeInKG and determining whether, over time, it exceeds the threshold value, we determine whether change and deviation detection is complementary as a data mining technique to the PPM-process.

For analysis purposes we aimed to identify articles which were consequently underperforming, we choose to use two earlier identified article and identify one article that had data for each of the 24 months. Considering the threshold value from the conducted case study, which was 250kg on a weekly basis, the threshold value was scaled up by a factor of 53 and divide by 12, resulting in 1104,167 kg per month. We identified article 605494 as an underperformer with regards to the scaled up threshold value and the presence of 24 months of data.

Table 11 TSA - Dataset structure

TimeStamp	Observed value
yyyy-mm-dd	340,798
N	600,145

Table 12 Time Series Analysis – parameters

Parameter	Data type
Alpha	Decimal (2,4)
beta	Decimal (2,4)
Phi	Decimal (2,4)

Article – 405245

For the application of TSA on article 405245, we extracted SalesVolumeInKG from the extended dataset, grouped per month and year, meeting the criteria of the newly defined threshold value. This resulted in a dataset comprising 9 sequential months in the year 2010. We transformed the derived dataset to a timeseries-object with the parameter *frequency* = 12 because we were dealing with data points observed per monthAfter applying the exponential smoothing-approach it was found that it did not generated adequate results as depicted in Fig. 4.

Article – 403953

For the application of TSA on article 403953, we extracted SalesVolumeInKG from the extended dataset, grouped per month and year, meeting the criteria of the newly defined threshold value. This resulted in a dataset comprising 15 sequential months in the year 2010 and 2011. We transformed the derived dataset to a timeseries-object with the parameter *frequency* = 12 because we were dealing with data points observed per month. After applying the exponential smoothing-approach it was found that it did not generated adequate results as depicted in Fig. 5.

Analysis – 605494

For the application of TSA on article 605494, we extracted SalesVolumeInKG from the extended dataset, grouped per mondht and year, meeting the criteria of the newly defined threshold value. This resulted in a dataset comprising 24 sequential months in years 2010 and 2011. We transformed the derived dataset to a timeseries-object with the parameter *frequency* = 12 because we were dealing with data points observed per month. After applying the exponential smoothing-approach it was found that it generated adequate results as depicted in Fig. 6.

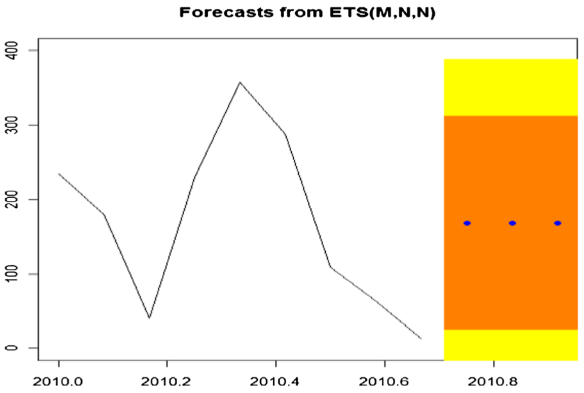


Fig. 4 Visualization - article 405245 - forecast

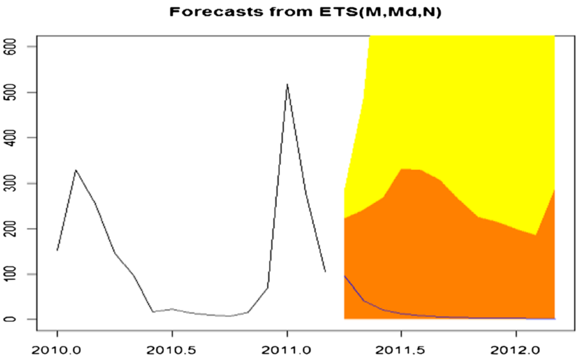


Fig. 5 Visualziation - article 403953 - forecast

Classification

As mentioned earlier we examined the capabilities of existing classification algorithms. However, those algorithms were lacking with regards to visualization of classified data. Hence, we evaluate the earlier presented conceptual BIPPMClass-algorithm.

Data preparation

Classficiation requires a dataset with the structure and format such as presented in Table 13.

