TIME-SERIES DATA MINING FOR E-SERVICE APPLICATION ANALYSIS

DATU IEGUVE LAIKA RINDĀS E-PAKALPOJUMU ANALĪZEI

A. Kirshners, Y. Kornienko

Keywords: E-service analysis, time series, data mining, classification, decision

Abstract - This paper provides application analysis of e-services available on the joint state and municipal e-service portal www.latvija.lv . The research is performed using a combination of time series analysis and data mining techniques. Time series analysis has enabled the determination of the count of clusters that represent services classification by application frequency. Meta-information is processed using data pre-processing methods and the values obtained are then discretised. The methods combinations examined in the paper are tested experimentally on the limited data amount available. The data describe the existing e-service requests by months. The clusters obtained are then added to the initial meta-information available when planning and developing services. E-service membership in the formed data set is determined using inductive classification trees. These algorithms represent knowledge in the form of classification trees through analysing feature values and cyclically split training instances into classes. As a result, based on the analysis conducted, recommendations for e-service developers and implementers are elaborated and basic parameters for successful introduction and application of e-services are determined.

Introduction

The development and introduction of electronic services of state and municipal institutions is a high-priority task for all developed countries of the European Union including Latvia. An approach using service-oriented architecture ensures a lot of advantages for integrating and developing public service system infrastructure. Currently there are available 26 e-services on the state portal http://www.latvija.lv; five more services are planned to be introduced by the end of 2009.

E-service is in essence just a new information supply kind for customers who save their resources and time when receiving the final service result (document, reference or data submission to the system). As a result of service electronisation a logical question arises to which there is usually no answer: How many residents will use a new e-service? Will the new service become popular among them? The task of the paper is to forecast demand for the new e-service using all available information including:

- Time-series – e-service demand over a certain time period;
- E-service description [11], where each e-service is described by 19 attributes; this information is already available and is systematised.

Within this paper the problem of demand evaluation during a certain time period is solved by mining and analysing the data available. The data are combined in groups (clusters) by adding service describing data with the help of decision trees and a connection between describing data and the groups derived, is determined.

Work Objective, Tasks and Plan of Experiments

The objective of this paper is to analyse application statistics of e-services available on the state and municipal e-service portal. It can be divided into these tasks:

- Discovering e-service usage patterns and grouping them according to regularities;
- Developing and analysing a model that determines e-service type;
- Forecasting the demand for a new e-service;
- Developing recommendations for further service electronisation as to which service kinds could be potentially on demand among inhabitants.

To solve the task of this paper, the following experiments have to be conducted:

- Available data analysis;
- Selected data pre-processing (e-service attribute value reduction) and transformation (time series);
- Discovering e-service types (clusters);
- Developing a model for discovering e-service membership in a certain type (class);
- Forecasting new e-service application.

Data Analysis and Pre-Processing

Each of the 26 e-services available on the joint state and municipal e-service portal is characterised by a certain attribute count. To develop a dataset for mining

information, it is necessary to reduce and transform the data related to this data service. Since six of the e-services offered are available within some months only, they are not used in data mining. Demand data are summarised for the period from July, 2008 to March, 2009. July 2008 is chosen as a support base because starting from that time authentication became accessible for users employing internet bank services, which rapidly enlarged the e-service user society.

The statistics available in essence contains available count of requests within e-services, which means how many requests of question-answer type, were executed, because to execute a single service, an institution usually needs several question-answer type requests which are dependent on the logic of service framework.

E-service Attributes Pre-Processing

The e-services offered to users on the portal are divided into groups (categories). Each e-service is described by 19 attributes, five of which were cancelled during pre-processing since it was impossible to discretise their parameters. The remaining set consists of 14 attributes whose values are discretised:

