ISSN 1407-7493 COMPUTER SCIENCE DATORZINĀTNE 2008-7493

INFORMATION TECHNOLOGY AND MANAGEMENT SCIENCE
INFORMĀCIJAS TEHNOLOĢIJA UN VADĪBAS ZINĀTNE

RULE INDUCTION FOR FORECASTING TRANSITION POINTS IN PRODUCT LIFE CYCLE DATA

Arnis Kirshners, B.sc.ing., Master's student, Department of Modelling and Simulation, Riga Technical University, 1 Kalku Street, Riga, LV- 1658, Latvia, e-mail: arnis@levisriga.lv.

Anatoly Sukov, Mg.Sc.ing., Lecturer, Riga Technical University, Department of Modelling and Simulation, 1 Kalku Street, Riga, LV- 1658, Latvia, e-mail: Anatolijs.Sukovs@cs.rtu.lv.

Keywords: time series, forecasting of transition points, grid-clustering method, classification of data, clustering of data

1. Introduction

One of the temporal data mining types is time series forecasting, which for example is used in trading. In order to successfully introduce a new product to the market, it is important to know how similar products' demand has fluctuated in the previous period of time, for how long this product was in demand and how long the product stayed on the market from the moment it was introduced until the moment its demand fell. With the use of data mining the producer can get some significant data related to the product in question. For data to be used in data mining, it has to be transformed to common relations. For example, if information about different products is collected in the same period of time, data mining provides the tools to process the data for further classification, using rule induction. The extracted rules describe the conditions for forecasting life phases of some product in the future.

This paper analyzes methods and approaches for data set pre-processing for rule induction that can be applied to the given problem area. Also the parameters of these methods will be analyzed to see their effect on the task-solving results.

2. Product life cycle represented by the time series

Product life cycle is a period that consists of five general development phases: product development, introduction, growth, maturity and decline. These phases exist for all types of products and services.

It is hard to foresee a decline in consumption. Usually it can be seen from decreasing sales volumes. But still it is more difficult to realize this transition point, the transition from one life cycle to another, than one can imagine because good marketing makes you think that this product can still reach high sell rates. It makes recognising this transition point even more important.

Since the purpose of this study is to forecast the transition points in the product life cycle data, the product life cycle is modified, according to the problem area (see Figure 1).

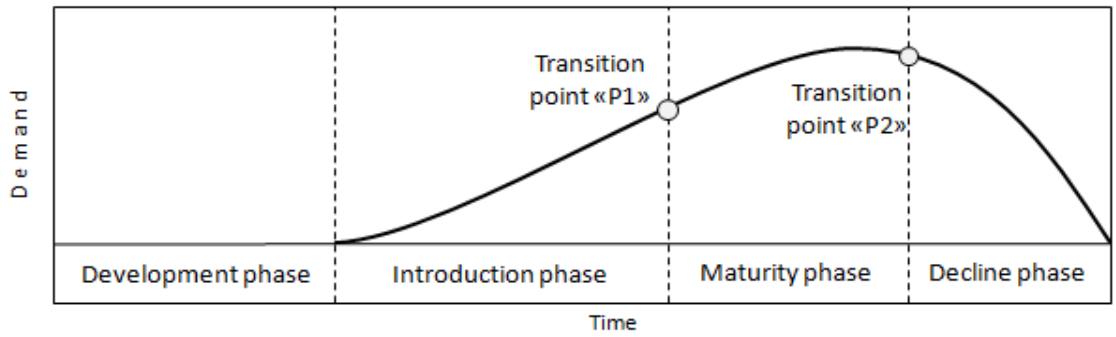


Figure 1. Modified product life cycle

The development phase is ignored because it is not used; introduction and growth phases are merged into an introduction phase because it will describe a state of a product until the maturity phase is reached. The forecasted transition point between introduction and maturity phases in the following text will be referred to as transition point «P1», transition point between maturity and decline phase – transition point «P2».

