RIGA TECHNICAL UNIVERSITY
Department of Computer Science and Information Technology
Institute of Information Technology

Arnis Kiršners
Student of the doctoral study program "Information Technology"

A SYSTEM FOR PROCESSING SHORT TIME SERIES AND THEIR CHARACTERISTIC PARAMETERS IN FORECASTING TASKS

Summary of the Doctoral Thesis

Scientific supervisor
Dr.habil.sc.comp., Professor
A. BORISOVS

RTU Press
Riga 2015

Kiršners A. A System for Processing Short Time Series and their Characteristic Parameters in Forecasting Tasks. Summary of the Doctoral Thesis. – R.: RTU Press, 2015. – 44 pages.

Printed according to the decision of the Institute of Information Technology, Faculty of Computer Science and Information Technology, Riga Technical University Board meeting on June 19, 2015, protocol No. 12100-4.1/4.

This work has been supported by the European Social Fund within the project „Support for the implementation of doctoral studies at Riga Technical University".

ISBN 978-9934-10-765-8

THE DOCTORAL THESIS
PROPOSED TO RIGA TECHNICAL UNIVERSITY FOR THE PROMOTION TO THE SCIENTIFIC DEGREE OF DOCTOR OF ENGINEERING SCIENCE

To be granted the scientific degree of Doctor of Engineering Sciences (Information Technology), the present Doctoral Thesis has been submitted for the defence at the open meeting at the Faculty of Computer Science and Information Technology, Riga Technical University, 1 Setas Street, auditorium 202, at 14^{30}, on February 17, 2016.

OFFICIAL REVIEWERS

Professor, Dr.habil.sc.ing. Jānis Grundspeņķis
Riga Technical University, Latvia

Professor, Dr.sc.ing. Egils Stalidzāns
Latvia University of Agriculture, Latvia

Professor, Dr.sc.comp. Pavel Ošmera
Brno University of Technology, Czech Republic

DECLARATION OF ACADEMIC INTEGRITY

I hereby declare that I have developed this thesis submitted for the doctoral degree at Riga Technical University. I confirm that this Doctoral Thesis has not been submitted to any other university for the promotion of other scientific degree.

Arnis Kiršners …….
signature

Date …….

The doctoral thesis is written in Latvian and includes introduction, 4 sections, result analysis and conclusions, list of bibliography, 2 appendices, 25 tables, 53 figures, overall – 144 pages. The bibliography contains 77 references.

CONTENTS

OVERALL DESCRIPTION OF THE THESIS

Background

As the field of Information Technology keeps advancing and coming into everyday life, the demand for its application in various scopes is growing. Information technologies are being increasingly often used in solving of complex tasks, which had previously only solved by a human expert. The use of information technology allows obtaining new information and performing analysis of historical information, which allows acquiring new knowledge that can be used in solving of complex tasks. The growing intensity of use of this technology increases the demand for obtaining fast and accurate decisions – forecast – in the shortest time possible. There are several fields, like sales, medicine and pharmacology, which use data in the form of time series that span across a short period of time and are obtained in interaction with the analysed object, as well as descriptive information about the object. To process these different data structures, it is necessary to develop a system that would be able to forecast events in the future based on data mining methods and algorithms.

Problem Statement

There are fields where experts operate with data in the form of short time series and their descriptive parameters. Short time series describe functional changes of an object in a period of time, whereas descriptive parameters represent features of the object. For example in healthcare a patient is given medicine to lower blood pressure and then gets his blood pressure measured every minute for 20 minutes. In this case the blood pressure measurements over time would be short time series and the height, weight, gender, age and other features of the patient would be descriptive parameters. This and other fields ask for solution of forecasting tasks (e.g., how the given medicine will influence blood pressure of a patient) using only descriptive parameters of the object (e.g., patient) to obtain the forecast. The presence of such data with different structure sets forth the following requirements towards the forecasting system to be developed:

- It should be able to process discrete and continuous data.
- While analyzing short time series, the clustering process has to include a step to determine the most suitable number of clusters for data set processing using clustering error calculations and determine association of the analyzed object with one of the clusters.
- The obtained clustering results should be merged with the descriptive parameters of the short time series while maintaining data integrity so they can be used in classification.
- The process of classification has to determine the relationship of the object clustering class and their descriptive parameters while maximizing classification accuracy, sensitivity and specificity.
- The classification results should be interpreted according to task set forth in the field.
- Classification of a new object should be carried out based on the developed system, which would also determine the future value of the forecast.

Research Object and Subject

Research object is forecasting system. **Research subject** is data mining and machine learning methods and algorithms.

Research Methods

The research for the Thesis is based on data analysis and data mining methods. Data pre-processing step uses *z-score normalization using standard deviation* and demand *normalization using life curve*. Attribute informativeness evaluation is carried out using *cfs subset evaluation* method and attribute search – using *BestFirst* method. Cluster analysis is realized using *k-means divisive, modified k-means divisive* proposed by the author, *Expectation-Maximization* and *hierarchical agglomerative* algorithms. Classification algorithms include *ZeroR*, *OneR*, *k-nearest neighbours*, *CN2*, *C4.5*, *Naive Bayes* and *JRip*. Classification accuracy evaluation is carried out using *10-fold cross-validation*, *leave one out* and *data stratification into training and test sets with 70:30 ratio*. The obtained classifiers are used to obtain conditional rules in connection with results of research carried out in the actual field. Transformation of clustering result class structure into the class structure used in the field is carried out using an approach developed by the author. Classification results are transformed into conditional rules, which characterize the group of an object, class determined in classification and value used in field research. A forecast for an object is calculated using mathematical expectation and a distance-based approach developed by the author.

Research Limitations

The limitations of the Thesis are related to the short time series analysis, which puts additional requirements towards the forecasting system to be developed by limiting the range of methods and algorithms. The forecasting system also has to be able to merge different data structures while still being able to analyze them using data mining methods and algorithms.

Research Goal and Tasks

The goal of the Thesis is to develop a system for forecasting tasks that works with short time series and their descriptive parameters using data mining methods and algorithms and that can be used in various fields. To reach the goal, the following **tasks** were defined for the research:

1. Analyze processing principles of short time series and their descriptive parameters.
2. Carry out analysis of short time series and their descriptive parameter pre-processing approaches.
3. Adjust and modify clustering algorithm to match the scope of processing of short time series and compare it to other clustering algorithms.
4. Develop an approach for merging the obtained clustering results and descriptive parameters of short time series.

5. Develop forecasting systems for different fields, which would process short time series and their descriptive parameters in order to make a forecast based only on newly entered descriptive parameters of an object.
6. Evaluate the accuracy of the developed forecasting system in various fields.
7. Develop approaches for conditional rule generation and use in different fields.
8. Develop guidelines for development of forecasting systems based on the developed forecasting system for various fields.

Research Hypotheses

The following hypotheses were proposed for this research:

1. Development of a system for processing of short time series and their descriptive parameters provides a solution for the difficult formalized task using data mining methods and algorithms.
2. *Modified k-means divisive* algorithm improves the determination of the most suitable number of clusters during clustering process when analyzing short time series.
3. The developed data processing system realizes forecasting task solution in various fields.

The first hypothesis points to data structures where data sources are short time series and their descriptive parameters. The hypothesis will be considered proven to be correct if the resulting system, which will be developed, will allow processing these differing data structures and make a prognosis as a result.

The second hypothesis points to comparative analysis of several clustering algorithms in various fields, evaluating how they can process data with short time series. If the modified clustering algorithm will show better results the hypothesis will be considered proven to be correct.

The third hypothesis is based on the idea that the developed system can be used in or adapted to various fields. The hypothesis will be considered proven to be correct if the developed system will be used to solve forecasting tasks in various fields.

Research Actuality and Scientific Novelty

The actuality of the Thesis is related to analysis of a different data structure type (short time series and their descriptive parameters). There is no known method or algorithm that could carry out analysis of short time series and their descriptive parameters. Therefore it is important to assess a set of approaches that use data mining methods and algorithms in order to process short time series and their descriptive parameters. It is proposed to carry out the processing of short time series using clustering to determine groups of similar objects. The descriptive parameters are processed using classification in order to find relationships between these parameters and the results of clustering. The prospective value of feature of a new object is determined by classifying the descriptive parameters of this object based on the obtained classifier.

The scientific novelty of the Thesis is based on the developed forecasting system for various fields, which realizes processing of short time series and their descriptive parameters. The developed system can analyze complex data structures. For this system:

1. A modified k-means divisive algorithm was developed, which provides processing of short time series in various fields and determines the most suitable number of clusters based on the mean absolute error of clustering.
2. An approach for merging two different data structures was developed.
3. Classification result visualization for various fields and tasks was developed.
4. Conditional rule construction and application approaches were developed for various fields.

Practical Value

The practical value of the Thesis is the developed demand forecasting system that realizes the forecast of possible demand for a new product based solely on the parameters describing the product. This system can be used in companies that need forecasts of possible demand of a product for future periods of time.

Also the developed heart necrosis risk forecasting system that determines risk of heart necrosis for a laboratory animal based on the descriptive information about this animal. The developed system can be used in research organizations that use laboratory animals in order to forecast the influence of a drug.

And the developed bacteria proliferation syndrome detection system that determines if an individual has a necessity to take lactose test by entering only health self-assessment parameters. The developed system can be used in healthcare – gastroenterology – to detect bacteria proliferation syndrome in small intestine and determine if a patient needs a lactose test. Accuracy of the system has been evaluated in various fields.

The system served as basis to develop forecasting system guidelines that contain recommendations for a developer concerning development of similar systems in various fields.

Approbation

During the research for this Thesis the following 13 articles have been written and published:

1. Parshutin S., Kirshners A. Research on Clinical Decision Support Systems Development for Atrophic Gastritis Screening// *Expert Systems with Applications.* – 2013. – Vol.40, Iss.15, pp. 6041–6046. Cited in: ScienceDirect, SCOPUS, Thomson Reuters ISI Web of Science.
2. Kirshners A., Parshutin S. Application of Data Mining Methods in Detecting of Bacteria Proliferation Syndrome in the Small Intestine // In: *European Conference on Data Analysis 2013: Book of Abstracts: European Conference on Data Analysis 2013.* – 2013. – pp. 139-139.
3. Kirshners A., Parshutin S., Leja M. Research in application of data mining methods to diagnosing gastric cancer// LNAI 7377. Proceedings of the 12th Industrial Conference on Data Mining ICDM'2012. – 2012. – pp. 24–37. Cited in: SpringerLink, SCOPUS.

4. Kirshners A., Liepinsh E., Parshutin S., Kuka J., Borisov A. Risk Prediction System for Pharmacological Problems// Automatic Control and Computer Sciences. – 2012. –Vol. 46, No.2. – pp. 57–65. Cited in: SpringerLink, SCOPUS.
5. Kirshners A., Borisov A. A Comparative Analysis of Short Time Series Processing Methods// Scientific Journal of Riga Technical University, Information Technology and Management Science, 2012. – Vol.15. – pp. 65–69. Cited in: VINITI, EBSCO, CSA/ProQuest.
6. Kirshners A., Borisov A., Parshutin S. Robust Cluster Analysis in Forecasting Task// Proceedings of the 5th International Conference on Applied Information and Communication Technologies (AICT2012). – 2012. – pp. 77–81.
7. Parshutin S., Kirshners A. Intelligent Agent Technology in Modern Production and Trade Management// Efficient Decision Support Systems: Practice and Challenges – From Current to Future/ Book Chapter. INTECH. – 2011. – pp. 21–42. Cited in: NetLibrary; Scirus; IntechOpen; WorldCat.
8. Kirshners A. Clustering-based Behavioural Analysis of Biological Objects// Environment. Technology. Resources: Proceedings of the 8th International Scientific and Practical Conference. – 2011. – Vol.2. – pp. 24–32. Cited in: SCOPUS.
9. Kirshners A., Borisov A. Multilevel Classifier Use in a Prediction Task// Proceedings of the 17th International Conference on Soft Computing. – 2011. – pp. 403–410. Cited in: Thomson Reuters ISI Web of Science.
10. Kirshners A., Borisov A. Processing short time series with data mining methods// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2011. – Iss.5, Vol.49. – pp. 91–96. Cited in: VINITI, EBSCO, CSA/ProQuest.
11. Kirshners A., Parshutin S., Borisov A. Combining clustering and a decision tree classifier in a forecasting task// Automatic Control and Computer Science. – 2010. – Vol.44, No.3. – pp. 124-132. Cited in: SpringerLink, SCOPUS.
12. Kirshners A., Borisov A. Analysis of short time series in gene expression tasks// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2010. – Iss.5, Vol.44. – pp. 144–149. Cited in: VINITI, EBSCO, CSA/ProQuest.
13. Kirshners A., Kuleshova G., Borisov A. Demand forecasting based on the set of short time series// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2010. – Iss.5, Vol.44. – pp. 130–137. Cited in: VINITI, EBSCO, CSA/ProQuest.

The results of the research have been presented at the 9 following conferences:

1. *European Conference on Data Analysis 2013,* Luxemburg, Luxemburg, July 10–12, 2013.
2. *53rd International Scientific Conference of Riga Technical University,* RTU, Riga, Latvia, October 10–12, 2012.
3. *12th Industrial Conference on Data Mining ICDM'2012,* Berlin, Germany, July 13–20, 2012.
4. *5th International Conference on Applied Information and Communication Technologies AICT2012*, LUA, Jelgava, Latvia, April 26–27, 2012.
5. *52nd International Scientific Conference of Riga Technical University,* RTU, Riga, Latvia, October 12–25, 2011.
6. *8th International Scientific and Practical Conference*, Rezekne, Latvia, June 20–22, 2011.

7. *17th International Conference on Soft Computing*, Brno, Czech Republic, June 15–17, 2011.
8. *Information Technology: Knowledge and Practice, University of Latvia*, Riga, Latvia, December 7, 2010.
9. *51st International Scientific Conference of Riga Technical University,* RTU, Riga, Latvia, October 11–15, 2010.