Where *"recordID" is* the unique identifier of the record, "PredictorAttrib#" is an attribute by which the classification algorithm seeks to identify to what class a record belongs, and "ClassLabel" is the attribute which indicates to what class a record is classified.

Modeling

With regards to the data mining model for classifying, plotting and visualizing each product on a portfolio matrix we use the earlier proposed algorithm.

Evaluation

For the evaluation of the proposed algorithm we've chosen the top 10 association rules of articles 405245 and 701024. In order to evaluate if the algorithm is useful as a visualization tool we've adopted the BCG-matrix (Hambrick, MacMillan, & Day, 1982). Based on the BCG-matrix we've constructed a product matrix, depicted in Figs. 7 and 8. The product matrix comprises 3 classes, "Keep", "FurtherAnalysis", and "Remove",

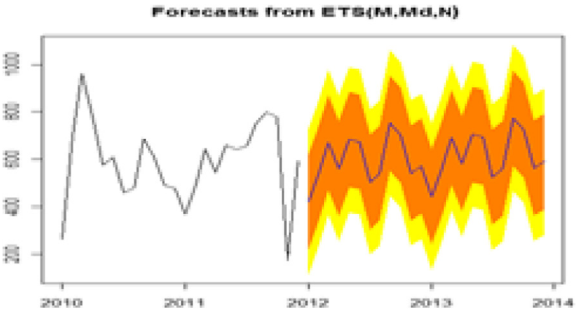


Fig. 6 Visualization - article 605494 - forecast

Table 13 Classification - dataset structure

RecordID	PredictorAttrib1	PredictorAttrib2	PredictorAttrib3	ClassLabel
1	12	41	8	Class1
N	76	8	12	Class2

respectively. The dominant metric is sold kilograms. Therefore, even if there is a strong association with an article (Y), but the SalesVolumeInKG of Y is low, then the article should be classified as “Remove”.

Analysis – 405245

For the analysis of article 405245 we’ve derived the top 10 association rules. Additionally, for all unique item Ys in the top10-dataset, the (maximum) lift and SalesVolumeInKG were retrieved with an MSSQL-query, resulting in the datasets “Lift-values” and “SalesVolumeInKG-values”. By combining the data of datasets “*LIFT-values*” and “*SalesVolumeInKG-values*”, “*Dataset – 405245 – Item Y’s*” resulted. The associated articles were plotted on the product matrix by means of coordinates(x,y) = (Lift,SalesVolumeInKG), as depicted in Fig. 7.

Based on the COUNT-results, it can be concluded that the majority of the items (4 in total) are clustered in Q2 and Q3. Keeping in mind the dominant metric is SalesVolumeInKG it is clear that article 405245 has an above average associations with three products but two of them generate low “SalesVolumeInKG”. Hence, article 405245 can, according to the data and proposed algorithm, be classified in Q2 and therefore removed from the product portfolio.