3. Time series mining

Time series is successive events that are organized by the time of their observation [6]. Events are recorded at regular intervals and are shown as a set. There are three directions in time series mining: *analysis*, *pattern recognition* and *forecasting* [1]. Time series forecasting is one of the most popular and demanded tasks. The forecasting model is made based on the data sets with previously known results. These results are used to forecast results that are based on other data sets. Therefore forecast model has to be statistically relevant and reasonable for forecasting to be as accurate as possible. To ensure this accuracy, classification tasks are used in these time series. Classification tasks describe the rules or a set of rules, according to which, any new object can be added to an existing class. These rules are created based on the information about the existing objects that are already put into classes [6]. Based on the switching in the time series from one product life cycle to another, a class is assigned in the classification tasks. The class is being described by «Ne» until the switching, but the switching itself is described by class «Ja», making classification tasks.

The classical time series forecasting passes three stages: building a model, that describes the data series, using statistical and classification methods; evaluating the created model, which includes splitting available data into training and test sets. The model is created using the training set and the forecast is made using the test set. The forecast is compared with real data and the model is evaluated using forecast error. If the first model passes the test, this model can be used for forecasting in the future.

4. Concept of the transition points

The concept of forecasting transition points is based on a forecast of two transition points, that determine the following transitions [3]: from *introduction* phase into *maturity* phase (described by point «P1») and from *maturity* phase into *decline* phase (described by point «P2»).

Previously gathered input information is used to forecast transition points «P1» and «P2». This data needs some input data pre-processing according to the rules. The input data describes time series that hold demand information of 24 months that is described by the following parameters:

- Field «*ID*» describes the sequence number of the time series;

- Field «Introduction» refers to the period when the product emerges in the market;
- Field «K1» refers to the period when transition between two life cycles happens;
- Field «K2» indicates the number of periods for which information will be collected;
- Fields «T1» to «T24» describe the demand for a given period.

The value of parameter «*K2*» is calculated by the following algorithm:

- If the value of the field «*Introduction*» is equal to 0, then *K2=K1*;
- If the value of the field «*Introduction*» is greater than 0, then *K2=(K1-Introduction)+1*.

Data is clustered using grid-clustering method with different numbers of clusters. Data classification is done through a classification task, which creates two classes - «Ne» and «Ja», where class «Ja» is assigned to the data set that indicates a transition between life cycle phases. In data discretization, where new data sets are created, data values are replaced with block numbers that were obtained in clusterization. These data sets are saved in *. arff format for further use in *Weka*, in classifier performance scoring.

4.1. Time series pre-processing

For the data to be used with data mining algorithms, it has to be initially processed to exclude noisy and missing data, as well as dominating values because it tunes down algorithm performance, contributes to inaccurate results and slows down the algorithm. Data preparation process can be divided into the following stages:

- time series selection based on the value of attribute «*K1*» that should to be in the range *2<K1<24*;
- time series cleaning – time series that don't match the value of «*K1*» are deleted;
- time series transformation – values are transferred to the left side without any changes in their meaning;
- time series normalization using Z-estimation normalization, because maximum and minimum limits of the attributes are unknown.

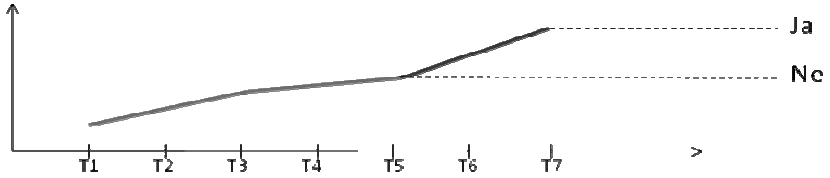
After pre-processing the data, values of attribute «*K2*» have to be recalculated so that pairs of given periods can be made, which enables data classification. It is recalculated by the following algorithm:

- if the value of attribute «*K2*» is an odd number, then *K2=K2+1*;
- if the value of attribute «*K2*» is an even number, then *K2=K2+2*.

4.2. Time series classification

Classification is adding new objects to the existing groups of objects [1]. Based on previously gained knowledge, classes are assigned to these object groups. Classification is done, taking pairs of time series in periods, where the first data set consists of four periods because the minimum period for transition from one life-cycle to another is greater than or equal to 3. If there are other pairs after the chosen pairs, data set is classified into «Ne» class; the values are passed to the next data set where the next pair is added. If after this addition there are no more pairs in the set, then the set is classified as «Ja», otherwise the values are passed to the next data set and classified as «Ne». All time series are searched in a similar manner until all possible data sets and their classes are found. An example of classification realization with one time series, which consists of six periods, is given in Figure 2.