Research results have been used in the following projects:

06.12.–03.13 – RTU research project No. ZP-1688 Young scientists in 2012/2013 "Research in designing decision support system for gastric cancer diagnosing support". Head, S. Parshutin.
01.10.–12.12. – European Social Fund Project "Interdisciplinary research group for early cancer detection and cancer prevention", Nr. 2009/0220/1DP/1.1.1.2.0/09/APIA/VIAA/016. Head, M. Leja.
06.10. – 12.11. – LATVIA – BELARUS Co-operation programme in Science and Engineering, Scientific Cooperation Project Nr. L7631. "Development of a complex of intelligent methods and medical and biological data processing algorithms for oncology disease diagnostics improvement". Head, Professor A. Borisov.

Structure and Contents of the Thesis

The Thesis contains Introduction, four Sections, Conclusions, Bibliography and appendices.
The **First section** describes the processing principles of short time series and their descriptive features. It gives a short insight into data mining and theoretic material about short time series and their descriptive features, as well as substantiation for algorithm selection for implementation of forecasting task solving and formal task description.
The **Second section** describes the applied methods and algorithms that are used in development of forecasting systems. This section also introduces the modification of k-means divisive algorithm that was developed by author and is intended for clustering of short time series with different numbers of objects.
The **Third section** proposes forecasting system for various fields that is used as a basis for: demand forecasting system, heart necrosis risk forecasting system and bacterial proliferation syndrome detection system. Construction and work principles of these systems are described. Experiments and their results are also presented. The section also describes the evaluation of accuracy of these systems and provides conclusions about the developed forecasting systems.
The **Fourth section** provides a description about the guidelines for forecasting system development, which help the developer to choose the best development process of the system for processing short time series and their descriptive parameters. System development process is based on the experience acquired during development of the three applied systems for different fields.
The **Last section** provides analysis of the obtained results and conclusions about the developed systems and guidelines.

SUMMARY OF THESIS CHAPTERS

1. PROCESSING PRINCIPLES OF SHORT TIME SERIES AND THEIR DESCRIPTIVE PARAMETERS

This section gives a short insight into data mining [2, 3, 4, 5, 44, 45, 52, 58, 59, 62, 75]. It describes tasks solved in data mining and the process of new knowledge acquisition [76]. It describes processing of short time series and their descriptive parameters using data mining methods and algorithms [7, 9, 11, 21, 22, 29, 31, 32, 33, 36, 42, 53, 54, 55, 57, 63, 64, 72, 70, 76].

Selection of Clustering Algorithms for Analysis of Short Time Series

A comparative analysis was carried out to select the most suitable clustering algorithms for short time series analysis using several criteria: does the algorithm belong to the list of 10 most popular algorithms [73]; interpretability of results obtained using this algorithm, where a positive rating was given to algorithms, whose results are interpretable without assistance of an expert and their implementation is available in the software used in experiments. The criteria for comparisons are evaluated using two values: positive rating (+) or negative rating (–). Evaluation results of comparative analysis of clustering algorithms are displayed in Table 1. The sum of positive ratings is given in column *Evaluation*.

Table 1

Comparative analysis of clustering algorithms

	Comparative criteria, possible values (+ or -)					
Clustering algorithms	Is in *Top10* algorithms (place in top)	Interpretability of results	Software for realization			Evaluation
			Weka	*Orange Canvas*	*Statistica*	
k-means divisive	+(3)	+	+	+	+	**5**
Expectation maximization (EM)	+(5)	+	+	–	+	**4**
Self-organizing neural networks (SOM)	–	–	–	+	–	1
Hierarchical agglomerative	–	+	+	+	–	3
C-means divisive	–	+	–	–	–	1

The comparative analysis shows that the most suitable algorithms for short time series analysis are *k-means divisive* and *expectation maximization* algorithms. Some experiments should be carried out using hierarchical agglomerative algorithm, because this algorithm is the next behind two selected algorithms.

Selection of Classification Algorithms for Descriptive Parameter Analysis

The selection of classification algorithm has to be carried out according to interpretability of the obtained results – if the result should be comprehensible for non-experts. Who will use the obtained results – an expert or a regular system user? How quickly should the algorithm result be obtained? Is the algorithm suitable for the available data structure? The answers to these questions point to several classification algorithms that can be analyzed experimentally to determine the best algorithm for solution of task. The selection of classification algorithms is carried out using comparative analysis that is displayed in Table 2.

Table 2

Comparative analysis of classification algorithms

	Comparative criteria, possible values (+ or –)				
Classification algorithms	Is in *Top10* algorithms (place in top)	Result interpretability (used method)	Software for realization		Evaluation
			Weka	*Orange Canvas*	
C4.5	+(1)	+ (inductive decision trees)	+	+	**4**
k-nearest neighbours (kNN)	+(8)	+ (distance metric)	+	+	**4**
Naive Bayes	+(9)	+ (probabilities)	+	+	**4**
CN2	–	+ (conditional rules)	+	+	**3**
OneR	–	+ (conditional rules)	+	–	2
ZeroR	–	+ (conditional rules)	+	–	2

Definition of Tasks

There are field data given that describe information about historical events and their descriptive parameters. The proposed data mining methods and algorithms are used to solve a forecasting task, whose output should be a value of the new object in the future that is based solely on the descriptive parameters of this object.

The use of data mining methods and algorithms should determine the most suitable that can be used to solve the tasks put forth in the field. The set of data mining approaches, which can be used for solution of specific tasks, is as follows:

1. Pre-processing of historical events (short time series) and their descriptive parameters (attributes).
2. Detection of relationships in historical data by creating a clustering model, which is used to determine similar groups of objects – clusters. Depending on the task put forth by the field research it is possible to create a visualization of the obtained similar object groups using prototype.

3. Transformation of the similar object group number to the number that is used in the dependent descriptive parameter variable (class).
4. Assessment of relationships between descriptive parameters and clustering results using classification.
5. Interpretation of classification results using conditional rules.
6. Evaluation of clustering and classification accuracy.
7. Determination of feature of a new object using only descriptive parameters of this object.

Formal Definition of the Task

The process of short time series processing in the Thesis is defined as a clustering task, whose goal is to determine groups of similar objects, which could be used to group objects in the analyzed data set. To test the suitability of the selected methods, it is necessary to carry out comparative analysis of short time series clustering algorithms and select the most suitable. There are clustering tasks where the selected methods provide only a partial solution to this task. To completely solve these tasks it is necessary to analyze the selected algorithms and develop their modifications for short time series clustering. Clustering uses unsupervised training.

The processing process of short time series descriptive parameters is defined as classification task, whose goal is to find relationships between descriptive parameters, their values and target attributes – classes – in supervised training. A testing process is used to evaluate the accuracy of the used classification algorithm. And the obtained classifier (classification model) is used to forecast the target attribute of the new object based on the model obtained in the training process.

Theoretical Model of the Forecasting System

Based on the formal definition of the task, data structure of the analyzed data and literature analysis, a theoretical model of the forecasting system is proposed based on data mining methods and algorithms (see Figure 1). The analyzed data that consists of two data types – short time series and their descriptive parameters – should be pre-processed. Data pre-processing involves removal of objects with missing values from the data set. Data normalization is used to avoid dominance of attributes in a data set that occurs from different ranges of attribute values (e.g., attribute *Age* and attribute *Income*) [1]. Descriptive parameters are evaluated for their information, which helps removing uninformative attributes from a data set, which have a negative impact on accuracy of classification [27]. In the process of clustering objects with short time series are clustered together into groups – clusters – based on their similarity using clustering algorithm [64]. The obtained clustering results are merged with descriptive parameters (attributes) and the class determined in clustering using one of classification algorithms [27, 64].

Determining the target attribute value or class (forecasting task) of a new object is carried out using solely descriptive parameters of this object and using the constructed classification model [27, 58, 64, 70].

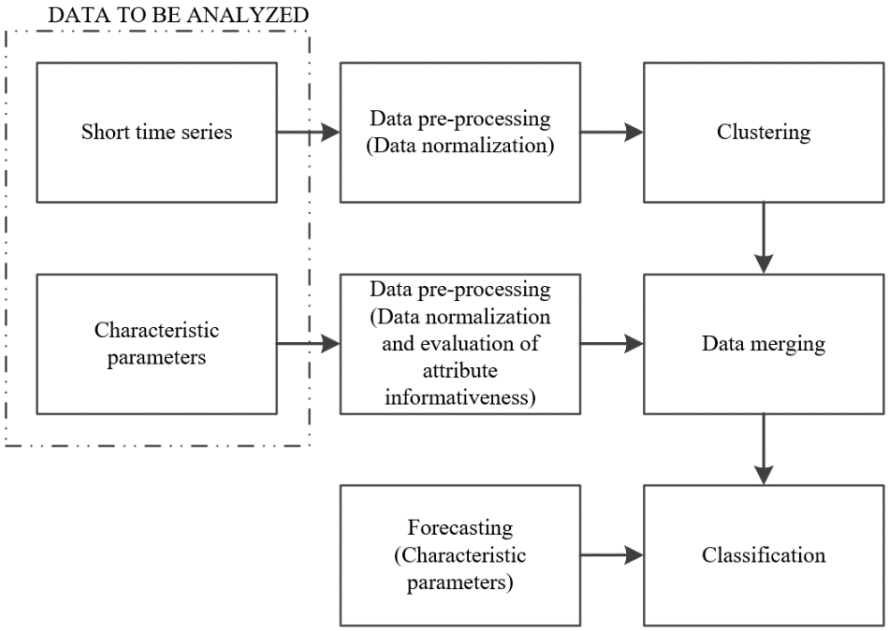

Figure 1. Theoretic model of forecasting system

2. REVIEW OF METHODS AND MODIFICATIONS USED IN THE DEVELOPMENT OF THE FORECASTING

This section describes the data mining methods, algorithms and algorithm modifications used in the development of the forecasting system and the principles of their functioning. Description is given for data pre-processing techniques used in system development: data cleaning and data transformation [1, 27, 57, 58, 63, 64, 76]. Information is given about data clustering, giving descriptions of algorithms like *k-means divisive* [28, 63], *expectation maximization* [17, 51], *hierarchical agglomerative* [63] and *modified k-means divisive* algorithm modification developed by the author.

The *modified k-means divisive* algorithm was developed based on the experiment results with short time series [30, 35] in various fields, which showed that, for example, the classic *k-means divisive* algorithm cannot determine the most accurate number of clusters for analysis of short times series [34, 37]. The results obtained with *k-means divisive* algorithm showed a unified trend – the increase in number of clusters decreases the overall clustering error, whereas *expectation maximization* algorithm shows the smallest error for the minimum number of clusters. Therefore a *modified version of k-means divisive* algorithm was developed, which allows avoiding these shortcomings.

Initially the maximum cluster number property is defined that will be used for clustering of the data set. This approach increases the calculation speed of the algorithm unlike the classical k-means divisive algorithm where the cluster number, until which the algorithm divides objects, is set manually [73]. Carrying out clustering with one of the algorithms, one has to determine the range of cluster numbers (from minimum until maximum number, where minimum is usually 2), in order for the clustering to be effective and less time-consuming. The optimal number of clusters has to be determined in the range from 2 until maximum number of clusters.

This maximum has to be large enough to accurately carry out clustering (can be determined using average absolute error or squared error) but it cannot be too large. In such case there is a chance of wrong result interpretation because, for example, a natural cluster is further divided into smaller clusters. Therefore the maximum number of clusters C_{max} [64, 70] can be calculated using the theoretical assumption that is calculated using the formula $C_{\max} = \sqrt{n},$ where n is the number of objects in the analyzed data set. The rest of the steps of the modified k-means divisive algorithm use the steps of *k-means divisive* algorithm until the algorithm reaches the step of sum of squares error calculations, where there are other modifications. There is a distance matrix calculated for each cluster, which characterizes distance d_n (calculated using *Euclidean* distance measure) between an object (time series {T1, T2, …, T12}) c_n and the nearest centroid. The obtained distance matrix results are used to calculate clustering error for a training set using mean absolute deviation calculation for each cluster and the overall clustering error calculation. Based on the minimum calculated overall clustering error value for each of the clusters, the most suitable number of clusters that are necessary for clustering of the data set is determined. If the number of objects in the analyzed data set is smaller than 200, clustering accuracy evaluation can be done using 10-fold cross-validation [43] and the calculation of clustering error should be done according to the section "Clustering error calculation for training data set" because dividing a data set into training and testing sets would result in a small number of objects, which can negatively influence clustering results. Whereas, if the number of objects is larger than 200, accuracy evaluation can be carried out using data set stratification into training and test sets using 70 % to 30 % ratio [70]. In this case clustering error calculation has to be carried out according to section "Clustering error calculation for training data set". Prototypes have to be constructed for each obtained cluster according to section "Prototypes in clusters". Then test data set clustering has to be carried out according to section "Clustering of test data set" and clustering error for the test data set is calculated according to section "Clustering error calculation for test set".

Clustering Error Calculation for Training Data Set

Mean absolute error is calculated from absolute deviation. It is calculated based on Equation (1), which produces the value of mean absolute deviation AD_i:

$$AD_i = \frac{d_1 + d_2 + \ldots + d_n}{c_n}, \tag{1}$$

where $d_1, d_2, \ldots, d_n$ – distance between an object and the centroid;

c_n – number of objects in a cluster;

AD_i – mean absolute deviation in the i-th cluster.

Then mean absolute deviations for all clusters AD_i are summed together and divided by the number of clusters C_i that were obtained in the cluster analysis. This gives the clustering mean absolute error *MeanAE* [52] according to Equation (2):

$$MeanAE = \frac{AD_1 + AD_2 + \ldots + AD_i}{C_i}, \quad (2)$$

where C_i – number of obtained clusters in the data set;

MeanAE – clustering mean absolute error.