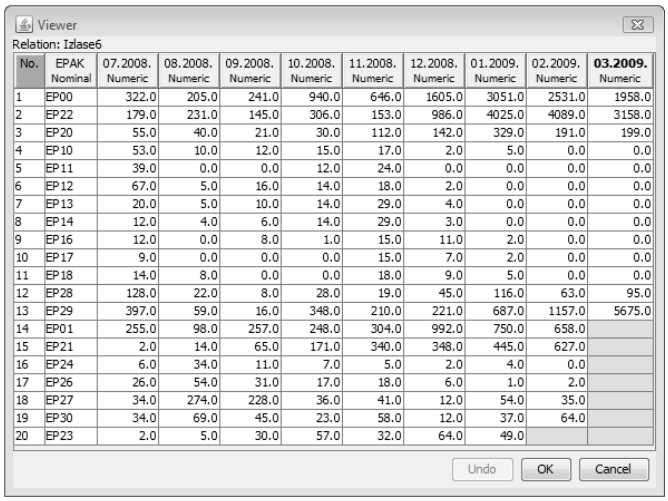
- Name of service (each service has its own unique name; it is used for parameter identification);
- A short description of a service (each service has its own description, so the attribute was not employed);
- Competent institution (values: 1 – state institution; 2 – municipal institution);
- Categories that characterise the service (values: 1 – place of residence, real estate and building; 2 - family, children, health, social services; 3 – commercial activity; 4 – rights defence, personal status, consumer rights, state purchases; 5 - transport, tourism, consular services, migration services). Besides, each service might belong to several categories at the same time; to describe this service, a conjunction of categories is used correspondingly;
- E-service type (values: 1 – synchronous; 2 - asynchronous);
- Authentication – characterises authentication type (values: 0 – anonymous user; 2.5 – smart card and bank; 3 – smart cards);
- Additional terms – indicate if the given service is subject to additional terms and conditions (value: 1 - yes; 2 - no);
- Charges (value: 1 – free; 2 - paid; 3 – partly free);
- Documents for service receipt (value: 1 – are not necessary; 2 – are necessary);
- Service result – a resident receives a document or submits data to the institution information system (value: 1 – various kinds of electronic activity; 2 – paper document; 3 – information reference);
- Stop points – stop points in the document flow system (the attribute will not be used);
- Identifier (unique value for each record - was not used because the unique identifier has already been employed);
- Service recipient is any person (value: 1 - no; 2 - yes);
- Service recipient is a physical person (value: 1 - no; 2 - yes);
- Service recipient is private authority (value: 1 - no; 2 - yes);
- Service recipient is public authority (value: 1 - no; 2 - yes);
- Connection with administrative process (value: 1 - no; 2 - yes).

E-service Demand Data Pre-Processing

E-service demand data can be characterised by 20 records that describe particular e-service demand within a certain time period. Attribute description contains this information:

- EPAK – E-service number;
- 07.2008. – 03.2009. – E-service provision period by months.

E-service demand data pre-processing is reduced via data transformation. Data transformation is value transfer to the left-hand part without changing data meaning [1]. Data transformation is necessary to enable analysis of time series with the same starting period. As a result of transformation, a data set is obtained that is shown in Fig.1. E-services EP01, EP21, EP24, EP26, EP27, EP30, EP23 were shifted because they were introduced later than other services.


Viewer

Relation: Izlase6

No.	EPAK Nominal	07.2008. Numeric	08.2008. Numeric	09.2008. Numeric	10.2008. Numeric	11.2008. Numeric	12.2008. Numeric	01.2009. Numeric	02.2009. Numeric	**03.2009.** Numeric
1	EP00	322.0	205.0	241.0	940.0	646.0	1605.0	3051.0	2531.0	1958.0
2	EP22	179.0	231.0	145.0	306.0	153.0	986.0	4025.0	4089.0	3158.0
3	EP20	55.0	40.0	21.0	30.0	112.0	142.0	329.0	191.0	199.0
4	EP10	53.0	10.0	12.0	15.0	17.0	2.0	5.0	0.0	0.0
5	EP11	39.0	0.0	0.0	12.0	24.0	0.0	0.0	0.0	0.0
6	EP12	67.0	5.0	16.0	14.0	18.0	2.0	0.0	0.0	0.0
7	EP13	20.0	5.0	10.0	14.0	29.0	4.0	0.0	0.0	0.0
8	EP14	12.0	4.0	6.0	14.0	29.0	3.0	0.0	0.0	0.0
9	EP16	12.0	0.0	8.0	1.0	15.0	11.0	2.0	0.0	0.0
10	EP17	9.0	0.0	0.0	0.0	15.0	7.0	2.0	0.0	0.0
11	EP18	14.0	8.0	0.0	0.0	18.0	9.0	5.0	0.0	0.0
12	EP28	128.0	22.0	8.0	28.0	19.0	45.0	116.0	63.0	95.0
13	EP29	397.0	59.0	16.0	348.0	210.0	221.0	687.0	1157.0	5675.0
14	EP01	255.0	98.0	257.0	248.0	304.0	992.0	750.0	658.0	
15	EP21	2.0	14.0	65.0	171.0	340.0	348.0	445.0	627.0	
16	EP24	6.0	34.0	11.0	7.0	5.0	2.0	4.0	0.0	
17	EP26	26.0	54.0	31.0	17.0	18.0	6.0	1.0	2.0	
18	EP27	34.0	274.0	228.0	36.0	41.0	12.0	54.0	35.0	
19	EP30	34.0	69.0	45.0	23.0	58.0	12.0	37.0	64.0	
20	EP23	2.0	5.0	30.0	57.0	32.0	64.0	49.0		