T1	T2	T3	T4	T5	T6	...	T24
-3,00629	-1,74833	-0,73562	0,35653	0,46171	0,51204		



T1	T2	T3	T4	T5	T6	...	T24	Class
-3,00629	-1,74833	-0,73562	0,35653					Ne
-3,00629	-1,74833	-0,73562	0,35653	0,46171	0,51204			Ja

Figure 2. Time series classification

4.3. Time series clustering

Clustering is cluster analysis or object grouping into clusters [1]. Objects of one cluster have to have similar qualities that serve as a cause for grouping. As a result of clustering, some information about grouped objects is extracted. *Clustering* means grouping data objects based on values of these objects. This task can be solved with different approaches: hierarchical, which means that a cluster can have sub-clusters, or partitioned, which means that one record can belong to only one cluster. In this case the grid-clustering method with 4, 9 or 16 clusters was used to cluster the data.

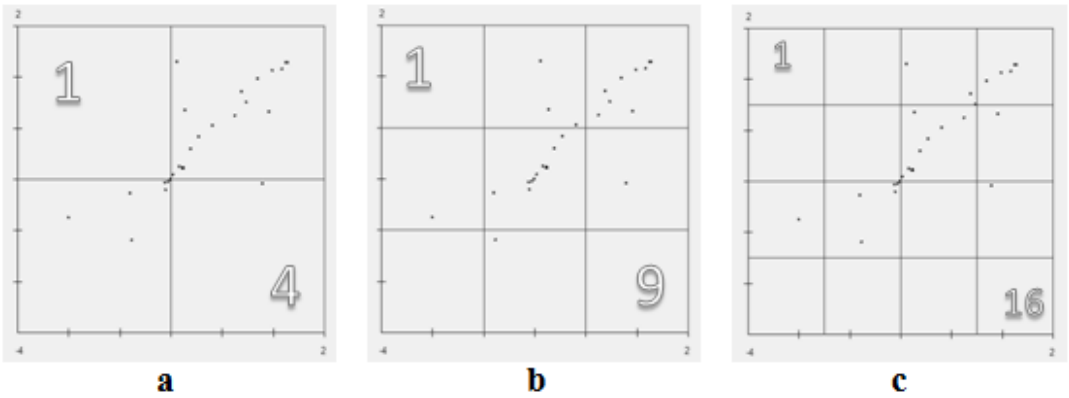


Figure 3. Dividing range of values into 4(a), 9(b) and 16(c) clusters and assigning numbers

Grid-clustering method is based on Statistical Information Grid, which divides the data value space into equal square cells. The space of data values is divided into cells that form the structure of the grid, which is a basis for clustering operations [4]. Grid-clustering method is based on distinguishing the maximum and minimum normalized value of the set of normalized values. Values are rounded to an integer (the lowest value is rounded down, the highest value is rounded up) using mathematical functions. The range of values is divided into 4, 9 or 16 parts, which corresponds to the chosen number of clusters. Division of the range of values can be seen in Figure 3. Cluster numbers or block indexes are assigned starting from top left corner, moving right and down the rows.