This approach implements the analysis of distances between all data set objects and centroids in each cluster and therefore ensures calculation of clustering mean absolute error. This is then used to choose the number of clusters among all clusters C_i with the smallest corresponding *MeanAE* value that will be used for clustering of the data set.

Prototypes in Clusters

Prototype is constructed for each cluster that is obtained in the process of clustering. The number of prototypes depends on the selected most suitable number of clusters. The prototype curve that is constructed based on mean values of objects at each period of time in the 6th cluster is displayed in Figure 2; the grey lines show objects of this cluster. The x axis shows period numbers, the y axis shows the normalized values of short time series. The obtained prototypes describe the behaviour of class objects at a certain period of time.

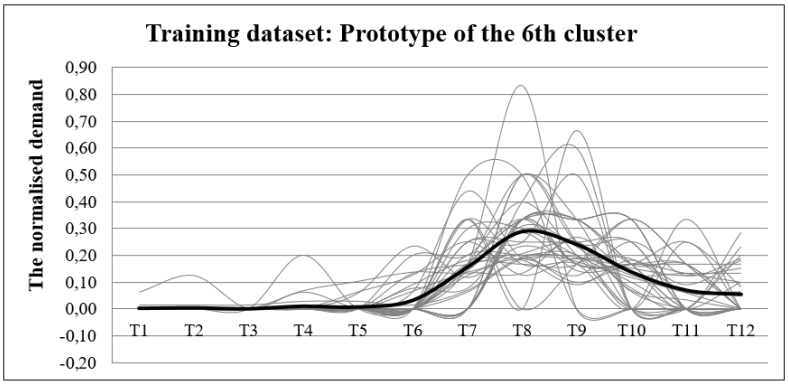

Figure 2. Prototype of the 6th cluster (drawn with bold line)

Clustering of Test Data Set

Clustering of test data set is based on the mean prototype values obtained in the clustering process of training data set and test set data (short time series). Distance between an object and the mean value of the corresponding cluster is calculated according to Equation (3). The obtained results are registered in a distance matrix, which is depicted in Table 3.

During clustering of the test data set, each object is assigned to one cluster based on the minimum distance [70]. Every test data set object is examined period after period and its distance to each cluster centroid (prototype of the training set) is calculated according to Equation (3).

$$d_{i,j} = \sqrt{\left(P_1 - C_1\right)^2 + \left(P_2 - C_2\right)^2 + \ldots + \left(Pm_n - Cz_n\right)^2}, \quad (3)$$

where $d_{i,j}$ – distance between a test data set object *i* and prototype *j*;

P_m – value of test data set object m in time period n;

C_z – mean value of prototype z obtained from training set in time period n.

When distance metric calculations for test data set objects are completed (results shown in Table 3), the minimum value of each row d_i (each object) must be determined and registered in the table field *Minimum distance*.

Table 3

Distance matrix of test data set for assessment of membership to a cluster

Object number	Distance between object and the corresponding cluster centre (prototype)				Minimum distance	Assigned class based on minimum distance
Nr.	***C1***	***C2***	**…**	***Cj***	d_i**(min)**	***C(j)***
d_1	$d_{1,1}$	$d_{1,2}$	…	$d_{1,j}$		
d_2	$d_{2,1}$	$d_{2,2}$	…	$d_{2,j}$		
…	…	…	…	…		
d_i	$d_{i,1}$	$d_{i,2}$	…	$d_{i,j}$		

The obtained minimum distance, which is calculated for each test data set object, is used to assign cluster number *C(j)*, which points to the cluster that this object is assigned to.

Clustering Error Calculation for Test Set

Evaluation of test set clustering results is based on mathematical metrics. It uses calculations of mean absolute deviation *MAD* and *MAE* mean absolute error for each cluster. Then a sum of mean absolute error values for each cluster. This provides evaluation for independent data sets. *MAD* [52, 76] is calculated according to Equation (4):

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\left|\left(Cvid_{n_i} - P_{m_i}\right)\right|, \tag{4}$$

where $Cvid_n$ – the obtained mean value of training data set prototype of cluster n in time period i;

P_m – value of a real test data set object m in time period i;

N – number of periods in a time series.

MAE [52, 76] for cluster i, which shows the overall error of a cluster regarding the training set, is calculated according to Equation (5):

$$MAE_i = \frac{MAD}{k}, \tag{5}$$

where k – number of test data set objects in cluster i.

Whereas the total absolute error *TAE* [52, 76] for the test data with n clusters can be calculated based on Equation (6).

$$TAE = \frac{\sum_{i=1}^{n} MAE_i}{n}, \tag{6}$$

where MAE_i – mean absolute error of the i-th cluster.

When the total absolute error is evaluated, the test data set is clustered and the obtained results are validated against prototypes, which were obtained by clustering the training data set, one can conclude about accuracy of clustering results.

The thesis also examines data classification [1, 3, 12, 58, 65, 76] and describes classification algorithms like *C4.5* [58], *k-nearest neighbours* [63], *CN2* [14], Naive Bayes [63], as well as criteria for accuracy evaluation of these algorithms. It also describes attribute information measurements and the corresponding process using *Cfs Subset Evaluation* and *BestFirst* methods [70].

3. PROCESSING SYSTEMS FOR FORECASTING TASKS

Based on the defined problem, which is described by the theoretical model that is shown in Figure 3, the system will solve short time series and their characteristic parameter processing in various fields using data mining methods and algorithms. The proposed forecasting system for various fields is shown in Figure 3. The analyzed data set, which consists of short time series and their descriptive parameters, is analyzed in pre-processing. The short time series are clustered making specific object groups named clusters using a clustering algorithm modification proposed by the author. The obtained clusters are used to create prototypes that characterize mean values of cluster objects in each time period. The obtained clustering results (numbers of object clusters) are then merged with the pre-processed characteristic parameters in the data merging blocks. If it is necessary (based on task definition), the clustering results with a different number of classes are processed using class transformation, which is proposed by the author. If it is necessary to split the merged data set into subsets, the data splitting block is used, where the split is obtained using the splitting attribute and its values (defined by an expert). The number of values represents the number of data subsets H that will be created. The obtained data sets are used for classification by applying a classification algorithm, determining the relationships between characteristic parameters and the class attribute. Conditional rules are used for comparison of two classifier results or generation of a knowledge base, which would store rules about object classes, splitting attribute value in the data set and the result of clinical research (transformed into a numeric value or class).

The specifics of forecasting are based on the task. In the forecasting that uses prototypes a number has to be obtained for each object (its class). It is determined using a trained classifier by classifying characteristic parameters of the object that is used for forecast. The obtained number points to a prototype that characterizes the demand of the analyzed object. Another forecasting task classifies the characteristic parameters of an analyzed object using two classifiers and compares the obtained results using conditional rules.

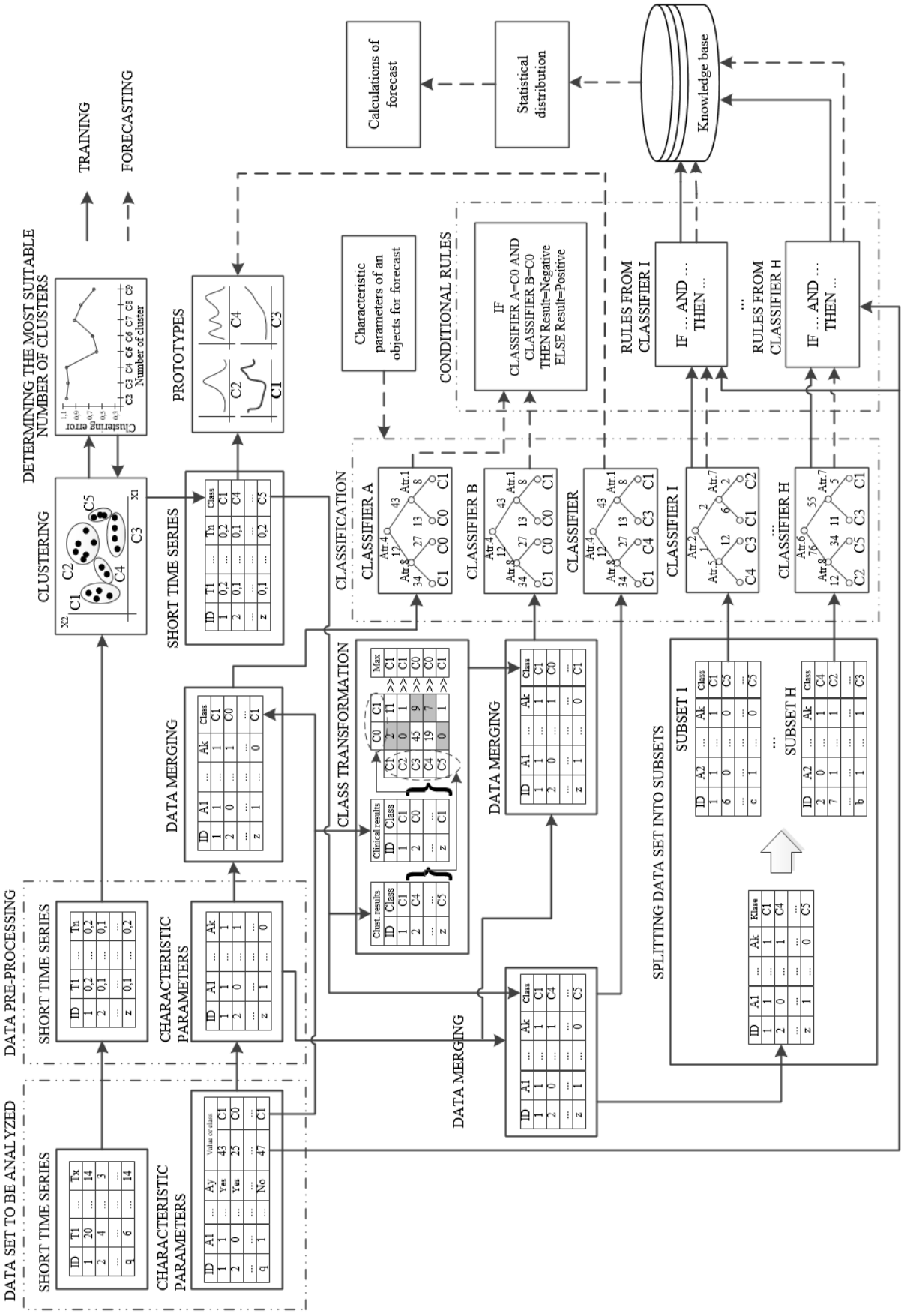

Figure 3. Forecasting system for various fields

Next approach creates temporary conditional rule that is obtained by classifying characteristic parameters of the analyzed object and determining it class, the other part of the condition is the value of the splitting attribute. Then the rules that apply to the created temporary conditional rule are filtered out of the knowledge base. The selected rules form a statistical distribution, which is then used to calculate the forecast value for the analyzed object using a method proposed by the author.

The proposed forecasting system for various fields is used as a theoretical model, which serves as a basis for systems for product demand forecast, heart necrosis risk forecast and bacteria proliferation syndrome forecast. The structure of these systems is described in the following sections.

3.1. Demand Forecasting System

Demand forecasting system (DFS) was developed for optimization of clothing company orders based on the product demand history of the previous years (short time series) and the characteristic parameters of a new product defined by the following attributes: price of the product, type, seasonality, life span of collection (in months), colour, size etc.

Structure of DFS

The structure of DFS is presented in Figure Figure 4. When the analysis is started (training process), the system has to receive the historical demand period (year={2006}) that is entered by a user – expert – based on his experience. After the period is entered, the system creates a query, which is sent to the data base using *PL/SQL* language.

The data base holds information about demand for products in various periods of time, product types, colours, prices, invoices, suppliers, clients, barcodes etc. The result of this query is a selected data set, which is then assigned to the data preparation block, where the data set is split into two data flows. The first is constructed from short time series data structure (product identifier and monthly demand for this product on annual basis) and the other data flow consists of the characteristic parameters (product identifier, category, type and price) that are then assigned to the forecasting system module.

In the forecasting system module the short time series are pre-processed. This results in data cleaning (decreasing noise by removing objects with noise-producing values) and normalization (equating value ranges). Clustering algorithm is used to detect relationships among short time series, which are accumulated into groups named clusters based on similarity measures. The most suitable number of clusters is determined by the clustering algorithm based on the smallest absolute error of clustering. Then a prototype is constructed for each obtained cluster, which describes the mean values of cluster objects in each period of time.

The second data flow is also pre-processed removing objects with missing values and normalizing attribute values of characteristic parameter data set. Then the most informative attributes are determined and the least informative attributes are removed from the data set.

Then the pre-processed characteristic parameters are merged with cluster numbers into one data set based on the cluster to which an object was assigned in the clustering process. Merging is

carried out based on the identification numbers of objects. The merged data set is assigned to the classification process, which uses inductive decision trees. This results in determined relationships between the prototype (cluster number) and the characteristic parameters. After the training process is completed, a forecast can be created for demand of a new product. The characteristic parameters of a new product (e.g., category={2}, type={6}, price={1,00}) are entered into the system. The inductive decision tree created in classification process is then used to project (forecasting process) the characteristic parameters of the new object onto the tree and determine the class of this product (e.g., 'C1'). Projection of data is carried out level by level moving from the root of the tree until a leaf node (lowest level) is reached and a class value can be determined. The obtained class (e.g., 'C1') points to the corresponding prototype, which was constructed during clustering.

The prototype (C1) describes the potential future demand for 12 months (periods from T1 till T12) for the product, whose characteristic parameters were entered into system for the forecasting process [40].