Undo OK Cancel

Fig. 1. E-service demand over a certain time period

Determination of Cluster Count

To determine the count of groups (clusters), in time series (request) comparison tasks [4,6] the k-mean algorithm is commonly employed that uses for that purpose the Euclidean distance and determines the similarity measure according to which classification into groups is made. Since group bounds are not expressed strictly in the case under consideration, the *Expectation maximisation algorithm* [7,8] (EM) is used. The algorithm of maximal likelihood employs probability measure instead of distance. The algorithm analyses classification curves for each dimension and each point is ascribed to a particular cluster with a certain probability. This technique is called soft clustering because the groups have no strongly expressed bounds and may overlap. Also, data mining free software *Weka 3.6* with built-in algorithm *Simple EM (Expectation maximisation class*) [6, 7] that checks if it is possible to combine data obtained during the pre-processing in clusters, is used. As a result of model run, these cluster groups were obtained (see Fig.2):

- Cluster 0 – e-services EP00, EP22, EP29, EP01 and EP21;
- Cluster 1 – e-services EP20, EP10, EP11, EP12, EP13, EP14, EP15, EP16, EP17, EP28, EP24, EP26, EP27, EP30 and EP23.

The clustering results obtained show that e-service demand division in clusters is carried out taking into account the number of e-service requests made each month. An analysis of e-service request graphs confirms the derived results. As can be seen from Fig.3, the graph depicts requests that are combined using cluster number 0 and the number of requests exceeds on average 330 times a month.

In its turn the graph shown in Fig.4 summarizes the requests combined in cluster 1; their demand does not exceed 330 times a month. In that way a hypothesis is confirmed that clusters were created on the basis of e-service request count.