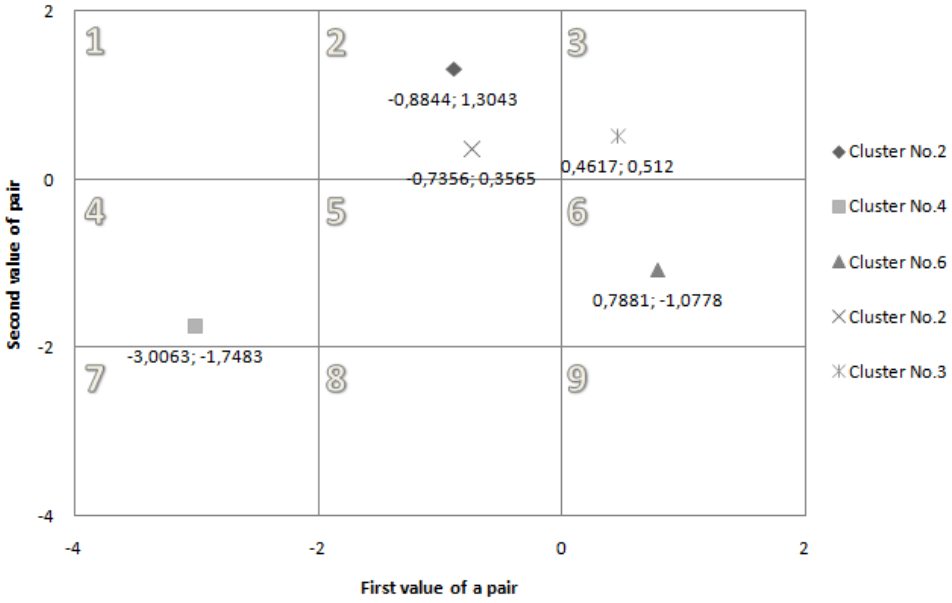
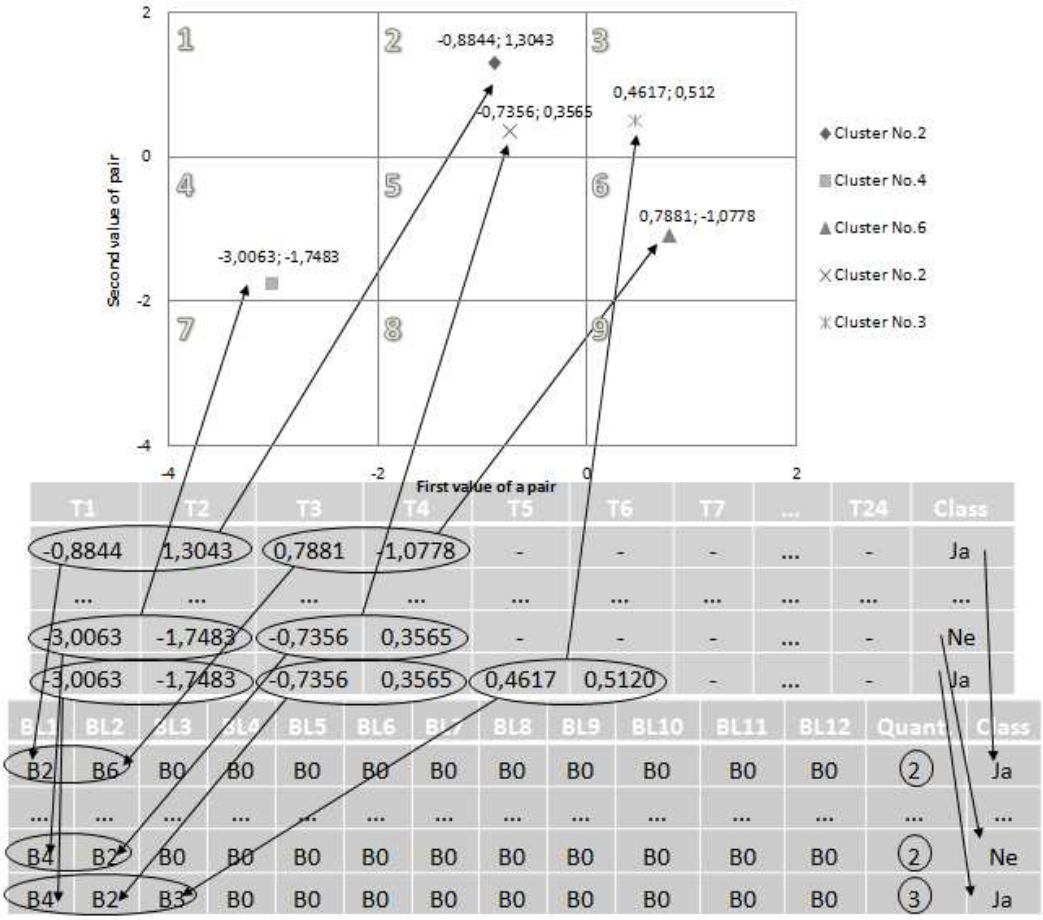


Figure 4. Realization of grid-clustering method

In clustering, time periods are arranged in pairs. The value of the first period is marked on «X» axis and the second period of the pair is marked on the «Y» axis, determining the number of a cluster or block in the clustering grid. Realization of this kind of clustering is shown in Figure 4.