DFS Experiment Results

Experimental validation of DFS was carried out using real historical demand data of clothing retail sales company for year 2005, which described 423 objects after data cleaning process and were used for model training, and 149 objects for year 2006, which were used for model testing. In order to use the data for comparison using the obtained prototypes, life span of the objects was reduced to 12 periods (months). The training and the test data sets also included characteristic parameters describing the objects (category, type and price of a product).

Normalization of historical demand training data set was carried out using *z-score normalization using standard deviation* that is generally used in data mining to even out dominating attribute values if the maximum value is not known. Another approach used for normalization is sales volume *normalization using life curve*, which has been used in studies of other authors [64] for normalization of similar data structures. Experimental evaluation using clustering with *k-means divisive* algorithm for ten clusters (experimentally determined as the most suitable number of clusters), different normalization approaches and 10-fold cross-validation, the following training errors were obtained: 3.28 for *z-score normalization using standard deviation* and 0.28 for *normalization using life curve*. The results show that the best accuracy was achieved using *normalization using life curve*. The clustering process for the training data set was carried out using *k-means divisive* and *expectation-maximization* algorithms. The results obtained with *k-means divisive* algorithm are presented in Figure 5, and the ones with *expectation-maximization* algorithm are presented in Figure 6.

Log-likelihood drop at the 17^{th} cluster (see Figure 6) is due to the small number of objects at this number of clusters, when compared to other cluster distributions, with Bayesian probability equal to 1. The obtained results show that the most suitable number of clusters in both cases is the maximum 21 because the mean absolute clustering error has to be minimized but the log-likelihood has to be maximized.

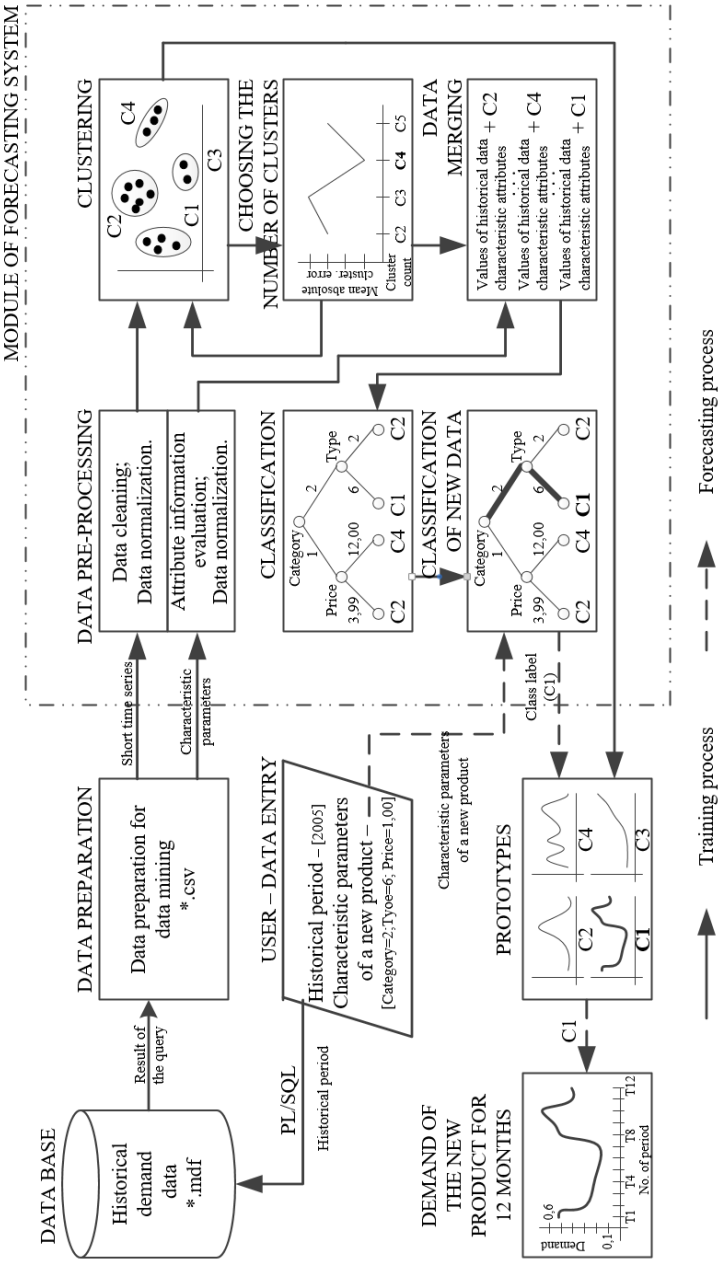

Figure 4. Demand forecasting system

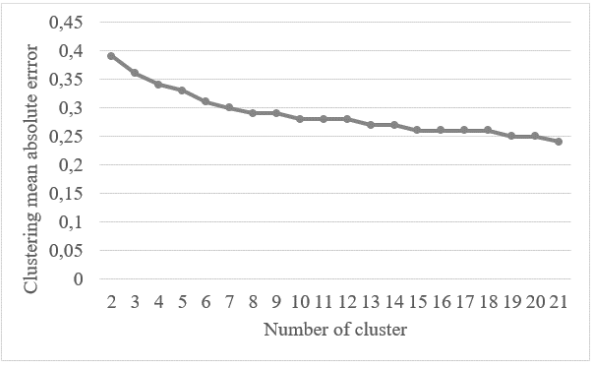

Figure 5. Error evaluation of the *k-means divisive* algorithm at different cluster counts

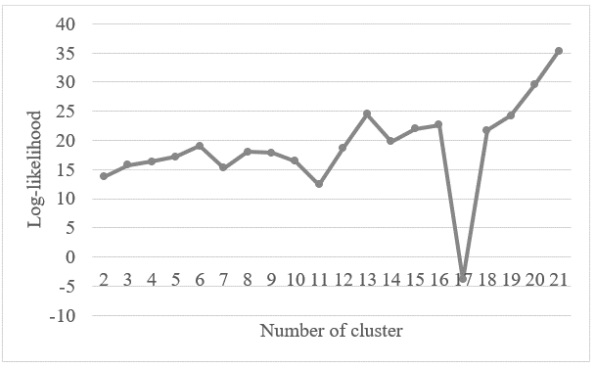

Figure 6. Log-likelihood evaluation of *expectation maximization* algorithm at different cluster counts

Both clustering algorithms are not suitable for cluster analysis of historical demand values – short time series. In the case of *k-means divisive* algorithm, the error decreases as the number of clusters grows but in the case with *expectation maximization* – log-likelihood increases as the number of clusters grows. This points to the low robustness of these algorithms, which means that they are not able to perform accurate data clustering. Therefore the *modified k-means divisive* algorithm was proposed in order to determine the most suitable number of clusters that would be necessary for clustering of the training set.

The most suitable number of clusters is determined using modified k-means algorithm based on the methodology described in subsection '*Clustering error calculation for training data set'* by finding mean absolute error of clustering. Its results are presented in Figure 7. Analysis of clustering mean absolute error results shows that the first significant minimum is reached at 10 clusters and, whereas the following fluctuations of the polygon are miniscule, the most suitable number of clusters for clustering of the training set is set to 10.

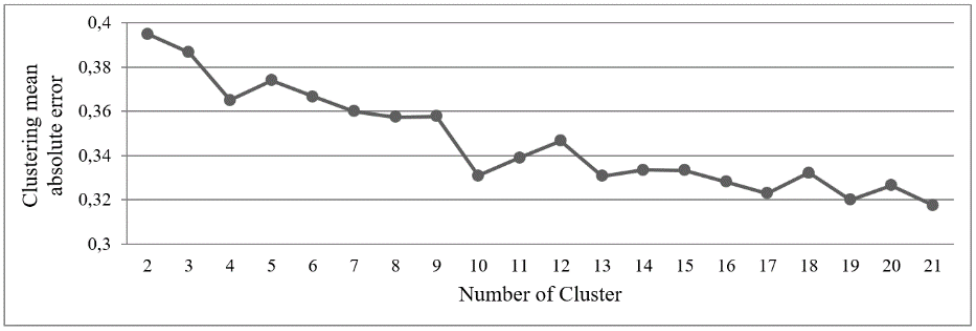

Figure 7. Clustering error calculations at different cluster counts

Evaluation of classification accuracy was based on mean absolute error and root mean squared error because different error calculation techniques provide more accurate interpretation of the obtained results [64]. To evaluate accuracy of a classifier, it is first trained using training data set and then tested using a test data set. *Orange Canvas* software was used to create an experiment set routine, which tested various classification algorithms to find the most suitable one. The obtained results are presented in Table 4. It shows that the smallest error value according to both error estimation methods was acquired using *C4.5* algorithm.

Table 4

Classification error of various classification algorithms in test data

Error estimation approach	Classification algorithm					
	ZeroR	*OneR*	*JRip*	*Naïve Bayes*	*k-nearest neighbour*	*C4.5*
Root Mean Squared Error (RMSE)	0,295	0,375	0,284	0,282	0,298	**0,264**
Mean Absolute Error (MAE)	17,4	14,1	16,1	15,6	16,0	**12,9**

According to classification accuracy results, C4.5 algorithm is selected for DFS.

3.2. Heart Necrosis Risk Forecasting System

Institute of Pharmacological Research that works on medication development for improvement of heart functionality and carries out laboratorial experiments to identify heart necrosis risk nominated a task to develop a heart necrosis risk forecasting system, which would determine the potential risk of heart necrosis for a laboratory animal when only characteristic parameters of this animal are known. Development of such system would help experts of pharmacology to save time spent on gathering research results and to decrease the number of laboratory animals required for experiments.

Heart necrosis risk forecasting system (HNRFS) provides a solution for bioinformatics experts that would help processing heart contraction power data (short time series) and parameters of laboratory animals (descriptive parameters) and determine the heart necrosis risk of a new individual. In the process of system development data of laboratory examinations of pharmacological research (obtained using the 'isolated heart' ischemia-reperfusion model [48, 49]) are analysed. The experiments are carried out using Wistar rat line. The animals are given heart function stimulating drugs with their food for a certain period of time (for eight weeks). Every group of the animals is given one of the drug types for a specific period of time. The goal of such pharmacological research is to determine the efficacy of the analysed drugs in heart cell protection against ischemia-reperfusion damage by estimating the degree of heart necrosis (amount of tissue that dies off). The Mildronāts® medication used in the studies is anti-ischemic substance developed in Latvia that optimizes heart energy metabolism [16]. Heart contraction power and heart rate are registered during occlusion using a tool and software developed by *ADInstruments*, which reads data with an interval of 60 seconds. It accumulates data describing 40 readings during occlusion (heart contraction poser is registered every minute), which include changes in heart contraction power and make up the first data set. The values of heart contraction power in each period are registered in the form of mm on mercury scale (mmHg). The obtained data set values are time series but since the time span of the observations is too short and it does not repeat, they are considered short time series. In the usual analysis of short time series it is almost impossible to find strong functional connections therefore this type of tasks is considered hard to formalize. The goal of the pharmacological experiment was to assess the percent ratio of dead heart tissue (necrosis) that depends on the used medications fed to animals.

The other group of data is characteristics parameters of laboratory animals and the heart necrosis risk assessment obtained in pharmacological experiments. The characteristic parameters describe a laboratory animal, e.g., its weight, studied medicine, blood plasma analysis results and the risk of heart necrosis.

Structure of HNRFS

First step is to carry out pre-processing of short time series where the first and last measurement are removed to avoid possible noise (erroneous readings). Then a time series is constructed with 38 periods for each object in the data set by selecting data about heart contraction power. Data

normalization is carried out using *normalization using life curve* [40, 64] and *z-score normalization using standard deviation* [1] approaches, which were also used in the development of DFS. Working principle of heart necrosis risk forecasting system is presented in Figure 8. Its structure is based on the application of several data mining methods. Methods of mathematical statistics have strict limitations in tasks that require analysis of patterns among short time series. Therefore this type of tasks is being solved using data mining approaches, which are considered more suitable [76].

Short time series are clustered using the modified k-means algorithm determining the most suitable number of clusters that is required to aggregate similar objects into clusters. Characteristic parameters of individuals are used to find correlation between attributes by determining level of their interaction and selecting most informative attributes for further steps of data analysis. The data set of characteristic attributes is divided into five equal data subsets because the laboratory animal food is mixed with five types of studied medicine. The divided data sets are merged with class attribute obtained in clustering process based on object identifier. Every obtained data subset is analyzed using classification to determine relationships between the clustering class and the characteristic parameters of objects by applying inductive decision tree algorithm *C4.5* [58]. During classification of objects a conditional rule is obtained from the classification tree in the form of "IF ... AND ... THEN ...", which holds information about the clustering class, the studied medicine and the heart necrosis risk assessment obtained in pharmacological experiments. The obtained conditional rules are stored in a knowledge base. Forecasting of heart necrosis risk is carried out by entering characteristic parameters of the 'new' individual. Based on the type of the studied medicine, characteristic parameters are projected onto the corresponding classifier to determine class value and creating a condition 'IF ... AND ...' (e.g. IF Group=2 AND Class=C2), which is transmitted to the knowledge base. Then all rules that fit the entered condition (IF Group=2 AND Class=C2) are selected from the knowledge base. Each condition is expanded using the information found in the knowledge base creating a set of temporary rules 'IF ... AND ... THEN ...', for example:

- IF *Group*=2 AND C*lass*=C2 THEN *Risk*=27;
- IF *Group*=2 AND *Class*=C2 THEN *Risk*=32;
- ….
- IF *Group* =2 AND *Class*=C2 THEN *Risk*=40.

The selected condition rule set is used to gather heart necrosis risk apparition frequency statistics (statistical distribution) that is used to create risk distribution function, which is used to obtain the number of apparitions of the corresponding risk value in the temporary rule set. Mathematical expectation calculation and a distance-based approach proposed by the author are used in order to determine heart necrosis risk. Statistical distribution of heart necrosis risk is used to calculate mathematical expectation. The distribution each frequency statistics value is assigned a value based on risk frequency number; the more this value occurs in the temporary rule set, the higher the probability would be. Then all mathematical expectation values of risk are summed together to obtain heart necrosis risk prognosis.