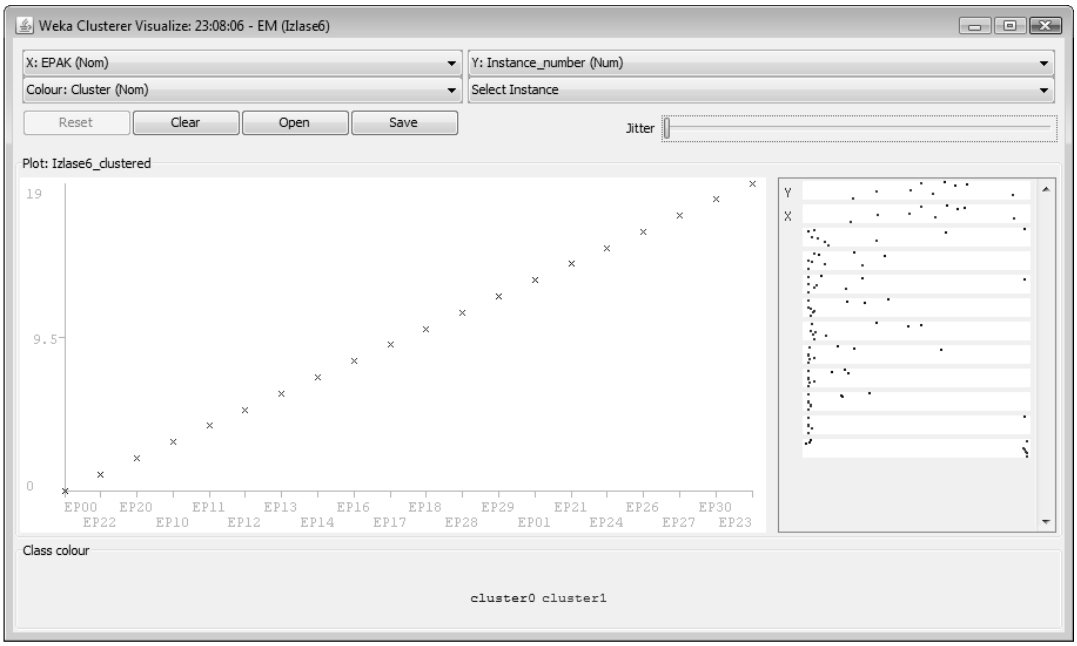


Fig. 2. Cluster count determination

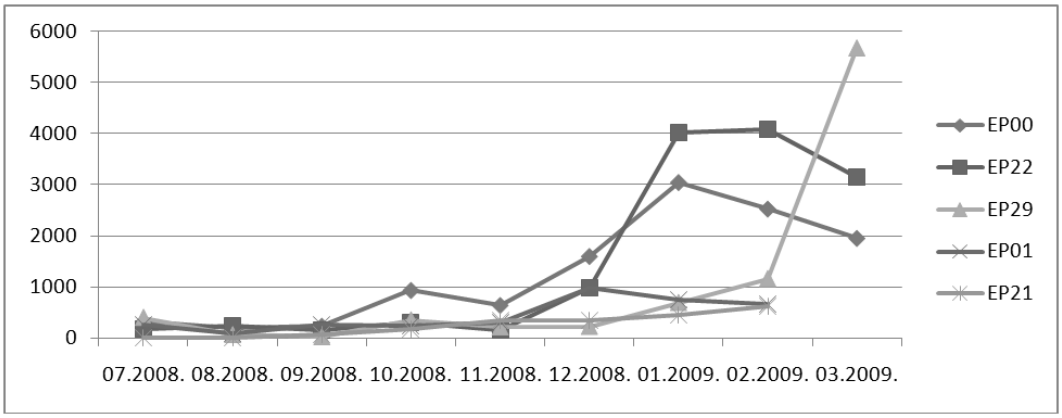


Fig. 3. E-service demand (cluster - «0»)

