Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 62 (2017) 123 – 128

www.elsevier.com/locate/procedia

10th CIRP Conference on Intelligent Computation in Manufacturing Engineering - CIRP ICME '16

Data mining techniques applied to a manufacturing SME

Michael S Packianather[a,*], Alan Davies[a], Sam Harraden[a], Sajith Soman[b], John White[b]

[a]*School of Engineering, Cardiff University, Cardiff CF24 3AA, UK*
[b]*Brick Fabrication Ltd, Pontypool NP4 6YW, South Wales, UK*

** Corresponding author. Tel.: +44-029-20875911 ; fax: +44-029-20874716. E-mail address: PackianatherMS@cf.ac.uk*

Abstract

This paper examines how data mining, an aspect of analytical science, can be applied to assist a Small to Medium Enterprise (SME) industry using unsupervised learning techniques, association rules and time-series analysis. Whilst recent developments have meant it is now possible for SME to compile large amounts of commercial data, this information is rarely utilised effectively. The study builds on a number of standard data mining techniques to produce a tailored set of analyses that provide maximum benefit to the company. Self-Organising Maps were utilised to visualise the core characteristics of the firm's customers. The study outlines a new technique to determine associations between customer variables using the arules package available within RStudios. Finally, time-series forecasting was conducted highlighting the seasonal variations and trends for potential growth in the coming year.

Peer-review under responsibility of the scientific committee of the 10th CIRP Conference on Intelligent Computation in Manufacturing Engineering

Keywords: Data mining; time-series analyses; Association rules; Unsupervised learning; K-means clustering; Hierarchical clustering and self-organising maps

1. Introduction

The advances in data processing and storage power has meant that data analytics is no longer the privilege of large multinational technology corporations but it can be used as a key component in the formation of strategy for companies of a range of sizes. In fact, data analytics is being used more and more by Small to Medium Enterprises (SMEs) to discover a wealth of information, including; customer purchasing patterns, sales forecasting, and efficient customer relationship management. Analysis by Bokman et al. [1] states a 126% profit improvement over competitors by companies that make extensive use of customer analytics. As such, failure to incorporate data analysis into strategy formulation can result in competitive disadvantage within the industry.

Brick Fabrication (BF) Ltd. is an SME involved in manufacturing various products made of cut bricks including brick clad chimneys, arches, etc., for the housing market. The BF headquarters is in Pontypool, Wales, with an additional site in Bromley, Essex, providing Brick Clad Chimney products to a range of clients, generating approximately three million pounds worth of revenue per annum. The company's mission is to become the fastest, most reliable and the most innovative supplier of building solution services in the UK and as such, need an innovative and real-time strategy to match. Whilst a wealth of commercial data is currently collected by the firm in order to process invoice and payment orders, this information has not yet been used as a means of improving the growth of the firm. This study uses BF as a case study into how data mining – an aspect of analytics – can be used to assist in the discovery of knowledge and formulation of strategy through analysis of commercial sales data.

The core aim of this study is to demonstrate how data mining can be applied to real world data, using BFs commercial database as a case study. In addition to this, the study sets out to produce a series of results from which the company can make informed decisions in a strategic manner.

The specific objectives of this study are:

- Provide an understanding of key aspects of data mining: Unsupervised learning, association analysis, and time-series analysis.
- Utilise unsupervised learning techniques to segment BFs customer list.

Peer-review under responsibility of the scientific committee of the 10th CIRP Conference on Intelligent Computation in Manufacturing Engineering
doi:10.1016/j.procir.2016.06.120

- Expand on association analysis to better understand what constitutes a typical order for a BF customer.
- Develop time-series forecasting techniques to discover patterns and trends of customers.
- Establish a business case for using long-term data mining analysis within BF Ltd.

The paper is organized as follows. Data mining techniques are described in section 2. The results of data mining techniques are presented in section 3. The paper is concluded in section 4.

2. Datamining Techniques

Data Mining is a general term, describing the method in which one can discover hidden trends, patterns and associations that would otherwise remain undiscovered [2]. In addition to this, it can be used to predict future occurrences, such as the growth of the market or the likelihood of customer loyalty. It relies on aspects of statistics, mathematics and computing, and should be used as part of an overall holistic approach to knowledge discovery.