4.4. Time series discretization

Time series *discretization* includes replacing values with numbers [2]. Before performing time series discretization the data is transformed into unified data set – data sets with block numbers from clustering and data sets with classes from classification.



T1	T2	T3	T4	T5	T6	T7	...	T24	Class
-0,8844	1,3043	0,7881	-1,0778	-	-	-	...	-	Ja
...	...	...	...	...	...	...	...	...	...
-3,0063	-1,7483	-0,7356	0,3565	-	-	-	...	-	Ne
-3,0063	-1,7483	-0,7356	0,3565	0,4617	0,5120	-	...	-	Ja

BL1	BL2	BL3	BL4	BL5	BL6	BL7	BL8	BL9	BL10	BL11	BL12	Quant	Class
B2	B6	B0	B0	B0	B0	B0	B0	B0	B0	B0	B0	2	Ja
...	...	...	...	...	...	...	...	...	...	...	...	...	...
B4	B2	B0	B0	B0	B0	B0	B0	B0	B0	B0	B0	2	Ne
B4	B2	B3	B0	B0	B0	B0	B0	B0	B0	B0	B0	3	Ja

Figure 5. Time series discretization and data transformation

By combining the results into one data set discrete time series with block numbers, the numbers of cluster blocks in the data set and the class that was assigned in classification are obtained. These data sets have to be of a specific length, maximum number of periods of a data set has to be divided by two because blocks are made of pairs. Missing values of a data set are replaced by «0». In order to use the created data set for rule induction, classification performance evaluation and transition point forecasting with *Weka* tool, letter indexes are put in front of the obtained values. Data transformation and discretization can be seen in Figure 5.

4.5. Classifier performance scoring with 3-fold cross-validation

To *evaluate performance of classification*, a measure of accuracy that characterizes the performance of classification, is needed. Clustering is performed using the *grid-clustering* method and *k-means algorithm* with different numbers of clusters. Classifier performance is measured using the *JRip* classifier of *Weka*. Performance is based on the technology of dividing data into training and test sets [1]. To evaluate classification performance *3-fold cross-validation* was used. That means that data set D is divided into three subsets (*D1, D2, D3)*. Division has a rule – membership to any of the subsets is random but the size of the subsets has to be equal. Cross-validation guarantees that every record will be used for training (n-1) times and once for testing [1]. Sometimes a given evaluation can be biased and true positive rate, which shows the proportion of positive cases that were correctly identified, must be used.

By evaluating classifier performance with different clustering approaches, the best method to forecast transition points is chosen (the choice is made on the basis of aggregated

confusion matrixes). The accuracy of the classification task (KKP) is described by proportion of the aggregated true positive and true negative rates and total classified records, which is determined for each cluster group using equation (1)

$$KKP = \frac{IP + IN}{IP + KN + KP + IN}, \tag{1}$$

where: IP – true positive (positive (Ja) cases that were correctly identified (Ja)); KN – false negative, (positive (Ja) cases that were incorrectly classified as negative (Ne)); KP – false positive, (negative (Ne) cases that were incorrectly classified as positive (Ja)); IN – true negative, (negative (Ne) cases that were classified correctly (Ne)).

Total classification error (KKK) is calculated by subtracting the accuracy from 1 or aggregated false positive and negative records divided by total number of classified records using the equation (2), but the calculation of total classification error for transition points «P1» and «P2» can be seen in Figure 6.