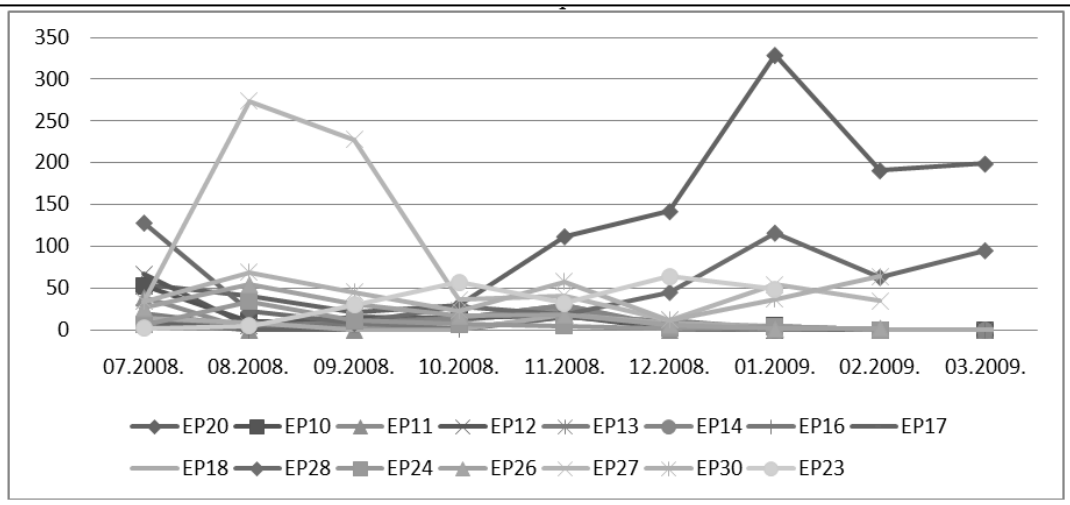


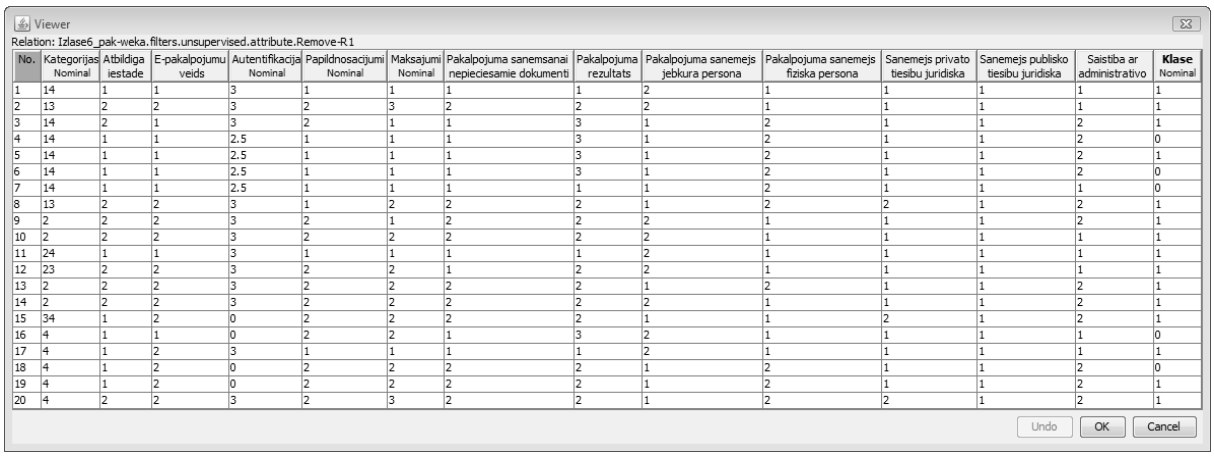
Fig. 4. E-service demand (cluster - «1»)

A model for Determining E-service Membership Type

To develop a model that defines e-service membership in a certain type (cluster), it is necessary to combine discretised e-service attribute and demand data obtained as a result of data pre-processing [2]. To e-service attribute data there is added the number of group (cluster) obtained in the previous experiment as a result of clustering and based on the service identification numbers of both datasets. The new dataset that is used in the experiments can be seen in Fig.5. The dataset is represented using tool's *Weka 3.6* module *Viewer* that ensures proper dataset representation.

Experiments and Results

Based on the developed dataset, e-service membership can be best determined employing inductive inference [3] which is used by classification tree algorithms ID3 [10] and ID3-GA [5]. These algorithms represent knowledge in the form of decision tree that helps to define object classification by analysing the values of its features. The ID3 algorithm cyclically splits instances of a training set into classes according to the variable (attribute) that has the largest classification power. Each subset of instances issued by that variable is anew divided into classes using the next variable with the largest classification power and so on. The splitting is finished when objects of only one class are present in the subset. As a result, a decision tree is created. To find an attribute with the least homogeneous class distribution, the algorithm is employing the notion of entropy [1,10]. Using a decision tree, forecasting models are constructed with whose help e-service membership in a certain type – cluster 0 or 1, is determined. Taking into account that test data are insufficient (only 20 records are available), training set data are used as test data. To evaluate the obtained model adequately, one has to check classifier performance. The model derived is evaluated using the 10-fold cross validation, which means that the set employed by the model is partitioned into ten equal parts, each of which once serves as a test set but nine times as a training set.


Viewer

Relation: Izlase6_pak-weka.filters.unsupervised.attribute.Remove-R1

No.	Kategorijas Nominal	Atbildiga iestade	E-pakalpojumu veids	Autentifikacija Nominal	Papildnosacijumi Nominal	Maksajumi Nominal	Pakalpojuma sanemsanai nepieciesamie dokumenti	Pakalpojuma rezultats	Pakalpojuma sanemejs jebkura persona	Pakalpojuma sanemejs fiziska persona	Sanemejs privato tiesibu juridiska	Sanemejs publisko tiesibu juridiska	Saistiba ar administrativo	**Klase** Nominal
1	14	1	1	3	1	1	1	1	2	1	1	1	1	1
2	13	2	2	3	2	3	2	2	2	1	1	1	1	1
3	14	2	1	3	2	1	1	3	1	2	1	1	2	1
4	14	1	1	2.5	1	1	1	3	1	2	1	1	2	0
5	14	1	1	2.5	1	1	1	3	1	2	1	1	2	1
6	14	1	1	2.5	1	1	1	3	1	2	1	1	2	0
7	14	1	1	2.5	1	1	1	1	1	2	1	1	1	0
8	13	2	2	3	1	2	2	2	1	2	2	1	2	1
9	2	2	2	3	2	1	2	2	2	1	1	1	2	1
10	2	2	2	3	2	2	2	2	2	1	1	1	1	1
11	24	1	1	3	1	1	1	1	2	1	1	1	1	1
12	23	2	2	3	2	2	1	2	2	1	1	1	1	1
13	2	2	2	3	2	2	2	2	1	2	1	1	2	1
14	2	2	2	3	2	2	2	2	2	1	1	1	2	1
15	34	1	2	0	2	2	2	2	1	1	2	1	2	1
16	4	1	1	0	2	2	1	3	2	1	1	1	1	0
17	4	1	2	3	1	1	1	1	2	1	1	1	1	1
18	4	1	2	0	2	2	2	2	1	2	1	1	2	0
19	4	1	2	0	2	2	2	2	1	2	1	1	2	1
20	4	2	2	3	2	3	2	2	1	2	2	1	2	1