The methodology used for the data mining analysis is a framework referenced from Giudici's Applications of Data Mining [3] and is as follows:

A. Definition of the objectives for analysis.
B. Selection, organisation & pre-treatment of the data.
C. Exploratory analysis of the data and subsequent transformation.
D. Specification of the statistical methods to be used in the analysis phase.
E. Analysis of the data based on the chosen methods.
F. Evaluation and comparison of the methods used and the choice of the final model for analysis.
G. Interpretation of the chosen model and its subsequent use in decision processes.

In an Introduction to Data Mining, Tan et al. [4] outlines five key aspects of data mining: data exploration, classification, clustering, association, and time-series analysis. To provide a comprehensive understanding of data mining, the theory of some of these major techniques will be discussed below.

2.1. Data Exploration

Data Exploration is an initial investigation of the data to understand its main characteristics and decide on the best approach to extract meaningful information. Its primary purpose is to help decide on the most appropriate pre-processing and data analysis techniques. Specific to this project, the aims of the data exploration were to: discover where the majority of revenue was generated for the firm, analyse general trends in growth, and determine the geographical location of customers.

2.2. Classification

Classification is a machine learning technique which classifies data into pre-defined groups. As such, distinct class labels are required. No such labels were available for the data used in this work and as a result classification will not be explored in this study.

2.3. Clustering

Clustering is the process of grouping together data points that are more similar in attributes than other data points. It is an unsupervised learning technique, meaning it can be used to analyse meaningful patterns without human supervision or intervention. To quantify the similarity of two data points, the distance between points is required. This is typically calculated using the Euclidean distance given in equation (1):

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad (1)$$

Where n is the number of dimensions and x_k and y_k are, respectively, the k^{th} attributes of features x and y.

There are two main Clustering techniques which group data based on similarity: Hierarchical and Partioning. Hierarchical Clustering builds a hierarchy of clusters through one of two approaches known as Agglomerative and Divisive clustering. Agglomerative clustering begins with all observations of a data set beginning as an individual cluster, before pairs are merged together into a hierarchy structure. Meanwhile divisive clustering begins with all observations under one cluster, before moving individual observations down the hierarchy.

Partioning on the other hand is a method that constructs various partitions, evaluating them by similarity and distance to one another. The k-means algorithm is one of the most popular methods to achieve this. The algorithm begins by assigning k different random data points as initial centroids. The distance between each data point and the cluster centre is subsequently calculated, and data points are assigned to the centre point with the minimum distance. New cluster centres are subsequently calculated, with the aim of minimising the Squared Error Function given in equation (2):

$$v = (1/c_i)\sum_{i=1}^{c_i} x_i \quad (2)$$

Where c_i represents the number of data points in *i*th cluster. The distance between each data point and cluster centre is recalculated, assigning data points to their new nearest cluster. This process is continued until iteration results in no further changes.

The advantage of the k-means algorithm is that it is simple to use and will automatically assign clusters. The primary disadvantage is that the analyst must pick an arbitrary number of clusters, without knowledge of the optimal number of groups to assign. In addition to this, the algorithm struggles to assign clusters when the data is not well separated or groups are not dissimilar from one another.

Self-Organising Maps (SOMs) are a form of Artificial Neural Networks (a system of statistical learning algorithms that replicate human biological brain patterns) that can be used to cluster multidimensional data, lowering the dimensional space and allowing it to be presented in a two dimensional form. SOMs create networks of "nodes" across a grid, processing the data in a way that allows relationship between individual items to be represented through a topological structure (i.e. one where properties remain preserved) [5].

The training process a SOM undertakes to train data begins by classifying weights for each node. The weightings assigned are small, random, standardized values that differ from one another. A vector (a quantified value of an item's variables) is then presented randomly to the grid of nodes (known as a lattice) and assigned to the node which it is most similar to, known as the Best Matching Unit. This is typically done using the Euclidean distance discussed earlier.

Once a Best Matching Unit has been determined, the node will assign itself a local neighbourhood. This neighbourhood will begin large, shrinking in radius over time by use of an exponential decay function.

The algorithm can be succinctly described step by step as outlined by Bacao and Lobo, [6] as follows:

1) Calculate the distance between the pattern and all neurons.
2) Select the nearest neuron as winner.
3) Update each neuron according to the rule:

$$w_{ij} = w_{ij} + \alpha\ h(w_{winner}, wij)\ \|x_k - w_{ij}\| \quad (3)$$

4) Repeat the process until a certain stopping criterion is met. Usually, the stopping criterion is a fixed number of iterations

Self-Organising Maps were tested on Brick Fabrication data where the process could automatically discover patterns between points based on a large number of variables. This could help discover patterns that the company are not currently aware of.