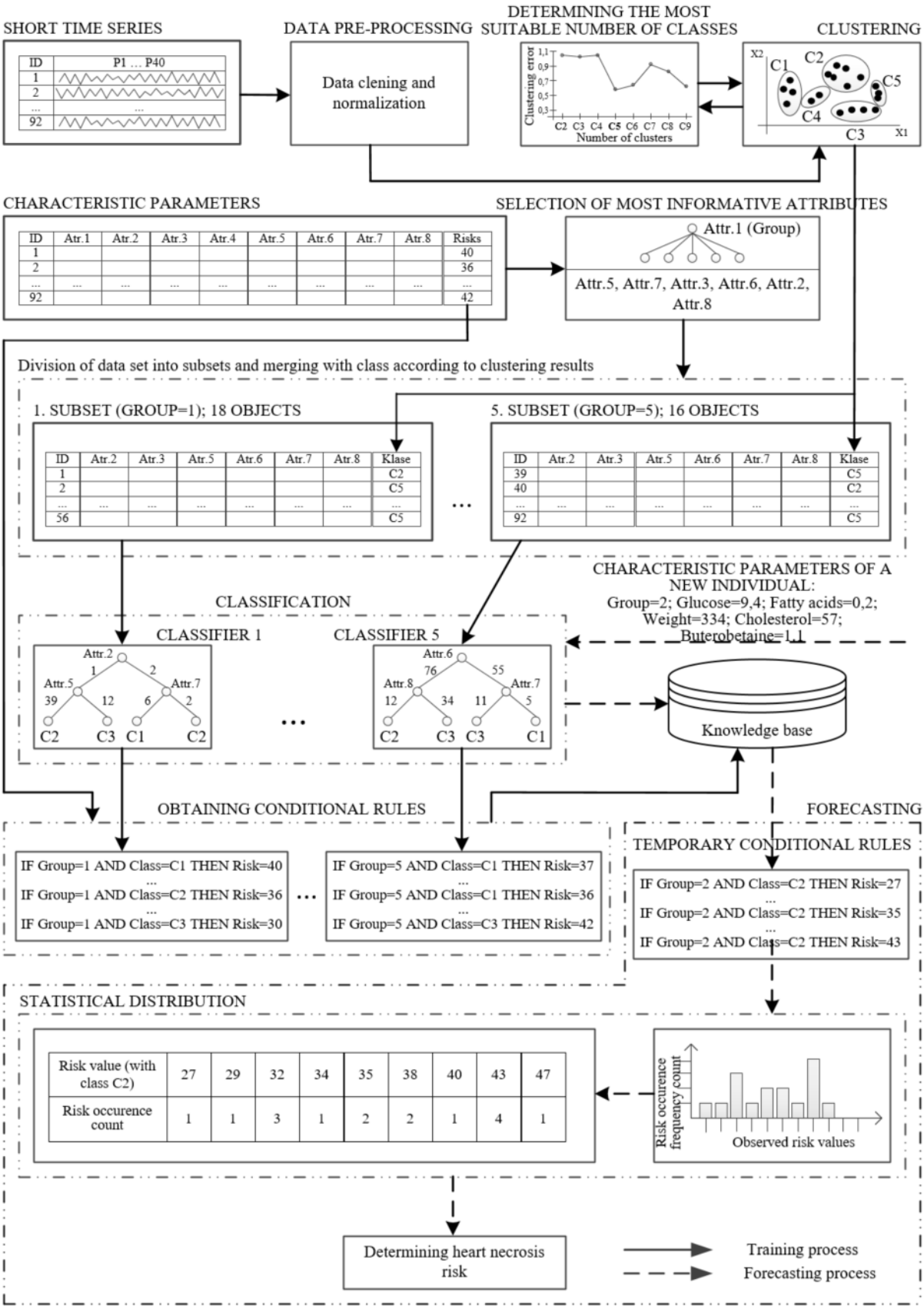

Figure 8. Structure of heart necrosis risk forecasting system

The approach proposed by the author uses distance metric between risk occurrence frequency number and risk values.

HNRFS Experiment Results

Experimental evaluation was carried out on a 92 object data set that describes heart contraction power values in ischemic state (short time series) and characteristic parameters of objects (attributes) like: *Group (Attr.1)* – type of the studied medicine that was given to the individuals; *Weight (Attr. 2);* blood plasma parameters: *Carnitine (Attr. 3), Triglycerides (Attr. 4), Fatty acids (Attr. 5), Glucose (Attr. 6), Cholesterol (Attr. 7), Butyrobetaine (Attr. 8)* and *Risk (Class)* – heart necrosis risk assessment obtained in pharmacological research. Length of short time series after data cleaning was 38 periods. The obtained short time series were normalized using *z-score normalization using standard deviation* and *normalization using life curve* approaches, acquiring two different data sets for each, which were experimentally evaluated and allowed to determine that the most suitable is *normalization using life curve*. The obtained data sets were clustered using *k-means divisive*, *expectation-maximization*, *hierarchical agglomerative* and *modified k-means divisive* algorithms. Results of the *k-means divisive* algorithm are presented in Table 5. The smallest sum of squares error for both normalization approaches is at 9 clusters that shows that this clustering algorithm is not suitable for clustering of short time series because the error decreases as the number of

Table 5

Sum of squares error for *k-means divisive* algorithm

	Number of clusters							
Normalization approach	2	3	4	5	6	7	8	9
Z-score normalization using standard deviation	141.67	138.42	128.36	125.89	122.64	120.56	116.34	**113.92**
Normalization using life curve	115.73	112.99	101.81	96.36	92.54	89.44	87.39	**85.79**

The *expectation-maximization* algorithm, whose results are provided in Table 6, using log-likelihood error estimation, shows the best result at 9 clusters using *z-score normalization using standard deviation* and at 8 clusters for *normalization using life curve*.

Table 6

Log-likelihood for *expectation-maximization* algorithm

	Number of clusters							
Normalization approach	2	3	4	5	6	7	8	9
Z-score normalization using standard deviation	-49.64	-48.64	-47.93	-47.39	-46.36	-45.47	-45.25	**-44.13**
Normalization using life curve	144.49	145.40	148.42	149.06	150.06	150.32	**151.13**	150.87

Results show that this algorithm is also not suitable to determine the best number of clusters because increase of cluster count correlates with the increased log-likelihood, which shows that the algorithm is not robust.

The results of the *modified k-means divisive* algorithm are presented in Table 7. The mean absolute error for *normalization using life curve* is the smallest for 5 clusters.

Table 7

Mean absolute error for the *modified k-means divisive* algorithm

	Number of clusters							
Normalization approach	2	3	4	5	6	7	8	9
Z-score normalization using standard deviation	1.233	1.192	1.158	1.150	1.128	1.095	1.093	0.946
Normalization using life curve	1.026	1.017	1.026	**0.59**	0.644	0.91	0.824	0.615

The characteristic parameters were first analyzed to find the most informative attribute subset, which could be used in classifier construction. Data classification depends on informativeness of the attributes. If there is correlation between attributes then they are considered to be connected [70]. While if there is no correlation it means that the connection between them is weak and they should not be used in classification. The experimental evaluation used *CfsSubsetEval* attribute evaluation and *BestFirst* search methods that are commonly used in data mining to assess informational content of attributes [70]. The following results were obtained: *Fatty acids*, *Cholesterol* and *Carnitine* showed 100%, *Glucose* 90%, *Weight* 80% and *Butyrobetaine* shows 20% correlation, while *Triglycerides* and *Group* showed 0%. The last two parameters should have been removed but *Group* attribute holds very important information about the type of the studied medicine, which is the basis for this research; therefore this parameter was not excluded. Then the whole initial data set was divided into subsets according to the values of *Group* parameter, which resulted in five data subsets. This division of the data set allows removing non-informative attribute from classifier training, while preserving the information about the type of the medicine that it held for conditional rule construction.

The most suitable classifier was selected based on classification accuracy results, which are presented in Table 8 for the case with the number of classes determined by the *modified k-means divisive* algorithm and Table 9 for the case with the number of classes determined by *k-means divisive* algorithm.

Experimental evaluation of *Naive Bayes*, *k-nearest neighbours*, *C4.5* and *CN2* algorithms shows that the most suitable algorithm is *C4.5*. The obtained classification results were used to create conditional rules that are stored to knowledge base. Forecast of heart necrosis risk is carried out based on the decision tree obtained in the classification process and the characteristic parameters of a ‘new’ individual. These characteristic parameters are classified using the

classifier with the corresponding *Group* value to determine the class. The obtained class value and value of *Group* attribute are used to construct a temporary conditional rule, which is used to select risk values from the knowledge base to generate conditional rules.

Table 8

Classifier accuracy of classification algorithms using the class count determined by the *modified k-means divisive* algorithm

	Classifier			
	Naive Bayes	*C4.5*	*k-nearest neighbours*	*CN2*
1. subset	0.45	0.7	0.65	0.6
2. subset	0.25	0.3	0.2	0.15
3. subset	0.25	0.6	0.4	0.5
4. subset	0.45	0.55	0.3	0.75
5. subset	0.2	0.25	0.5	0.3
Mean value	0.32	**0.48**	0.41	0.46
Whole data set		0.38		

Table 9

Classifier accuracy of classification algorithms using *k-means divisive* algorithm and the most suitable class count determined previously

	Classifier			
	Naive Bayes	*C4.5*	*k-nearest neighbours*	*CN2*
1. subset	0.35	0.45	0.4	0.1
2. subset	0	0.1	0.05	0.05
3. subset	0.45	0.35	0.4	0.2
4. subset	0.05	0.05	0.2	0.05
5. subset	0.27	0.4	0.25	0.25
Mean value	0.224	0.27	0.26	0.13
Whole data set		0.3		

The obtained conditional rules are used to create statistical distribution, which is then used to calculate heart necrosis risk using distance metric shown in Table 10.

Table 10

The calculation of the possible heart necrosis risk according to risk occurrence frequency distance assessment

	Calculations								
Risk occurrence frequency counts (ROFC)	1	1	3	1	2	2	1	4	1
Risk values	27	29	32	34	35	38	40	43	47
Distance between risk values	-	2	3	2	1	3	2	3	4
Difference between ROFC min and ROFC values	0	0	-2	0	-1	-1	0	-3	0
Sum: **Distance** + **difference**		2	1	2	**0**	2	2	**0**	4
Assessment					35			43	
Mean assessed value	**39**								

3.3. Bacteria Proliferation Syndrome Detection System

Bacteria proliferation in small intestine is colonization of the intestines with microorganisms of the colon that can lead to a wide spectrum of clinical reactions, starting with light and vague symptoms and ending with severe indigestion. Bacteria proliferation in the small intestine in its initial form, when bacteria move from the colon to the distal small intestine, provokes complaints only in patients with chronic digestive system diseases. Symptoms that are caused by bacteria proliferation in the small intestine are very atypical. Patients suffering from it complain about discomfort in the intestines, increased gas formation (meteorism), changes in stool [50, 56].

Bacteria proliferation syndrome (BPS) detection involves diagnostic tests that can be divided into invasive and non-invasive test. Detection of *BPS* has not yet been standardized [20], therefore nowadays doctors in clinical praxis use glucose (*GET*) and lactose (*LET*) breath tests. Clinical algorithm of BPS detection states that an individual is initially assigned *GET* test. If it is positive, the individual is assigned *LET* test. Only if both tests show positive results an individual is diagnosed with *BPS*. The breath tests use specific hydrogen concentration in the exhaled air. Individuals suffering from *BPS* usually show one 'early' maximum of exhaled hydrogen in the *GET* test. For the test procedure to be correct, the individual has to abstain from eating at least 6 hours before undergoing the procedure. At least 30 minutes before taking the substrate the individual has to refrain from smoking and physical activities. Alveolar air, i.e. the last part of one exhale (around 150 ml), is used for measurements. *GET* procedure: patient takes 75 g of glucose that is dissolved in 400 ml of water [46]. If the concentration of H_2 in the exhaled air rises more than $\Delta = 20$ Pm, it is considered to be a positive test result and *BPS* in the small intestine is confirmed.

Since *GET* and *LET* tests are time-consuming and require a long period of preparation, they usually cause negative attitude in individuals towards this procedure. Therefore the aim is to offer alternative solution for *BPS* detection in the small intestine using data mining methods and algorithms. A *BPS* detecting system is created that is based on examination results from *GET* testing and filling out a questionnaire about symptom self-assessment of an individual. *GET* test measurements are considered short time series [21], because number of measurements in one time series is 10 (10 time periods), which are obtained from exhaled breath samples (taken directly before the procedure, as well as every 20 minutes after taking the substrate during the following 3 hours). The health self-assessment of an individual is described by several questions that are answered by the individual during *GET* test. The questions are asked by a qualified medical expert, who also registers the answers in a specific protocol.