$$KKK = 1 - KKP \text{ or } KKK = \frac{KN + KP}{IP + KN + KP + IN}, \tag{2}$$

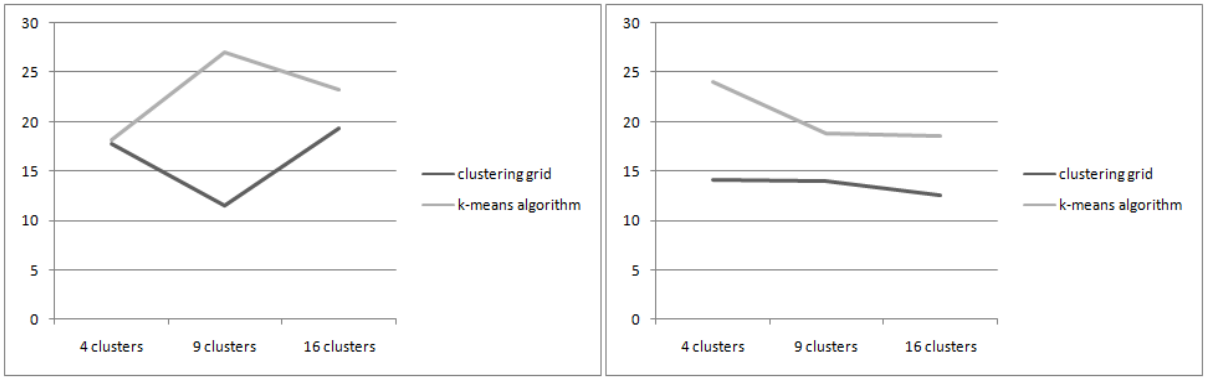


Figure 6. Total classification error for transition points «P1» (on the left side) and «P2» (on the right side)

Classification error for the class «Ja» can be determined by the positive records that were classified as negative and that are described by true positive rate (*IPN*) which can be calculated using the equation (3), true positive rates (IP) for transition points «P1» and «P2» are shown in Figure 7.

$$IPN = \frac{IP}{IP + KN}, \tag{3}$$

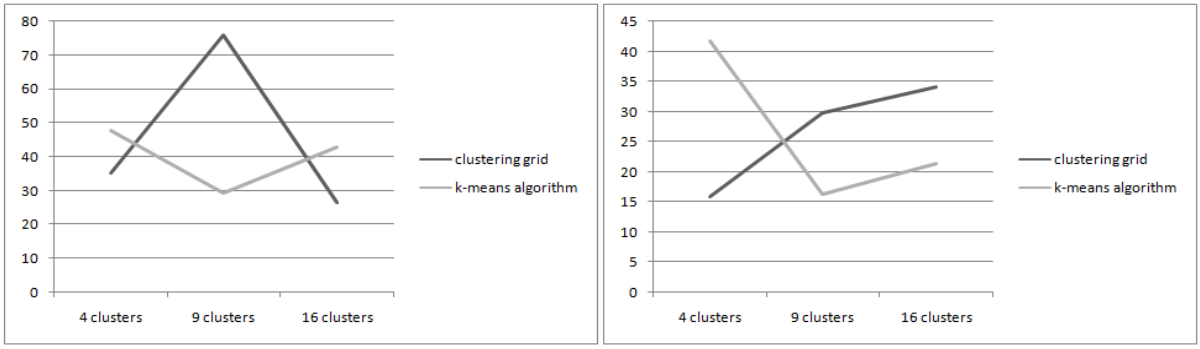


Figure 7. True positive rates for transition points «P1» (on the left) and «P2» (on the right)

5. Conclusions

After evaluating classification performances, it can be concluded that for forecasting transition point «P1» for data sets used in this study it is better to use grid-clustering method with 9 clusters, and for forecasting transition point «P2», grid-clustering method with 16 clusters is recommended. Since the classifier's performance is evaluated, it is important to evaluate new classifier's performance when working with new time series to reduce forecast error that can be caused by specific nature of time series and the choice of clustering algorithms.

References

1. Datu ieguve: Pamati / A. Sukovs, L. Aleksejeva, K. Makejeva [u.c.]. – Rīga: SIA Drukātava, 2007.
2. Das, G., Lin, K., Manilla, H., Renganathan, G., Smyth, P. Rule Discovery from Time Series. // Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining, 1998, p. 16-22.
3. Sukov, A. Recognition of Transitions between Different Phases of the Product Life Cycle //Proceedings of the 20th European Modelling and Simulation Symposium (EMSS 2008) September 17-19, 2008, Briatico, Italy. – Rende (CS), Italy, 2008. – p.81-86.
4. Han, J., Kamber, M., Tung, A.K.H., Spatial Clustering Methods in Data Mining – Canada: School of Computing Science, Simon Fraser University Burnaby, BC Canada V5A 1S6. Retrieved 31 March, 2008 from: http://www-faculty.cs.uiuc.edu/~hanj/pdf/gkdbk01.pdf.
5. Komninos, I., Product Life Cycle Management. Thessaloniki: Aristotle University of Thessaloniki, 2002. Retrieved 5 April 2008 from: http://www.urenio.org/tools/en/Product_Life_Cycle_Management.pdf.
6. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. Барсегян, М. Куприянов, В. Степаненко – Санкт-Петербург: БХВ- Петербург, 2007.