2.4. Association Rules

Association analysis deals with discovering hidden relationships within data sets. Fundamentally, it allows one to discover the probability of one specific event occurring as a direct result of another. Association can be performed in a number of ways, one of which is market basket analysis. Market basket analysis inspects transaction history, looking at cases of multiple items being bought in one single transaction. This information could be used for multiple purposes at Brick Fabrication, such as altering the website catalogue so customers view complimentary products together or assisting with inventory.

In addition to providing an understanding of what items customers are purchasing together, market basket analysis can also be used to highlight multiple items that may not be purchased together where one would expect them to, potentially discovering a gap in marketing which needs to be covered.

To be able to interpret association rules, it is important that one has a basic understanding of the computations used. Pang-Ning Tan (Tan, 2003) describes an association rule as "an implication expression of the form X ➔ Y, where X and Y are disjoint item sets, i.e. $X \cap Y = \emptyset$)." This definition is better explained by examining the two key criteria used in assessing the value of an association rule: The Support and Confidence. Support signifies the frequency of a specific rule within a dataset. The higher the support, the more that rule is involved in the database. According to Agrawal et al. [7] support is defined as:

$$Support, s(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \quad (4)$$

Meanwhile, confidence is defined as the percentage of occurrences that contain X which also contains Y.

$$Confidence, c(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (5)$$

The support helps one appreciate how prevalent a rule is within the data, whilst the confidence assesses the conditional probability of that rule occurring.

The challenge for association analysis is that 100,000s of different rules can be generated from a small number of different products, even when a minimum support threshold and confidence threshold are set. As the majority of these rules are of little interest to the analyst, algorithms such as Apriori are used to eliminate sub-set of rules and rule duplication.

2.5. Time-series Analysis

A time-series is defined by Esling and Agon as "a collection of values obtained from sequential measurements over time" [8]. Data Mining can be of use for a huge range of purposes relating to time-series due to the high dimensionality of the data. Application of data mining can include areas such as trend alignment, similarity modelling, anomaly detection and forecasting.

Time-Series forecasting involves quantifying a series of observations into a pattern and using this to predict future events. One of the most common means of forecasting within data mining is Exponential Smoothing. According to Kalekar, Exponential smoothing can be defined as "a procedure for continually revising a forecast in the light of more recent experience" [9]. Essentially, the more recent an observation is, the higher the importance given to that observation over older observations.

Exponential Smoothing can be performed at three levels: Single order, double order and triple order. Single exponential smoothing is used when the data oscillates around a fixed average, that is, no trend is observed. Double exponential smoothing observes a general trend where no recurrent pattern across a fixed cycle is witnessed. Third exponential smoothing is used when both a trend and seasonality are apparent in the data.

The basis of triple exponential smoothing relies on a series of equations known as Holt-Winters. These equations will differ depending on whether the dataset is Multiplicative or Additive. The basis of multiplicative equations are outlined due to the likelihood of Brick Fabrications time-series data displaying this trend. Firstly, the time series (y_t) is represented as in equation 6, by decomposing the time-series into a permanent component (b_1), a trend component (b_2), a seasonal component (S_t) and a random error component (ϵ_t).

$$y_t = (b_1 + b_2 t)S_t + \epsilon_t \quad (6)$$

Smoothing constants α (the overall smoothing constant), β (the trend constant), and γ (the seasonality constant) are then introduced to the model to obtain smoothing parameters for R_t (the overall smoothing parameter), G_t (the trend parameter) and S_t (the seasonal parameter) respectively. Equation 7 can subsequently be applied to calculate a forecast for the next time period.

$$y_t = (R_{t-1} + G_{t-1})S_{t-L} \quad (7)$$

Time series analysis is applied to Brick Fabrication data to observe upward, downward, cyclic and seasonal trends and use this information for forecasting purposes.

3. Datamining Results

Two years' worth of commercial sales data was available for this study. The format of the data was one in which rows signified individual product orders, whilst columns represented the attributes associated to each individual product order. The attributes were as follows: Unique transaction code, stock description, quantity of stock ordered, price, account name, and date. Additional attributes on each account such as region and annual turnover were also obtained through pre-processing and external search.