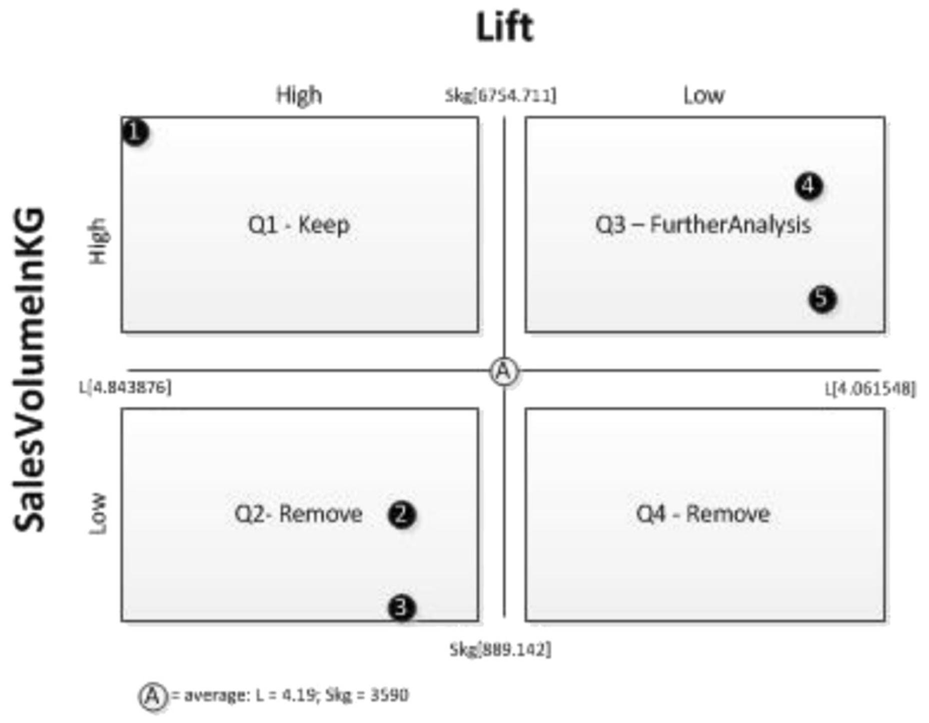


Fig. 7 Portfolio matrix - 405245

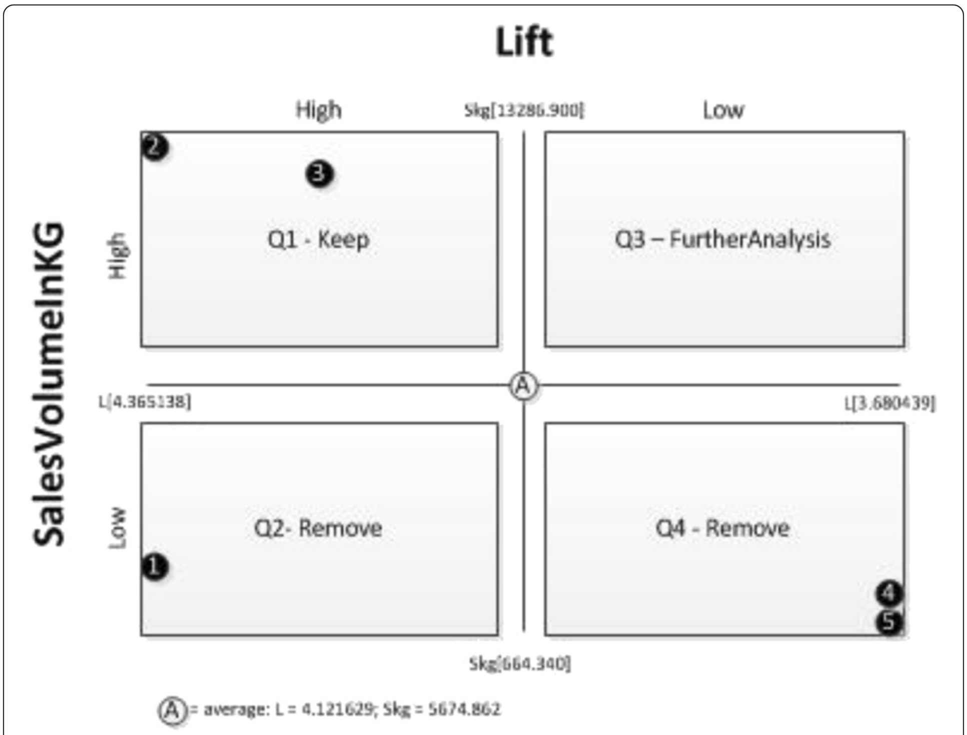


Fig. 8 Product matrix - 701024

Analysis – 701024

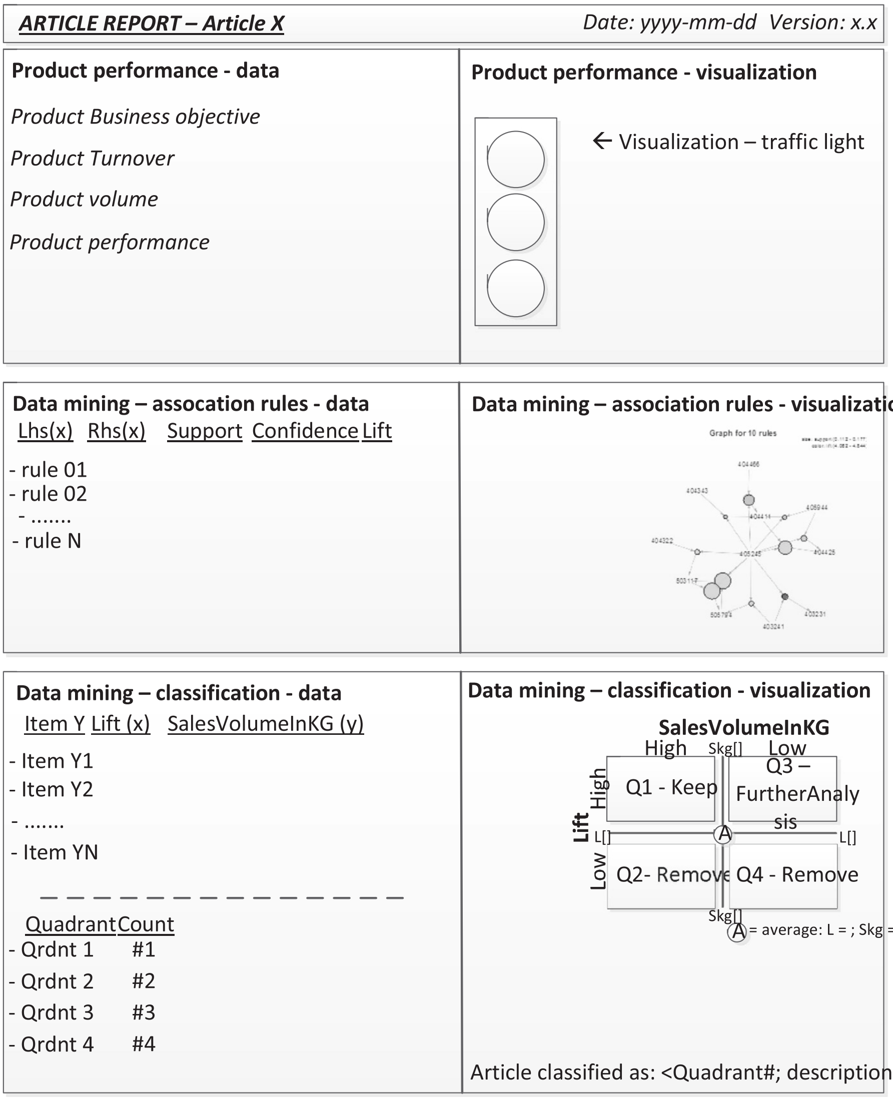
For the analysis of article 701024 we've derived the top 10 association rules. Additionally, for all unique item Ys in the top Ten-dataset, the (maximum) lift and SalesVolumeInKG were retrieved with an MSSQL-query, resulting the datasets "Lift-values" and "Sales VolumeInKG-values" respectively. By combining the data of datasets "*LIFT-values*" and "*SalesVolumeInKG-values*", "*Dataset – 405245 – Item Y's*" resulted. The associated articles were plotted on the product matrix by means of coordinates(x,y) = (Lift,SalesVolume InKG), as depicted in Fig. 8.

Based on the COUNT-results it can be concluded that the majority of the items (4 in total) are clustered in Q1 and Q4. However, due to SalesVolumeInKG being the dominant metric and article 2 and 3 each have a SalesVolumeInKG of 13286.900 and 12006.000, it is appropriate to classify article 701024 in quadrant 1. Hence, article 701024 is kept in the product portfolio.

Deployment

In order to provide the decision makers within a PPM-process with the information generated by the selected data mining techniques we propose a deployment-template (concept) called "ARTICLE REPORT". The ARTICLE REPORT comprises the concepts PRODUCT PERFORMANCE, ANALYSIS RESULTS and PRODUCT MATRIX found in the BIPPM-method. In Fig. 9 the proposed layout of the concept ARTICLE REPORT is depicted.