Kiršners Arnis, Sukovs Anatolijs. Pārejas punktu prognozēšanas likumu indukcija produkta dzīves cikla datos

Dotajā rakstā tiek aprakstīta netradicionāla pieeja problēmas risinājumam, kas saistīta ar jauna produkta dzīves cikla prognozēšanu. Šī pieeja ir vērsta uz pārejas punktu prognozēšanu produkta dzīves ciklā. Raksta aktualitātes pamatā ir risku samazināšana, pieņemot lēmumus, kas saistīti ar jauna produkta dzīves cikla prognozēšanu nākotnē, analizējot līdzīgu produktu vēsturiskus pieprasījuma datus. Raksta autori ir analizējuši vēsturiskus produkta pieprasījuma datus par diviem gadiem. Par raksta teorētisko pamatojumu kalpo produkta dzīves cikla un laika rindu analīzes metodes. Rakstā tiek izmantotas datu pirmapstrādes, normalizācijas, klasifikācijas, klasterizācijas un diskretizācijas metodes. Klasifikācijas veiktspēja un īstās pozitīvās klases atpazīšana tiek vērtēta ar 3-kārtīgu šķērsvalidāciju, pielietojot rīku Weka – tājā iebūvēto klasifikatoru JRip, klasterizācijai izmantojot klasterizācijas režģa metodi un k-vidējo sadalošo algoritmu pie dažādu klasteru skaita. Eksperimentāli tiek noteikta labākā klasterizācijas metode un klasteru skaits pārejas punktu prognozēšanai.

Kirshners Arnis, Sukov Anatoly. Rule induction for forecasting transition points in product life cycle data

This paper describes a novel approach to solve problems related to new product lifecycle. The approach is based on search for transition points on the curve of product life cycle. The topicality of the study is based on the decrease in risks in decision-making process. This new approach becomes even more important if decisions are connected with new product life cycle forecasting in the market. The authors have analyzed historical demand data for a two-year period. Methods of product life cycle and time series analysis are the theoretical basis of the study; data pre-processing and normalization methods are also used. Data classification, clustering as well as discretization of the available data are made. Classification performance is evaluated by the method of three-fold cross – validation using Weka's built-in classifier, JRip. Data clustering is made using grid-clustering

method and the k-means algorithm at different numbers of clusters. The most appropriate clustering method and the number of clusters for forecasting transition points P1 and P2 are determined experimentally.

Киршнерс Арнис, Суков Анатолий. Индукция правил для прогнозирования точек перехода в данных жизненного цикла продукта.

Данная работа описывает нетрадиционный подход к решению проблемы прогнозирования жизненного цикла нового продукта. Этот подход основан на поиске точек перехода на кривой жизненного цикла продукта. Актуальность работы основывается на снижении рисков при принятии решений, связанных с прогнозами жизни того или иного продукта на рынке в будущем, анализируя исторические данные спроса на схожие продукты. Авторы работы проанализировали исторические данные спроса на товары за двухлетний период. Теоретическим основанием работы служат методы анализа жизненного цикла продукта и анализа временных рядов. В работе также использованы методы первичной обработки данных, нормализации, классификации, кластеризации и дискретизации имеющихся данных. Оценка эффективности классификации и нахождения положительного класса проведена при помощи метода трехкратной перекрестной валидации, используя программное обеспечение Weka, с встроенным классификатором JRip. Кластеризация данных проведена с помощью метода кластеризационной решетки и алгоритма k-средних при разном числе кластеров. Экспериментально выведен наилучший метод кластеризации и число кластеров для прогнозирования переходных точек P1 и P2.