3.1. Data Exploration Results

Data exploration was conducted with three objectives: Discover the distribution of order frequency by accounts, determine the key products sold by the firm and identify key regions where revenue was generated in geographical terms.

Firstly, a histogram was plotted based on the orders made per account to understand the purchasing habits of clients. Fig. 1 shows that the majority of account holders made between 1-50 orders in the two year period. It can also be seen that one firm made over 600 orders which represented 9.25% of the total revenue.

Secondly, the top 10 products (ranked in terms of total revenue) were inspected. These made up 65.05% of total revenue. The most frequently purchased products were Flat Gauge Arch Panelites and Segmental Arch Panelites, which were the two highest revenue generating products for over 90% of individual accounts. As seen in Fig. 2, the top 10 highest revenue generating products are a range of prices, representing good portfolio diversity. The standard deviation also shows the variation in the average cost of a product.

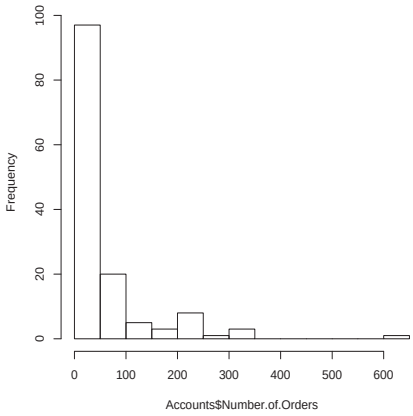

Fig. 1. Frequency of orders per account.

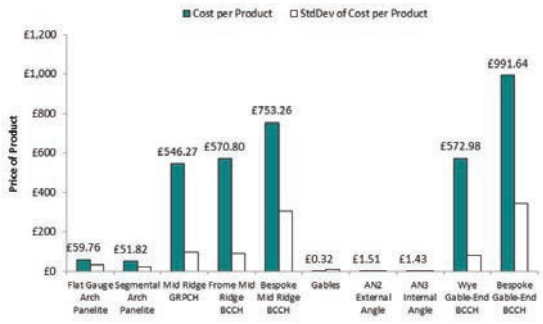

Fig. 2. Top 10 most frequently ordered products.

Lastly, using the lookup table each account was assigned to a region, and this was used to generate Table 1 giving the breakdown of revenue created by account holders per region. This showed South Central generating the most revenue per account, whilst Wales generating the highest share of revenue overall.

Table 1. Breakdown of account holders per region and revenue.

Region	Number %	Revenue %
East England	11.45%	9.03%
Midlands	21.37%	17.76%
North West	19.08%	13.29%
South West	13.74%	13.70%
Wales	19.85%	24.88%
South Central	8.40%	13.05%
Other	6.11%	8.29%

3.2. Customer Segmentation Results

The aim of this analysis was to apply data clustering using unsupervised learning techniques to discover which of Brick Fabrication customers were the most profitable and valuable. The unsupervised learning techniques used were k-means clustering, and Kohonen's Self Organising Map.

The first analysis that was carried out was clustering using the k-means algorithm. To achieve this, each variable was scaled on a value of 0-100, where the maximum value of each

variable was assigned a value of 100, and all others as a proportion of that. This meant each variable had equal weighting. K-means clustering was subsequently carried out, calculating the similarity between items using the Euclidean distance based on three key variables, Unique Orders, Revenue, and Quantity of Products Ordered. The k-means algorithm was decided upon as a starting point due to its simplicity and the globular dense structure of the smaller items of data.

A plot of the clusters on a Revenue by Account vs. Quantity Ordered by Account scatter plot is shown in Fig. 3, whilst Table 2 displays financial characteristics for each cluster group.

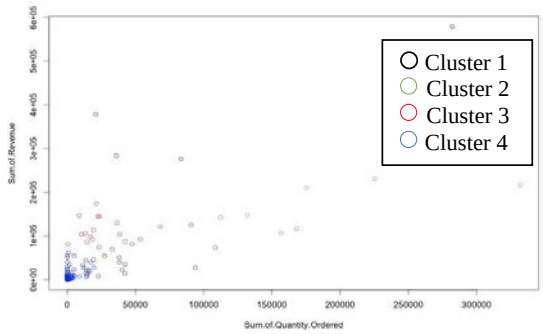

Fig. 3. Scatterplot of Revenue Vs. Quantity per Account.

Table 2. Breakdown of revenue per cluster.