The ARTICLE REPORT comprises a data and visualization section per integrated concept. On the left hand side the data section is presented to the decision maker with actual statistics. On the right hand side the visualization section is presented to the



ARTICLE REPORT – Article X *Date: yyyy-mm-dd Version: x.x*

Product performance - data

Product Business objective

Product Turnover

Product volume

Product performance

Product performance - visualization

← Visualization – traffic light

Data mining – assocation rules - data

Lhs(x) Rhs(x) Support Confidence Lift

- rule 01
- rule 02
- ....
- rule N

Data mining – association rules - visualization

Graph for 10 rules

Data mining – classification - data

Item Y Lift (x) SalesVolumeInKG (y)

- Item Y1
- Item Y2
- ....
- Item YN

Quadrant	Count
- Qrdnt 1	#1
- Qrdnt 2	#2
- Qrdnt 3	#3
- Qrdnt 4	#4

Data mining – classification - visualization

A = average: L = ; Skg =

Article classified as: <Quadrant#; description>

Fig. 9 Article report

decision maker. The visualization per integrated concept is based on the data residing in the data section of the ARTICLE REPORT.

Discussion

With regards to the scope of this research, the investigation of the applicability of data mining techniques in a product portfolio management context, the selected data mining techniques (classification, dependency modeling, and change and deviation detection) were empirically validated by applying them to a transactional dataset comprising 12 months of sales statistics.

Change and deviation detection, in particular seasonal time series forecasting, was found to be of no added value in a product portfolio management context. The algorithm used, exponential smoothing, forecasts data based on historical data. The forecasted data points did not exceed the defined threshold value due to the fact that the algorithm calculates future data points based on available historical data. Hence, if a product consequently underperforms with regards to a certain criteria, the exponential smoothing-algorithm will forecast future data points of the same magnitude. Thefore, they do not exceed the defined threshold value. However, despite the fact that seasonal time series analysis does not provide useful input in a product portfolio management context, it could prove to be useful in other organizational processes such as procurement or sales.

With regards to dependency modeling and classification, both provide useful input for a product portfolio management. In the context of this research, analysis outcome shows that the best approach is to analyze each product separately. It was also determined that deriving associations on a whole dataset (comprising all articles and product groups) or a constructed dataset per product group yields unacceptable association rules.

Zooming in on classification, it was found that existing algorithms did not contain a visualization step for plotting products on a portfolio matrix. Hence, a new conceptual algorithm for classifying and plotting items on a portfolio matrix was presented. The validation of the algorithm was done by hand. No prototype was developed. However, the results of the application of the algorithm with regards to classification and visualization are promising. However, this algorithm needs to be validated in practice on several datasets comprising transactional data and association rules. In addition, it is desirable to perform more case studies. Furthermore, the scaling of the X- and Y-axis in a (portfolio) matrix is arbitrary. Anyone and everyone are free to determine the definition and thereby scaling of the X- and Y-axis on a matrix. If done improperly, the information and outcome the algorithm generates could be false and misinterpreted, which could lead to erroneous decision making in a product portfolio management context. Hence, other visualization options should be explored also.

Conclusions

For the evaluation of the selected data mining techniques three analysis approaches were executed, (1) whole dataset analysis, (2) separate datasets per identified product group, (3) separate dataset per product up for analysis.

The first data mining technique to be was association rule mining (dependency modeling). With regards to the first analysis approach it was found that the parameters c.q. thresholds of the association rule mining-algorithm, APRIORI, are greatly influenced by the size of the dataset and the distribution of records among each product group. The first analysis approach yielded unacceptable results. The association between products derived; with both thresholds (support and confidence) at a value of 0.1 all belonged to product group 1 and 2. After inspection of the dataset it was clear that the majority of the records were distributed among product group 1 and 2. Therefore, by taking into account the parameters and using the first analysis approach, the APRIORI-algorithm only derived associations for products residing in product group 1 and 2.

By following the second analysis approach in combination with the APRIORI-algorithm, associations were identified between products which do not belong to product group 1 and 2. However, a limitation occurred due to the fact that each product group has its own transactional dataset. the inability of this analysis approach is to derive cross-product group association rules. This limitation could lead to misinterpretation of results and result in wrongful decision making. Hence, the second analysis approach is classified as undesirable.