Working Principle of the BPS Detection System

The BPS detection system (presented in Figure 9) has to process two types of data flows. One is the historical results of GET tests that are represented by short time series with 10 intervals of time, the other – health self-assessment of an individual, which is described by following attributes: *gender*, *feeling sick*, *eructation*, *flatulence*, *fullness* and the results of clinical

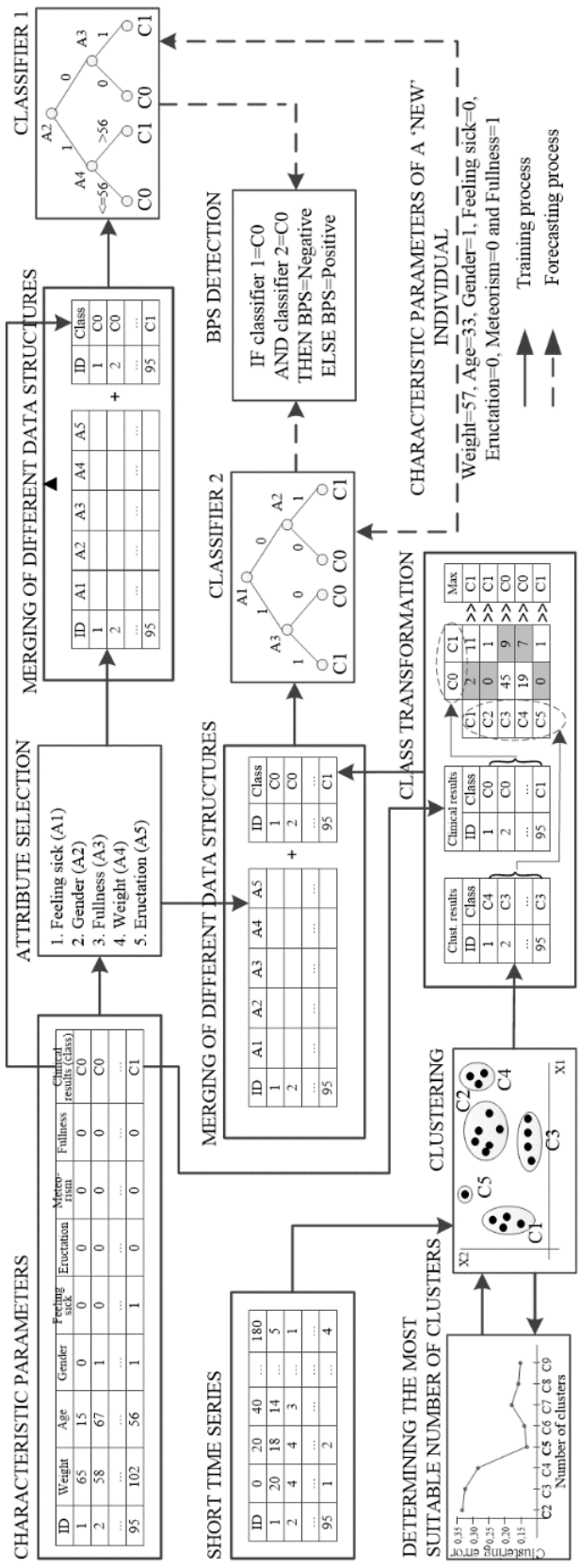

Figure 9. Bacteria proliferation syndrome detection system

parameters of individuals and clustering results.

examinations that are represented by *class* attribute. The system has to perform: selection of the most informative attributes, data clustering, determination of the most suitable number of clusters, class transformation, creation of a new data set by merging different data structures, classification of two types of data flows and *BPS* detection.

Evaluation of informativeness of attributes is carried out using *CfsSubsetEval* attribute evaluation and *BestFirst* searching methods that are often used in data mining [70]. After selection of most informative attributes among health self-assessment of an individual, the data set of characteristic parameters, which will be called *DS1* in this study, is extended with results of clinical examinations (attribute *class* – clinically approved bacteria proliferation syndrome). The created data set DS1 is processed by the 1st classifier using *k-nearest neighbours*, *C4.5* and *CN2* classification algorithms determining connections between characteristic parameters of an individual and classes. Short time series are processed in the clustering model using *modified k-means divisive* algorithm to determine the most suitable number of clusters for clustering of the data set. Clustering results are classes that are transformed to class structure of clinical examinations in the transformation model. The obtained class structure is added to the selected most informative attribute set based on the identification of an individual. This data set will be called *DS2* in this study and is then processed in the 2nd classifier using *k-nearest neighbours*, *C4.5* and *CN2* classification algorithms determining connections between characteristic

Results of the 1st classifier that processes data set *DS1* describe the connection between characteristic parameters of health self-assessment of an individual and the clinical results. Results of the 2nd classifier that processes data set *DS2* describe the connection between breath test measurements and the class obtained in clustering, which is reduced to the class structure of clinical results. The model of processing of two different data flows provides analysis of the obtained results from different sides. There are two independent assessments, which are compared in the model of *BPS* detection, which produces a notice stating whether the individual needs further examination. This proposed model for processing of two data flows guarantees that, if an individual has not answered health self-assessment questions honestly, there is a chance that the second classifier will detect the false information.

BPS Detection System Experiment Results

Experimental validation was carried out using data from a retrospective study that included individuals of both genders with no age limitation. The research included 95 individuals who were tested using glucose breath test. A protocol was developed including readings of glucose breath test hydrogen levels without CO_2 correction in a time period from 0 till 180 minutes with 20 minute intervals, as well as *Weight* and *Age* of an individual at the moment of uptake into research, *Gender*, conclusion of the glucose test or result of the clinical study – *Class* (negative – 0, positive – 1), parameters of health self-assessment of and individual: *Feeling sick*, *Eructation*, *Flatulence* and *Fullness*. A negative glucose test shows that there is no need for further examinations, whereas a positive test result shows that further examination is needed.

The most informative attributes were determined using *CfsSubsetEval* attribute evaluation and *BestFirst* search methods. From the seven attributes of the initial data set (*Gender, Weight, Feeling sick, Eructation, Flatulence* and *Fullness*) five attributes were selected: *Feeling sick* (100 %), *Gender* (70 %), *Fullness* (50 %), *Weight* (40 %) and *Eructation* (40 %). The obtained results in the brackets show percentage of times when the attribute was selected into the combination with the best evaluation.

The experimental evaluation determined the most suitable number of clusters that should be used for clustering of the data set, based on the mean clustering error for different cluster counts. As it can be seen in Table 11, the most suitable number of clusters is five because this number of clusters shows the smallest mean absolute error (*MAE*).

Table 11

Clustering results using the *modified k-means divisive* algorithm

	Number of clusters							
	2	3	4	5	6	7	8	9
Mean absolute error	0.335	0.329	0.283	**0.125**	0.146	0.183	0.167	0.154

The results also show that this algorithm is robust in short time series clustering unlike the classical *k-means divisive* algorithm, whose results are presented in Table 12.

Table 12

Clustering results using *k-means divisive* algorithm

	Number of clusters							
	2	3	4	5	6	7	8	9
Sum of squares error	20.56	19.15	17.42	15.93	15.23	14.74	13.76	**12.59**

To determine the most suitable classification algorithm, a set of experiments was performed using *C4.5*, *CN2* and *k-nearest neighbours* classification algorithms. The obtained results of classification experiments for the 1st classifier and data set *DS1* are presented in Table 13. The accuracy was evaluated using different approaches.

Table 13

Classification results using data set *DS1*

Classification algorithms	10-fold cross-validation			Leave one out		
	Classification accuracy	Sensitivity	Specificity	Classification accuracy	Sensitivity	Specificity
C4.5	0.65	0.87	0.08	0.64	0.87	0.04
kNN	0.63	0.86	0.04	0.68	0.87	0.12
CN2	0.67	0.91	0.04	0.63	0.86	0.04

It can be seen that the used classifiers have determined the negative class better ("C0"), i.e. the individuals who do not need any further examinations. If one evaluates it from the perspective of the defined task, it is a positive result, but viewed from the perspective of positive class accuracy ("C1") the result is not satisfactory because specificity, e.g. of *C4.5* algorithm with 10-fold cross-validation, is only 0.08.

The next set of experiments was carried out for the 2nd classifier using data set *DS2* and the obtained results are presented in Table Table 14. The results show that *kNN* algorithm recognized both, the positive and the negative, classes equally well. *CN2* shows almost 100% recognition of the negative class but the accuracy for the positive class is not satisfactory. *C4.5* algorithm showed similar results with both accuracy evaluation approaches, recognizing the negative class with sensitivity 0.65 and 0.69 accordingly and the positive class with specificity 0.39 in both cases.

Table 14

Classification results using data set *DS2*

Classification algorithms	10-fold cross-validation			Leave one out		
	Classification accuracy	Sensitivity	Specificity	Classification accuracy	Sensitivity	Specificity
C4.5	0.53	0.65	0.39	0.55	0.69	0.39
kNN	0.46	0.51	0.39	0.42	0.43	0.41
CN2	0.6	0.98	0.16	0.63	0.98	0.23

Since *C4.5* algorithm showed similar results with both accuracy evaluation approaches, classification of *DS1* and *DS2* data sets using 1st classifier and 2nd classifier correspondingly was carried out using the system with *C4.5* classifier performing BPR detection.

4. GUIDELINES FOR FORECASTING SYSTEM DEVELOPMENT

Guidelines for forecasting system development (GFSD) provide instructions for realization of data processing systems where the data source is in the form of short time series and their characteristic properties. Knowledge that GFSD are based on is drawn from the experience in development of similar systems – product demand, heart necrosis risk and bacteria proliferation syndrome forecasting systems. The guidelines advise a developer about the recommended structure of the model based on the data description provided by a customer. A customer is a person or organization that defines requirements and provides description of the analyzed data. A developer is a person or an organization that evaluates the submitted requirements, analyzes them and prepares an answer to the customer. If both sides agree on the next step, the developer develops the most suitable system implementation based on the submitted requirements and the guidelines. After choosing the forecasting system implementation, the developer implements the system and validates its work based on the classification accuracy evaluation. If the evaluation is sufficient, the developer executes system introduction and integration with the information systems of the customer.

Structure of a Forecasting System

The structure of a forecasting system is presented in Figure 10. It consists of five steps. The first step involves defining the requirements that are based on the dialog of a customer and the developer, analyzing the conformity of the customer data and the system to be developed and the defined task, the aim of the forecast. The second step involves choosing the most suitable scenario for system implementation. The third step involves construction of the system concept. The fourth step involves testing of the system concept; if the accuracy of this concept is sufficient, the concept is turned into a system. The fifth step involves integration of the developed system into the information system of the customer.

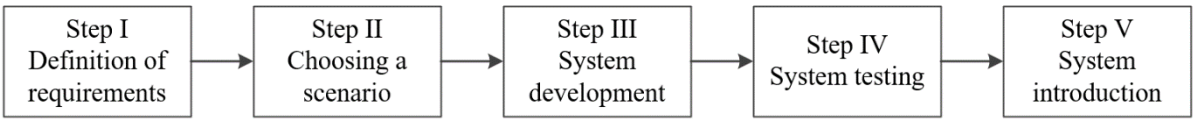

Figure 10. Structure and steps of a forecasting system

The step of requirements definition is showed in detail in Figure 11. The customer submits data set to be analyzed to the developer and defines the aim of the forecast. In order to create a forecasting system, the data to be analyzed have to represent values of historical events and their descriptive parameters. The aim of the forecast has to be clearly defined. It has to be reached by only using characteristic parameters of a ‘new’ analyzed object. In the case when system has to be trained using a specific period of historical periods (short time series with a specified period of time, e.g., product sales volume data in year 2012), then the customer has to ask for this option to be integrated into the system. The definition of requirements submitted by the customer is used to evaluate compliance of data structures with the defined aim. If there is a concordance then the developer can move on to the second step. If there is none, the

developer writes a report about the reason of refusal and performs inadequacy aversion in collaboration with the customer.

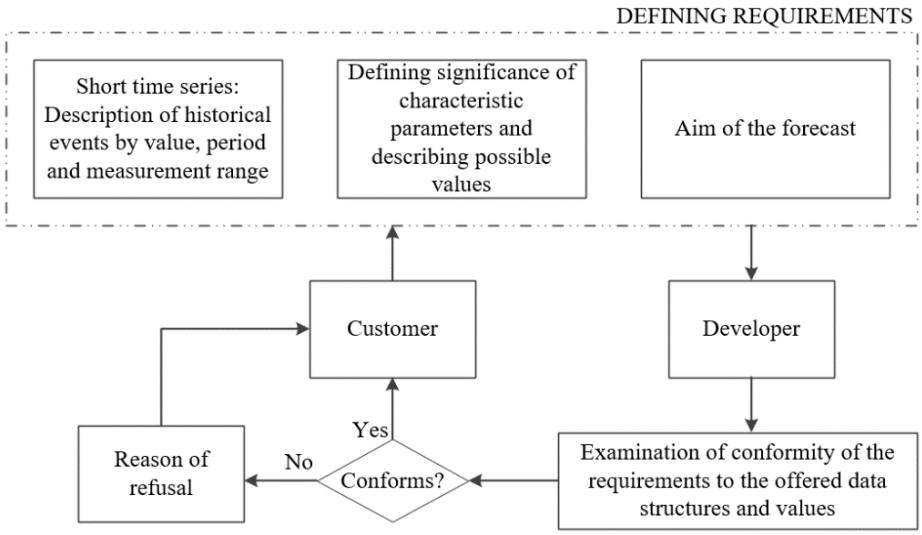

Figure 11. Definition of requirements for a forecasting system (step I)

Choosing the System Scenario

Choosing the forecasting system scenario starts from data base of a user (see Figure 12), which is sent a query to select information that is necessary for the analysis (historical short time series and their characteristic parameters), which is submitted to data splitting block. Data splitting block splits data into two flows: historical time series and their characteristic parameters. The historical time series can be processed performing data preparation and/or data pre-processing. Data preparation process involves period equalization, processing of missing values. Data normalization involves *normalization using life curve*, processing dominant values. Detection of groups of similar objects is carried out using cluster analysis and *modified k-means divisive* algorithm, determining the most suitable number of clusters for clustering of a data set. Based on the task specifics, it is possible to transform the number of classes obtained in the clustering process into the number of classes used in the characteristic parameter data. It is also possible to transform clustering results into prototypes that represent mean values of a cluster in each period of time. In the process of characteristic parameter data preparation it is advised to perform processing of missing values, data normalization using *z-score normalization using standard deviation* and data discretization. The classification process of the characteristic parameters can be carried out in several ways:

a) For classification use the data set of characteristic parameters and a class describing these parameters (specified by clinical, pharmacological or other research and is considered the ‘golden standard’).
b) If necessary, split data set of characteristic parameters into subsets based on the distribution of attribute significance that is determined experimentally or defined by the customer.
c) For classification use the data set of characteristic parameters and the class obtained in clustering.