Cluster	Size	Mean No. of Orders	Revenue Generated (£)	Revenue / Order (£)	Revenue / Quantity (£)
1	4	342	1,517,416	1,109.2	3.6
2	21	124	2,248,895	866.6	3.2
3	7	236	1,172,516	710.2	0.9
4	106	16	1,310,317	793.7	2.1

The cluster groupings provided some interesting results. Firstly, the algorithm automatically grouped the top four accounts in terms of revenue to Cluster 1, indicating the results to be accurate. Analysis of these four accounts show they generated 25% of total revenue across the 2 year period and as such are likely to be of high importance to Brick Fabrication. The second point of interest was the relatively lower Revenue per Quantity value (£0.90) attributed to Cluster 3. This suggests Cluster 3 made low cost orders (below £750) with a preference for low cost products. This was looked into deeper and showed Cluster 3 has similar spending habits to the other groups, however was much more likely to purchase low value products (below 0.50p) such as gables. Meanwhile, Cluster 4 purchased less high end products such as BCCH products than other groups. The growth of each cluster was then calculated. Interestingly, all groups saw strong growth except Cluster 1. This should be taken into account by Brick Fabrication for risk management. If this cluster group continues to decline, 25% of the company's revenue could be at stake.

3.3. Kohonen Self-Organising Maps Results

Whilst the clustering techniques proved useful at grouping customers based on the three key variables, introducing multiple variables made the data more difficult to visualise. For this reason, it was decided to use Kohonen's Self Organising Map to explore a much larger degree of account characteristics, possibly discovering further hidden relationships. For comparison with the k-means analysis, the algorithm was driven by the same three variables as before: Unique Orders, Revenue, and Quantity of Products Ordered (again, these variables were scaled in the same way as before). Growth, Company Size and Number of Different Products ordered were also added to the data set to be visualised on the grid. A fan diagram shown in Fig. 4 was created, providing an instant and clear view of the distribution of each variable for the nodes within the grid. It can be seen that a dense cluster of nodes in the top right hand corner of the grid exists, similar to Cluster 1.

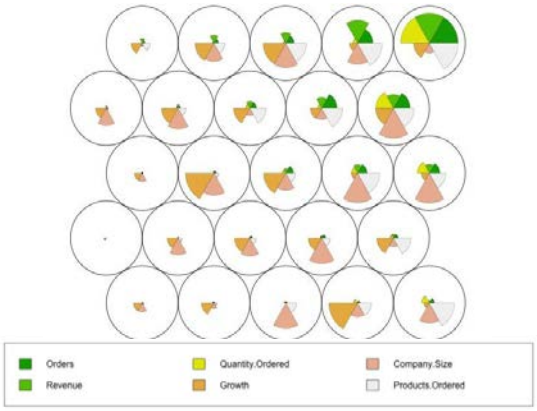

Fig. 4. Kohonen Self Organising Map with 5x5 nodes shown in Fan diagram displaying breakdown of variables.

3.4. Association Rules Results

This section presents the results of association analysis, conducted using the arules package in R. There were approximately 10 top products which made up over half of all revenue. The top 5 products ranked by revenue were discovered to be: Flat Gauge Arch Panelite, Segmental Arch Panelite, Mid-Ridge BCCH, Frome Mid Ridge BCCH, and Bespoke Mid Ridge BCCH.

The first analysis that was run was standard market basket analysis on the transaction data. The first requirement was to pre-process the data, converting the raw transactions from a data frame format to a transaction format. To begin the process, minimum thresholds of 0.01 Support and 0.1 Confidence were set and the arules algorithm was implemented generating 327 rules. The Support Threshold specified that the rule had to be apparent in at least 1% of transactions (71 unique orders), whilst the confidence threshold of 0.1 meant that when item A (LHS) was purchased, item B was also purchased at least 10% of the

time. Rule pruning was then carried out using the apriori algorithm. This algorithm filtered out repeating rules (i.e. IF 'A → B' prune B → A) and subset rules (i.e. IF 'A, B → C' PRUNE 'A, C → B'), reducing the number of rules to 66. Inspection of the pruned rules showed when two or more products were purchased in the same order, they are typically dominated by two products and their variations: AN3 Internal Angles and AN2 External Angles. This is visualised by the top 50 rules matrix shown in Fig. 5. Analysis showed this combination had a 66% confidence rating and that this combination (including variations) appeared in 47.94% of all orders. Beyond this, Segmental Arch Panelites and Gables also featured highly in a typical basket. This is similar to the findings of the data exploration that showed all these to be highly popular and regularly purchased orders, so it is not surprising they have high levels of confidence to one another.