The third analysis approach used a dataset comprising all transactions and thereby transactional data in which the product under analysis occurred. Three products were selected for evaluating association rule mining in combination with this approach. By analyzing each article separately, it was found that this analysis approach coped with the limitation of cross-product group-association rule generation. It derived association rules between products that belonged to different product groups. Hence, this approach and the data mining technique association rule mining are classified as desirable and applicable in a product portfolio management context. Evidently, for the second and third data mining technique, , the product based approach was used.

For seasonal time series forecasting the exact same three products used in the association rule mining evaluation were used. In order to properly analyze this data mining technique, the transactional dataset provided by the case study organization was extended to 24 months of transactional data, residing in a two year time-period (2010 and 2011). For evaluating this data mining technique the exponential smoothing algorithm was selected. Applying the exponential smoothing algorithm on each of the three dataset yielded undesirable results. For the proper functioning of the exponential smoothing algorithm it is required that in each period (in this case year) per observed data point (in this case months) data are present. For each article this was not the case, which resulted in failed forecasts (flat lines) or partial forecasts (three months in advance). In order to properly evaluate the exponential algorithm, two articles were derived from the dataset comprising 24 months of data observed per similar data point per period. By applying the exponential smoothing algorithm it was able to properly forecast future data points. However, the exponential smoothing algorithm forecasts data points based on historical data. The articles and thereby the generated datasets were selected by means of the threshold value. After inspecting the forecasted data points, none of the data points exceeded the pre-defined threshold value. Concluding on the evaluation of seasonal time series forecasting, it is safe to say that, in a product portfolio management context, the technique is not applicable.

With regards to the data mining category classification, it was desirable to not only classify but in addition also visualize product classification by using a portfolio matrix (i.e. BCG-matrix). However, none of the available classification algorithms support visualization of data. Hence, during this research, a new classification algorithm was proposed. The algorithm was applied on datasets which combined association rule mining metrics with the SalesVolumeInKG-metric. By plotting the products with which the product under analysis is most strongly associated, one is able to classify and visualize the product under analysis on a portfolio matrix. The results are promising with regards to classification and visualization possibilities. However, more refinement and testing of the algorithm is needed.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SO made substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; has been involved in drafting the manuscript. MS supervised the research; made substantial contributions to conception and design of data analysis and interpretation of data; has been involved in revising the manuscript critically for important intellectual content and co-authored the paper. RH secondary advisor of the research; has been involved in revising the manuscript critically for important intellectual content and co-authored the paper. All authors read and approved the final manuscript.

Author details
[1]Department of Information and Computer Sciences, Utrecht University, Princetonplein 5, Utrecht, The Netherlands.
[2]Faculty of Management, Science and Technology, Open University, Valkenburgerweg 177, Heerlen, The Netherlands.

Received: 23 March 2015 Accepted: 11 May 2015
Published online: 04 June 2015

References

Agrawal, R, & Srikant, R. (1994). *Fast algorithms for mining Association Rules in Large Databases* (Proceedings of the 20th International Conference on Very Large Data Bases, pp. 478–99). Santiago de Chile: Citeseer.

Agrawal, R, Imieliski, T, & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record, 22*(2), 207–16.

Aitkenhead, MJ. (2008). A co-evolving decision tree classification method. *Expert Systems with Applications, 34*(1), 18–25.

Ankerst, M, Breuning, MM, Kriegel, HP, & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. *ACM SIGMOD Record, 28*, 49–60.

Archer, NP, & Ghasemzadeh, F. (1999). An integrated framework for project portfolio selection. *International Journal of Project Management, 17*(4), 207–16.

Balkin, SD, & Ord, JK. (2000). Automatic neural network modeling for univariate time series. *International Journal of Forecasting, 16*(4), 509–15.

Bekkers, W, Weerd, I, Spruit, MR, & Brinkkemper, S. (2010). *A framework for process improvement in software product management* (Proceedings of the 17th European Conference on Systems, SOftware and Services Process Improvement (EUROSPI), pp. 1–12). Grenoble, France: Springer.

Berkhin, P. (2006). *Survey of clustering data mining techniques* (pp. 25–71). Grouping Multidimensional Data: Recent Advances in Clustering.

Borgelt, C. (2005). *An Implementation of the FP-growth Algorithm. Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations* (pp. 1–5). Chicago: ACM.