Merging of different data types is implemented for options ‘b’ and ‘c’. To determine connections between data of different types or classification, it is advised to use *C4.5* algorithm. The specifics of a task may ask for more than one classifier.

Construction of conditional rules is based on assembling of results of several processes in a form of IF … AND … THEN rules where:

a) The first condition is the splitting parameter value of the created subset (e.g., the type of an analysed medication that was added to food – attribute Group; value range [1…5]).
b) The second condition is the class of the analysed object (determined in the process of clustering).
c) The consequent – attribute *Risk* is the value of “golden standard”.

Classification of each object results into one conditional rule that is stored to rule data base, e.g., IF Group=1 AND Class=C2 THEN Risk=35.

System scenarios developed for different scopes in the Thesis are displayed in Figure 12: ‘○ ▬▬’ shows the scenario for demand forecasting system (DFS), ‘× ▬▬’ shows heart necrosis risk forecasting system (HNRFS) and ‘Δ ▬▬’ shows bacteria proliferation syndrome detection system (BPSDS).

Forecasting for a new object is carried out by entering the characteristic parameters into the system where a classifier is used to determine association of an object with one of the classes and then the obtained classification results are interpreted according to the task specifics:

a) If forecasting is carried out using prototypes, the class obtained from the classifier points to the number of the corresponding prototype, which describes demand of the product in the future at a certain period of time. The user can choose the specific historical demand period, which will be used to train the system before forecasting.
b) If forecasting is carried out using conditional rules, the class obtained from the classifier and the value of the splitting parameter define specific conditional rules that have to be selected from the data base of conditional rules. The selected rules that describe frequency of conditional rule appearance serve as a basis to calculate the forecast value.

If forecasting is carried out data sets with and without the use of clustering results, the forecast is determined based on the comparison of results acquired from different classifiers. If both classifiers point to class ‘C0’, the prognosis is that there is no need for further action. If the result is class ‘C1’ or a result could not be obtained, the forecast means that there is a need for further action. Correctness of the choice of scenario can only be determined experimentally by creating a system concept and carrying out testing with the real data set available to the customer.

Development and Testing of Forecasting System

After a scenario is chosen, a concept of the forecasting system is developed. It is used for experimentation using the real data set available to the customer. Results of experiments show the accuracy, sensitivity and specificity of the classifier. In some cases, especially in healthcare, some additional accuracy evaluation methods might be necessary that can be calculated from the extended confusion matrix. If the analysed data set represents 200 or more objects, the classifier of the system is trained using a training data set and then tested using a test data set. Whereas, if the analysed data set represents less than 200 objects, a 10-fold cross-validation is used, which guarantees credibility of results when the number of objects is small.

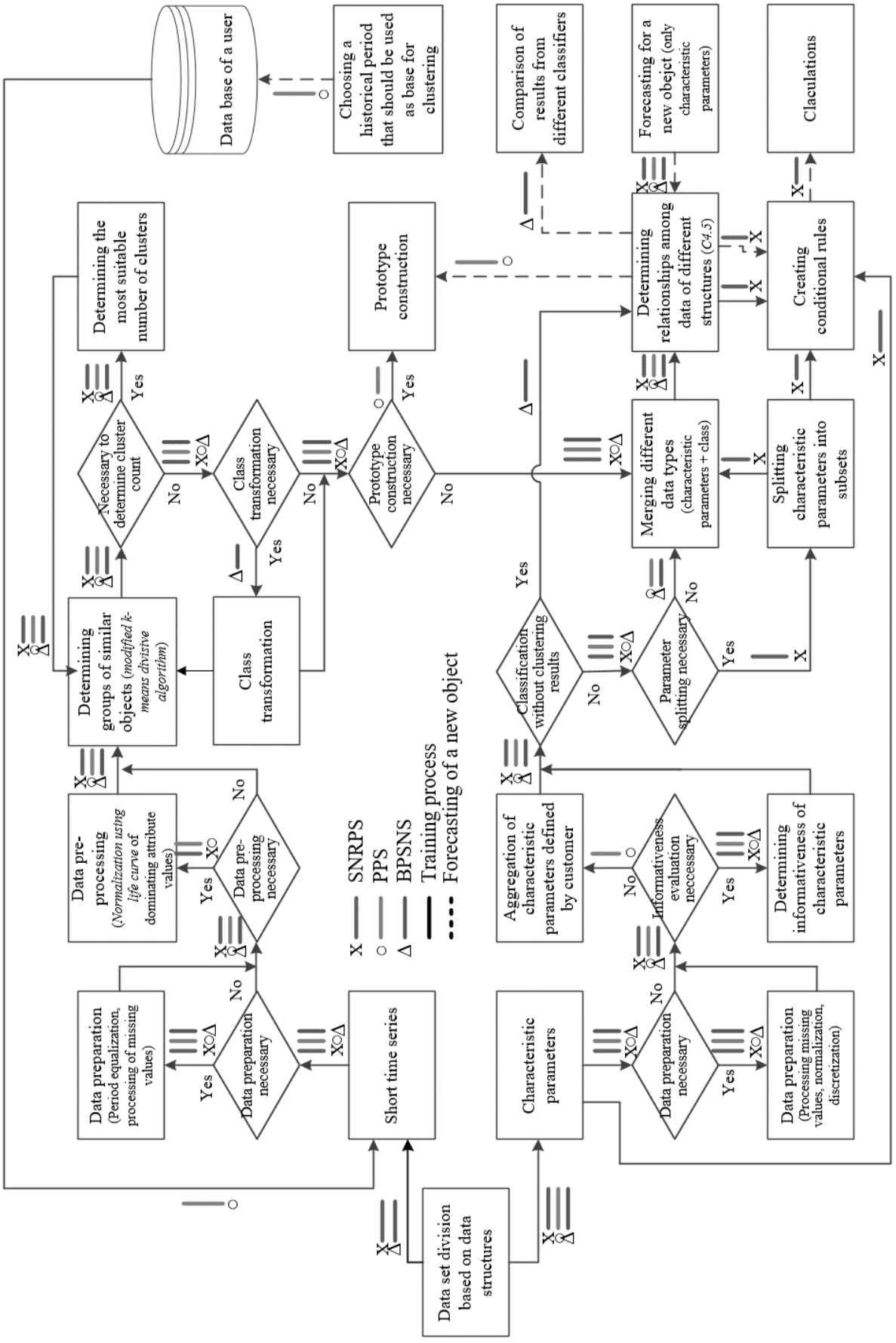

Figure 12. Selecting a scenario for development of a forecasting system (step II)

Number 200 is an estimated limit. If the obtained evaluation of classification results acquired by the developed system is acceptable for the expert – system developer, then this concept can be implemented into an application which can be then integrated into information system of the customer.

In the case if the evaluation of classification results obtained from the system is insufficient (as concluded by the expert – the system developer), other solutions can be chosen based on the guidelines and carry out repeated accuracy evaluation of the developed system. It is important to correctly select the most suitable data preparation and data pre-processing approaches because the most significant inaccuracies in the process of data mining stem from an incorrect choice of data pre-processing techniques and poor pre-processing of the data set. The proposed guidelines for processing of short time series and characteristic parameters could be imperfect for development of new systems in different fields because they are created based on the experience obtained in demand forecasting for products in sales, heart necrosis risk forecasting in pharmacology and bacteria proliferation syndrome detection in medicine. Therefore, if a system is being developed for another field, the guidelines could definitely be complemented with new blocks, possible events and other data flows.

RESULTS AND CONCLUSIONS

The Thesis proposes a system for forecasting tasks in different fields and processing short time series and their descriptive parameters based on data mining. During this study the following tasks were completed:

1. Processing principles of short time series and their characteristic parameters were analysed, determining methods that can be used in system implementation.
2. As a result of analysis the most suitable data pre-processing methods for processing of short time series and characteristic parameters were selected.
3. A *modified k-means divisive* algorithm was developed according to the field specifics; it is used to cluster short time series and determine the most suitable number of clusters. The developed algorithm modification was compared to other clustering algorithms.
4. A data merging technique was developed to merge clustering results and characteristic parameters of short time series, which provides class transformation approaches to equalize numbers of classes in the data structures to be analysed.
5. Forecasting systems for sales, pharmacology and medicine fields were developed. They carry out forecasting for a new object based on the characteristic parameters of an object that are entered into the system.
6. An accuracy evaluation of clustering and classification models that are used in the system was carried out.
7. Conditional rule construction and application approaches were developed for different fields.
8. Guidelines to guide developers of new forecasting systems, which use short time series and their characteristic parameters as data source, were developed.

The product demand forecasting system implements visualization of clustering results using prototypes, which are later used in forecasting. The heart necrosis risk forecasting system and the bacteria proliferation syndrome detection system implement the use of several classifiers. A result merging approach that merges the results obtained in classification training process and results of pharmacological research and that was implemented using conditional rules.

The developed modification of a clustering algorithm and all three of the developed forecasting systems were tested experimentally to test the previously defined hypotheses:

1. The developed system for various fields points towards accepting the stated hypothesis that development of a processing system for short time series and their characteristic parameters provides a solution of a task, which is difficult to formalize, using data mining methods and algorithms.
2. The second hypothesis is accepted due to the developed modified k-means divisive algorithm, which improves the determination of the most suitable number of clusters in the clustering process by analysing short time series. The developed modification of the algorithm was experimentally tested in different scopes and compared to other clustering algorithms.
3. The third hypothesis was accepted due to the fact that there had been three systems developed (product demand forecasting, heart necrosis risk forecasting and bacteria proliferation syndrome detection systems) that are used in different fields: sales, pharmacology and medicine to process data sets that hold short time series and their characteristic parameters.

The experimental analysis carried out for this study allows concluding the following:

a) The most suitable normalization method is *normalization using life curve*, which showed the best results in all of the developed systems.
b) The most suitable algorithm for short time series clustering is the *modified k-means divisive* algorithm; it also maintained its robustness in the range of all analyzed clusters and showed good results in all of the developed systems.
c) The most suitable algorithm for classification is *C4.5* because it showed the best results in classification accuracy evaluations.
d) For data sets with less than 200 records the recommended approach for classification accuracy evaluation is 10-fold cross-validation; for data sets with 200 or more records the recommended approach is to divide the data set into training and test sets with 70:30 ratio.
e) When calculating the possible heart necrosis risk from a statistical distribution, it is recommended to use the approach proposed by the author of this Thesis that is based on distance metrics.

Scientific Results Achieved in the Study for the Thesis:

1. A modification of clustering algorithm for analysis of short time series was developed.
2. A transformation approach to transform class structures with different numbers of classes into a unified structure of classes.
3. Guidelines were developed to guide a developer through development of similar forecasting systems.

Practical Results Achieved in the Study for the Thesis:

1. A product demand forecasting system was developed, which determines the possible demand for an object; the system was tested using real product demand data.
2. A heart necrosis risk forecasting system was developed, which determines the possible value of heart necrosis risk of an object; the system was tested using real data from pharmacological research.
3. A bacteria proliferation syndrome detection system was developed, which determines if the analysed individual needs to undergo lactose test; the system was tested with real medical data.

During the study for this Thesis the following conclusions about the developed system for processing short time series and their characteristic parameters were made:

1. During development of DFS and HNRFS it was experimentally proven that data normalization gave best results when *normalization using life curve* was used.
2. Determining informativeness of characteristic parameters and selection of parameters improves classification results, decreases the size of an analyzed data set and increases the speed of algorithm execution.
3. The developed modification of the clustering algorithm can be used also in solution of other cluster analysis tasks where the data source holds short time series.
4. The developed modification of the clustering algorithm allows using different evaluation approaches: 10-fold cross-validation or splitting of a data set into training and test sets.
5. The developed approach to class transformations allows comparing data structures with different numbers of classes.
6. When developing similar forecasting systems in medicine, classification accuracy evaluation has to be carried out by using also sensitivity and specificity in parallel because these parameters have a significant influence on selection of a classifier.
7. In all three of the developed systems the best classification algorithm (determined experimentally) was *C4.5*.
8. Using several classifiers in system development increases the overall classifier accuracy; this was experimentally proven using divided and whole data sets.
9. The developed forecasting systems for short time series and their characteristic parameters motivate and ensure solution of a complex formalized task using a combination of clustering algorithms, their modifications and classification algorithms.
10. The developed forecasting systems for short time series and their characteristic parameters realize forecasting for a new object based only on the characteristic parameters of this object.
11. The developed forecasting system guidelines allow guiding development of new forecasting systems, as well as provide a chance to expand the existing guidelines.

The following research would be associated to medicine – developing a screening system for decreasing gastric cancer risk [41]. This would allow using the forecasting system guidelines developed in this study to develop a gastric cancer screening system by connecting results of non-invasive examinations (short time series) with characteristic parameters of a respondent. This type of screening systems can be integrated into healthcare centres, which would be of help to field experts when determining diagnoses or appointing patients to examination with a specialist.