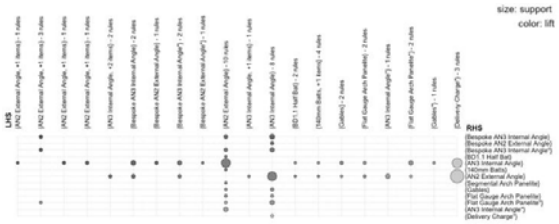

Fig. 5. Top 50 matrix of rules.

3.5. Time Series Results

The aim of conducting Time Series analyses was to provide an insight into what growth Brick Fabrication could expect to see in 2015. As only two years' worth of sales data was available on Brick Fabrication, a regression model was created using Construction data from the Office of National Statistics as an independent variable.

As a starting point, total monthly revenue was plotted for 2013-14, and decomposed to display: observed, trend, seasonality and randomness as shown in Fig. 6. The graphs show a high level of seasonality and a rising trend pattern between July 2013 and July 2014.

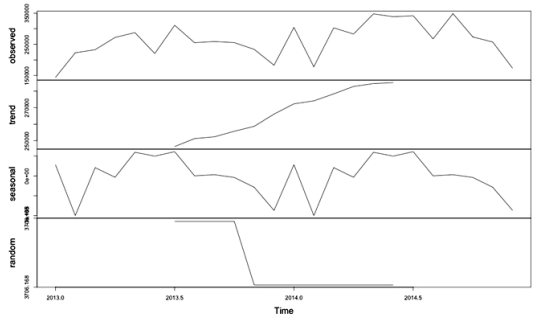

Fig. 6. Decomposition of Time-Series.

Table 3 displays standardised seasonal trend values for the calendar year. This information is of use to the BF as it can help display when demand will be highest and plan accordingly. To achieve standardised values, July was assigned a base value of 100 (as it had the highest seasonality), and February a value of 0 (due to the lowest seasonality). It can be seen from Table 3 that the months of May, June and July are extremely busy, and as such resources should be managed effectively around this time. The second half of the year typically experiences a steep slow down, and this time should be used for expansion planning, as it will minimise the effect on production.

Table 3. Monthly Indexed Demand.

Month	Index Demand	Month	Index Demand	Month	Index Demand
January	80	May	99	September	64
February	0	June	93	October	60
March	75	July	100	November	44
April	60	August	62	December	8

4. Conclusion and future work

This paper has presented some datamining techniques namely data exploration, customer segmentation, Kohonen's Self organising map, association rules, and time series to real data to extract knowledge and patterns for strategic decision making and forecasting. Future work would look into data classification of manufacturing time and costs required to manufacture each product in order to assign a profitability label to each account based on purchases.

Acknowledgements

The authors would like to thank Innovate UK, ASTUTE 2020 project and CAMSAC for supporting this work.

References

[1] Bokman A, Fiedler L, Perrey J, Pickersgill A. Using customer analytics to boost corporate performance. McKinsey & Company; 2014.
[2] Pham DT, Packianather MS, Dimov S, Soroka AJ, Girard T, Bigot S, Salem Z. An application of data mining and machine learning techniques in the metal industry. In Proceedings of the 4th CIRP International Seminar on Intelligent Computation in Manufacturing Engineering (ICME-04), Sorrento (Naples), Italy;2004.
[3] Giudici P. Applied Data Mining. Wiley; 2003.
[4] Tan PN, Steinbach M, Kumar V. Association analysis: basic concepts and algorithms. Introduction to data mining, 2005; 327-414.
[5] Kohonen T. Self-organizing maps, Series in Information Sciences, Heidelberg: Springer, 3rd Ed. 2001.
[6] Bação F, dan Lobo V. Introduction to Kohonen's Self-Organizing Maps, Instituto Superior de Estatistica E Gestao de Informacao (ISEGI), Universidade Nova de Lisboa, Portugal; 2010.
[7] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. ACM SIGMOD Record. 1993; 22(2), 207-216.
[8] Esling P, Agon C. Time-series data mining. ACM Computing Surveys (CSUR). 2012; 45(1), 12.
[9] Kalekar PS. (2004). Time-Series Forecasting using Holt-Winters Exponential Smoothing. Kanwal Rekhi School of Information Technology. 2004; 4329008, 1-13.