Brijs, T, Swinnen, G, Vanhoof, K, & Wets, G. (1999). *Using association rules for product assortment decisions: A case study. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 254–260). San Diego: ACM.

Brijs, T, Swinnen, G, Vanhoof, K, & Wets, G. (2004). Building an association rules framework to improve product assortment decisions. *Data Mining and Knowledge Discovery, 8*(1), 7–23.

Chao, RO, & Kavadias, S. (2007). A theoretical framework for managing the NPD portfolio: when and how to use strategic buckets. *Management Science, 54*(5), 907–21.

Chapman, P, Clinton, J, Kerber, R, Khabaza, T, Reinartz, T, Shearer, C, et al. (2000). *CRISP-DM 1.0: Step-by-step data mining guide* (CRISP-DM Consortium).

Chen, MS, Han, J, & Yu, PS. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866–83.

Cooper, RG. (1999). The invisible success factors in product innovation. *Journal of Product Innovation Management, 16*(2), 115–33.

Cooper, RG, & Kleinschmidt, EJ. (1995). Benchmarking firms'new product performance and practices. *Engineering Management Review, 23*(3), 1–12.

Cooper, RG, Edgett, SJ, & Kleinschmidt, EJ. (2000). New problems, new solutions: making the portfolio management more effective. *Research - Technology Management, 43*(2), 18–33.

Cooper, R, Edgett, S, & Kleinschmidt, E. (2001). Portfolio management for new product development: results of an industry practices study. *R&D Management, 31*(4), 361–80.

Cooper, RG, Edgett, SJ, & Kleinschmidt, EJ. (2002). *Portfolio management: fundamental to new product success*. New York: Wiley.

Dickinson, MW, Thornton, AC, & Graves, S. (2001). Technology portfolio management: optimizing interdependent projects over multiple time periods. *Engineering Management, 48*(4), 518–27.

Ester, M, Kriegel, HP, Sander, J, & Xu, X. (1996). *A density - based algorithm for discovering clusters in large spatial databases with noise* (Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining, pp. 226–31). Portland: AAAI Press.

Fayyad, U, Piatetsky-Shapiro, G, & Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37–54.

Fayyad, U, Piatetsky-Shapiro, G, & Smyth, P. (1996b). *Knowledge discovery and data mining: Towards a unifying framework. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 82–88). Portland: AAAI.

Fayyad, U, Piatetsky-Shapiro, G, & Smyth, P. (1996c). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27–34.

Fisher, DH. (1987). *Knowledge acquisition via incremental conceptual clustering* (Machine learning, pp. 139–72).

Goldberg, DE. (1989). *Genetic Algorithms in Search, Optimizations and Machine Learning*. Waltham: Morgan-Kaufmann.

Guba, S, Rastogi, R, & Shim, K. (1998). *CURE: an efficient clustering algorithm for large databases*. ACM Sigmod Record: ACM.

Hahsler, M, Grun, B, & Hornik, K. (2005). A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software, 14*(1), 1–25.

Hambrick, DC, MacMillan, IC, & Day, DL. (1982). Strategic attributes and performance in the BCG Matrix–A PIMS-based analysis of industrial product businesses. *Academy of Management Journal, 25*(3), 510–31.

Han, J, Pei, J, & Yin, Y. (2000). *Mining frequent patterns without candidate generation. Proceedings of the 2000 ACM SIGMOD international conference on Management of data. 29* (pp. 1–20). Dallas: ACM.

Han, J, Kamber, M, & Pei, J. (2011). *Data mining: concepts and techniques*. Waltham: Morgan-Kaufmann.

Hill, T, O'Connor, M, & Remus, W. (1996). Neural network models for time series forecasts. *Management Science, 42*(1), 1082–92.

Hylleberg, S. (1992). General Introduction. In S Hylleberg (Ed.), *Modelling Seasonality* (pp. 3–14). Oxford: Oxford University Press.

Hyndman, RJ, & Khandakar, Y. (2007). *Automatic time series forecasting: the forecast package for R. Monash Econometrics and Business Statistics Working*. Papers: Monash University.

Jain, AK, & Dubes, RC. (1988). *Algorithms for Clustering Data*. Englewood Cliffs: Prentice-Hall.