REFERENCES

1. Datu ieguve: Pamati/ A. Sukovs, L. Aleksejeva, K. Makejeva u.c. – Rīga: Rīgas Tehniskā universitāte, SIA „Drukātava", 2006. – 130 lpp.
2. Dravnieks J. Matemātiskās statistikas metodes sporta zinātnē. – Rīga, 2004. – 76 lpp.
3. Klasifikācija un klasterizācija izplūdušajā vidē/ L. Aleksejeva, O. Užga-Rebrovs, A. Borisovs – Rīga: Rīgas Tehniskā universitāte, 2012. – 248 lpp.
4. Latvijas Zinātņu Akadēmijas TK ITTEA terminu datubāze/ Internets. – http://termini.lza.lv - Resurss apskatīts 2015. gada 22. janvārī.
5. Smotrovs J. Varbūtības teorija un matemātiskā statistika. – Rīga: Apgāds Zvaigzne ABC, 2004. – 264 lpp.
6. Varbūtību teorijas un matemātiskās statistikas elementi medicīnas studentiem/ U. Teibe, U. Berķis – Rīga: AML/RSU, 2001. – 88 lpp.
7. Alba E., Mendoza M. Bayesian forecasting methods for short time series// Foresight: The International Journal of Applied Forecasting, International Institute of Forecasters. – 2007. – Issue 8. – pp. 41–44.
8. Armstrong J.S., Collopy F. Error measures for generalizing about forecasting methods: Empirical comparisons// International Journal of Forecasting 8. – 1992. – pp. 69–80.
9. Armstrong J. S., Collopy F., Yokum J. T. Decomposition by causal forces: A procedure for forecasting complex time series// International Journal of Forecasting 21. – 2005. – pp. 25–36.
10. Armstrong J.S., Fildes R. Correspondence on the selection of error measures for comparisons among forecasting methods// International Journal of Forecasting 14. – 1995. - pp. 67–71.
11. Berndt D. J., Clifford J. Using dynamic time warping to find patterns in time series// Association for the Advancement of Artificial Intelligence, Workshop on Knowledge Discovery in Databases (AAAI), – 1994. – pp. 229–248.
12. Berry M.W., Browne M. Lecture notes in data mining. – World Scientific Publishing Co. Pte. Ltd., 2006. – 222 p.
13. Boyer K. K., Verma R. Operations and Supply Chain Management for the 21st Century. – USA: South-Western Cengage Learning, 2010. – 560 p
14. Clark P., Niblett T. The CN2 induction algorithm. Machine Learning, 3(4), – 1989. – pp. 261–283.
15. Cost S., Salzberg S. A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features // Machine Learning. – 1993. – Vol.10. – pp. 57-78.
16. Dambrova M., Liepinsh E., Kalvinsh I., Mildronate. Cardioprotective Action through Carnitine-Lowering Effect// Trends Cardiovasc. Med. – 2002. – Vol.12. – pp. 275-279.
17. Dellaert F. The Expectation Maximization Algorithm. College of Computing, Georgia Institute of Technology, Technical Report number GIT-GVU-02-20, – 2002.
18. Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm// Journal of the Royal Statistical Society. Series B (Methodological). – 1977. – Vol.39 (1). – pp. 1–38.
19. Devisscher M., De Baets B., Nopens I. Pattern discovery in intensive care data through sequence alignment of qualitative trends: proof of concept on a diuresis dataset// Appearing in the Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications, Helsinki, Finland. – 2008.
20. Donald I.P., Kitchingmam G., Donald F., Kupfer, R. M. The diagnosis of small bowel bacterial overgrowth in elderly patients// J. Am. Geriatr. Soc. – 1992. – Vol.40(7). – pp. 692–696.
21. Ernst J., Nau G. J., Bar-Joseph Z. Clustering short time series gene expression data// Bioinformatics. – 2005. – Vol.21. – pp. 159–168.

22. Flores J. J., Loaeza R. Financial time series forecasting using a hybrid neural-evolutive approach// Proceedings of the XV SIGEF International Conference, Lugo, Spain. – 2009. – pp. 547–555.
23. Gardner E. S. Jr., Exponential Smoothing: The State of Art// Journal of Forecasting. – 1985. – Vol.4. – pp. 1–28.
24. Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., Lander E. S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring// Science. – 1999. – Vol. 286 (5439). – pp. 531–537.
25. Grabusts P. The choice of metrics for clustering algorithms// Environment. Technology. Resources: Proceedings of the 8th International Scientific and Practical Conference. – 2011. – Vol.2. – pp. 70–76.
26. Graves S.C., Kletter D.B., Hetzel W.B. A dynamic model for requirements planning with application to supply chain optimization// Operation Research. – 1998. – Vol.46(3). – pp. 35–49.
27. Hall M. A. Correlation-based feature selection for machine learning/ Doctoral Thesis, Hamilton: University of Waikato. – 1999. – 178 p.
28. Han J., Kamber M. Data Mining: Concepts and Techniques. Second Edition. – Morgan Kaufmann, Elsevier Inc., 2006. – 800 p.
29. Hsieh C. H., Anderson C., Sugihara G. Extending nonlinear analysis to short ecological time series// Am Nat. – 2008. – Vol.171(1). – pp. 71–80.
30. Kirshners A. Clustering-based behavioural analysis of biological objects// Environment. Technology. Resources: Proceedings of the 8th International Scientific and Practical Conference. – 2011. – Vol.2. – pp. 24–32.
31. Kirshners A., Borisov A. A Comparative analysis of short time series processing methods// Scientific Journal of Riga Technical University, Information Technology and Management Science. – 2012. – Vol.15. – pp. 65–69.
32. Kirshners A., Borisov A. Analysis of short time series in gene expression tasks// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2010. – Iss.5, Vol.44. – pp. 144–149.
33. Kirshners A., Borisov A. Multilevel classifier use in a prediction task// Proceedings of the 17th International Conference on Soft Computing. – 2011. – pp. 403-410.
34. Kirshners A., Borisov A. Processing short time series with data mining methods// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2011. – Iss.5, Vol.49. – pp. 91–96.
35. Kirshners A., Borisov A., Parshutin S. Robust cluster analysis in forecasting task// Proceedings of the 5th International Conference on Applied Information and Communication Technologies (AICT2012). – 2012. – pp. 77–81.
36. Kirshners A., Kornienko Y. Time-series data mining for e-service application analysis// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2009. – Iss.5, Vol.40. – pp. 94–100.
37. Kirshners A., Kuleshova G., Borisov A. Demand forecasting based on the set of short time series// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2010. – Iss.5, Vol.44. – pp. 130–137.
38. Kirshners A., Liepinsh E., Parshutin S., Kuka J., Borisov A. Risk prediction system for pharmacological problems// Automatic Control and Computer Sciences. – 2012. – Vol.46, No.2. – pp. 57–65.

39. Kirshners A., Parshutin S. Application of data mining methods in detecting of bacteria proliferation syndrome in the small intestine// In: European Conference on Data Analysis 2013: Book of Abstracts: European Conference on Data Analysis 2013. – 2013. – pp. 139–139.
40. Kirshners A., Parshutin S., Borisov A. Combining clustering and a decision tree classifier in a forecasting task// Automatic Control and Computer Science. – 2010. – Vol.44, No.3. – pp. 124–132.
41. Kirshners A., Parshutin S., Leja M. Research in application of data mining methods to diagnosing gastric cancer// LNAI 7377. Proceedings of the 12th Industrial Conference on Data Mining ICDM'2012. – 2012. – pp. 24–37.
42. Kirshners A., Sukov A. Rule induction for forecasting transition points in product life cycle data// Scientific Proceedings of Riga Technical University, Information Technology and Management Sciences. – 2008. – Iss.5, Vol.36 – pp. 170–177.
43. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection// Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI-95). – San Mateo, CA: Morgan Kaufman. – 1995. – pp. 1137–1143.
44. Kohavi R., Quinlan J. R. Decision-tree discovery// Handbook of Data Mining and Knowledge Discovery. – Klosgen W., Zytkow J.M., Eds. – Oxford: Oxford University Press. – 2002. – pp. 267–276.
45. Koza J. R. Genetic Programming: On the Programming of Computers by Means of Natural Selection. – Cambridge, MA: The MIT Press, 1992. – 840 p.
46. Lembcke B. Atemtests bei Darmkrankheiten und in der gastroenterologischen Funktionsdiagnostik. Schweiz// Rundschau Medizin (Praxis). – 1997. – Vol.86. – pp. 1060–1067.
47. Liepinsh E., Vilskersts R., Loca D., Kirjanova O., Pugovichs O., Kalvinsh I., Dambrova M. Mildronate, an inhibitor of carnitine biosynthesis, induces an increase in gamma-butyrobetaine contents and cardioprotection in isolated rat heart infarction// J Cardiovasc Pharmacol. – 2006. – Vol.48(6). – pp. 314–319.
48. Liepinsh E., Vilskersts R., Skapare E., Svalbe B., Kuka J., Cirule H. et al. Mildronate decreases carnitine availability and up-regulates glucose uptake and related gene expression in the mouse heart// Life Sci. – 2008. – Vol.83. – pp. 613–619.
49. Liepinsh E., Vilskersts R., Zvejniece L., Svalbe B., Skapare E., Kuka J. et al. Protective effects of mildronate in an experimental model of type 2 diabetes in Goto-Kakizaki rats// British Journal of Pharmacology. – 2009. – Vol.157. – pp. 1549–1556.
50. Lupascu A., Gabrielli M., Lauritano E. C., Scarpellini E., Scantoliquido A., Cammarota G., Flore R., Tondi P., Pola P., Gasbarrini G.,Gasbarrini A. Hydrogen glucose breath test to detect small intestinal bacterial overgrowth: a prevalence case-control study in irritable bowel syndrome// Aliment Pharmacol Ther. – 2005. – 22(11-12). – pp. 1157–1160.
51. McLachlan G., Krishnan T. The EM algorithm and extensions, 2nd edition. Wiley series in probability and statistics. – John Wiley & Sons, 2008. – 400 p.
52. Montgomery D. C., Jennings C. L., Kulachi M. Introduction to Time Series Analysis and Forecasting. – Wiley-interscience, 2008. – 445 p.
53. Parshutin S., Aleksejeva L., Borisov A. Forecasting product life cycle phase transition points with modular neural networks based system// Proceedings of 9th Industrial Conference on Data Mining ICDM`2009, Springer-Verlag. – 2009. – LNAI 5633. – pp. 88–102.
54. Parshutin S., Kirshners A. Intelligent agent technology in modern production and trade management// Efficient Decision Support Systems: Practice and Challenges – From Current to Future/ Book Chapter. INTECH. – 2011. – pp. 21–42.

55. Parshutin S., Kirshners A. Research on clinical decision support systems development for atrophic gastritis screening// Expert Systems with Applications. – 2013. – Vol.40, Iss.15. – pp. 6041–6046.
56. Posserud I., Stotzer P. O., Bjornsson E. S., Abrahamsson H., Simren M. Small intestinal bacterial overgrowth in patients with irritable bowel syndrome// Gut. – 2007. – Vol.56(6). - pp. 802–808.
57. Pyle D. Data Preparation for Data Mining. – San Francisco etc.: Morgan Kaufmann, 1999. - 540 p.
58. Quinlan J. R. C4.5: Programs for Machine Learning. – San Mateo: Morgan Kaufmann Pub., 1993. – 302 p.
59. Russell S. J., Norvig P. Artificial Intelligence: A Modern Approach – Prentice-Hall, Inc., 1995. – 932 p.
60. Salam A. Najim, Zakaria A. M. Al-Omari, Samir M. Said. On the application of artificial neural network in analyzing and studying daily loads of Jordan power system plant// Computer Science and Information Systems. – 2008, – Vol.5, Iss.1. – pp. 127–136.
61. Sjakste N., Gutcaits A., Kalvinsh I. Mildronate: An antiischemic drug for neurological indications// CNS Drug Reviews. – 2005. – Vol.11(2). – pp. 151–168.
62. Starzyk, J. A., Haibo H., Yue L. A Hierarchical Self-organizing Associative Memory for Machine Learning// Advances in Neural Networks. – 2007. – pp. 413–423.
63. Tan P. N., Steinbach M., Kumar V. Introduction to Data Mining. – Boston: Pearson Addison-Wesley, 2006. – 769 p.
64. Thomassey S., Fiordaliso A. A hybrid sales forecasting system based on clustering and decision trees// Decision Support Systems. – 2006. – Vol.42, Iss.1. – pp. 408–421.
65. Thomassey S., Happiette M. A neural clustering and classification system for sales forecasting of new apparel items// Applied Soft Computing. – 2007. – Vol.7. – pp. 1177–1187.
66. Thomassey S., Happiette M., Castelain J. A global forecasting support system adapted to textile distribution// International Journal of Production Economics. – 2005. – Vol.96. – pp. 81–95.
67. Thomassey S., Happiette M., Castelain J. A short and mean - term automatic forecasting system – application to textile logistics// European Journal of Operations Research. – 2005. – Vol.161. – pp. 275–284.
68. Toshniwal D., Joshi R. C. Similarity search in time series data using time weighted slopes// Informatica, An International Journal of Computing and Informatics. – 2005. – Vol.29, No.1. – pp. 79–88.
69. Ward J. H., Jr. Hierarchical Grouping to Optimize an Objective Function// Journal of the American Statistical Association. – 1963. – Vol.58. – pp. 236–244.
70. Witten I. H., Frank E. Data mining: Practical machine learning tools and techniques (Second edition). – San Francisco, CA: Morgan Kaufmann, 2005. – 560 p.
71. Wiener N. Cybernetics: Or Control and Communication in the Animal and the Machine. – Boston, MA: Technology Press, 1948. – 219 p.
72. Wang X., Wu M., Li Z., Chan C. Short time-series microarray analysis: Methods and challenges// BMC Systems Biology – 2008. – 2:58. – pp. 1–6.
73. Wu X., Kumar V., Quinlan J.R., et al. Top 10 algorithms in data mining// Knowl. Inf. Syst. – 2007. – 14. – pp. 1–37.
74. Zhu W., Zeng. N., Wang N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementation// NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland. – 2010. – pp. 1–9.
75. Zurada J. M. Introduction to Artificial Neural Systems. – West: St. Paul, MN, 1992. – 679 p.
76. Барсегян А., Куприянов М., Степаненко В. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – Санкт-Петербург: БХВ - Петербург, 2007. – 384 с.
77. Гринглаз Л., Копытов Е. Математическая статистика с примерами решения на компьютере: Учеб. Пособие. – 2-е изд. – Рига: ВШЭК, 2002. – 326 с.