Jain, AK, Murty, MN, & Flynn, PJ. (1999). Data clustering: a review. *ACM Computing Surveys, 31*(3), 264–323.

Kacprzyk, J, & Zadronzy, S. (2005). Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Information Sciences, 173*(4), 281–304.

Karypis, G, Han, EH, & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer, 32*(8), 68–75.

Kotsiantis, S, & Kanellopoulos, D. (2006). Association rules mining: a recent overview. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 71–82.

Larose, DT. (2005). *An Introduction to Data Mining*. Hoboken: Wiley Online Library.

Lawrence, MJ, Edmundson, RH, & O'Connor, MJ. (1985). An examination of the accuracy of judgmental extrapolation of time series. *International Journal of Forecasting, 1*(1), 25–35.

Lippmann, RP. (1988). An introduction to computing with neural nets. Artificial neural networks. *Theoretical Concepts, 209*(1), 36–54.

Maaß, D, Spruit, M, & Waal, P. (2014). Improving short-term demand forecasting for short-lifecycle consumer products with data mining techniques. *Decision Analysis, 1*, 4.

Makridakis, S, & Wheelwright, SC. (1987). *THe handbook of Forecasting: A manager's Guide*. New York: Wiley.

Niewiadomski, A. (2008). A type-2 fuzzy approach to linguistic summarization of data. *IEEE Transactions on Fuzzy Systems, 16*(1), 198–212.

Ordonez, C. (2006). Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine, 10*(2), 334–43.

Ordonez, C, & Omiecinski, E. (1999). *Discovering association rules based on image content* (Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries, pp. 38–50). Baltimore: IEEE Computer Society.

Pachidi, S, Spruit, M, & Weerd, I. (2014). Understanding users' behavior with software operation data mining. *Computers in Human Behavior, 30*, 583–94.

Park, HS. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications, 36*(2), 3336–41.

Piatetsky-Shapiro, G, & Frawley, WJ. (1991). *Knowledge Discovery in Databases*. Cambridge, MA, USA: MIT Press.

Quinlan, JR. (1993). *C4.5: Programs for Machine Learning*. Waltham: Morgan-Kaufmann.

Scott, DW. (1992). *Multivariate Density Estimation*. New York: Wiley.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal, 16*(1), 30–4.

Smyth, GK. (2002). Nonlinear regression. In AH El-Shaarawi & WW Piegorsch (Eds.), *Encyclopedia of Environmetrics* (pp. 1405–11). Chichester: Wiley & Sons.

Spruit, M, Vroon, R, & Batenburg, R. (2014). Towards healthcare business intelligence in long-term care: an explorative case study in the Netherlands. *Computers in Human Behavior, 30*, 698–707.

Triki, A, Pollet, Y, & Ben Ahmed, M. (2008). *Database summarization approach based on description logic theory. Proceedings of the 6th IEEE International Conference on Industrial Informatics* (pp. 1285–1288). Daejeon: IEEE.

Vleugel, A, Spruit, M, & Daal, A. (2010). Historical data analysis through data mining from an outsourcing perspective: the three-phases method. *International Journal of Business Intelligence Research, 1*(3), 42–65.

Weiss, SI, & Kulikowski, C. (1991). *Computer Systems That Learn: Classification and Predicition Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems*. San Francisco: Morgan Kaufmann.

Wu, D, Mendel, JM, & Joo, J. (2010). *Linguistic summarization using IF-THEN rules. Proceedings of the 2010 IEEE International Conference on Fuzzy Systems* (pp. 1–8). Barcelona: IEEE.

Xu, X, Ester, M, Kriegel, HP, & Sander, J. (1998). *A distribution-based clustering algorithm for mining in large spatial databases* (Proceedings of the 14th International Conference on Data Engineering, pp. 324–31). Orlando: IEEE.

Yager, RR. (1982). A new approach to the summarization of data. *Information Sciences, 28*(1), 69–86.

Zhang, PG. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*(1), 159–75.

Zhang, GP, & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research, 160*(2), 501–14.

Zhu, RF, & Takaoka, T. (1989). A technique for two-dimensional pattern matching. *Communications of the ACM, 32*(9), 1110–20.