Undo OK Cancel

Fig. 5. A model for making experiments

Testing the Model with the ID3 Algorithm

The constructed model is tested with the help of tool's *Weka 3.6* built-in algorithm ID3; as a result, a forecasting module is produced whose classification tree is depicted in Fig. 6. When testing the given decision tree, using the 10-fold cross validation, 14 of 20 instances (or 70%) were classified correctly. The total classification error is 30%. Recognition support performance for class 1 is 86.7%, but for class 0 is 20%.

Model analysis has provided the following expert's comments: the main attribute according to which the data have been distributed is Authentication. This attribute could serve as the key one, since it indicates that less popular services need authentication with an electronic signature. However, it has to be concluded that the selected second level attributes are very subjective and cannot be determined unambiguously when analysing one or another service that is a candidate for electronisation.

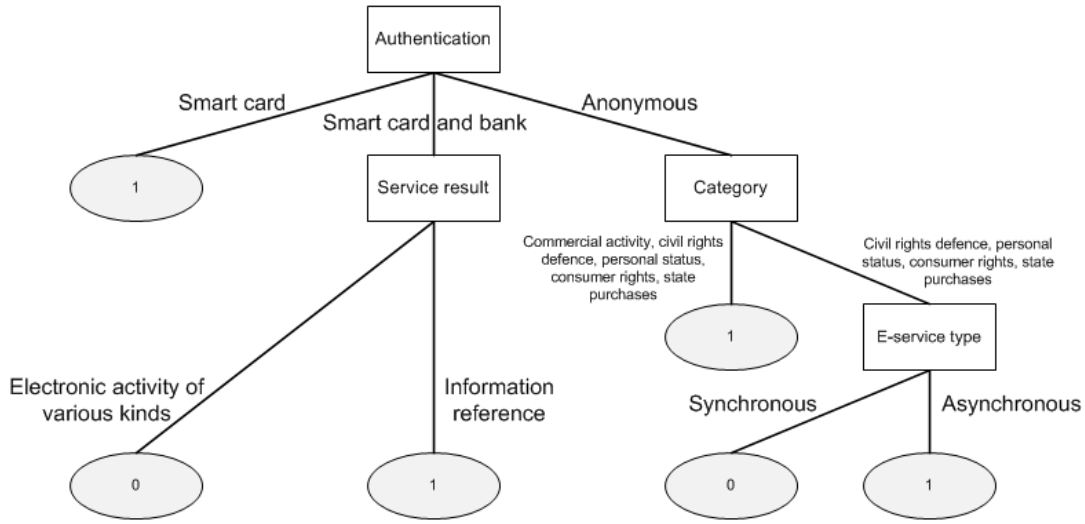


Fig. 6. Decision tree constructed with the ID3 algorithm

Model Testing with the ID3-GA Algorithm

In the experiment that is following, the *ID3-GA* algorithm has been employed (implemented as an extension of *MLC++* library) that combines genetic algorithms and decision trees [5]. This algorithm differs from standard ID3 algorithm in that it creates a classifier ensemble that applies less informative attributes to construct classifiers. Similarly to the previous experiment, training set data were used as test data and 10-fold cross validation was employed. As a result, a forecasting module was obtained (see Fig. 7). According to test results, 15 of 20 instances (or 75%) were classified correctly. The total classifier error is 25%. Recognition support performance for class 1 is 80%, but for class 0 is 60%.

The ID3-GA algorithm has chosen „*Competent institution*" as a major attribute; if the competent institution is municipal institution, the algorithm always selects cluster number 1, but if cluster number is 0, the classification tree continues branching and selects attribute „Service recipient is private authority". If service recipient is always private authority, then the selected services will have cluster number 1; otherwise further branching of the tree occurs and the algorithm chooses attribute „*Authentication"*.

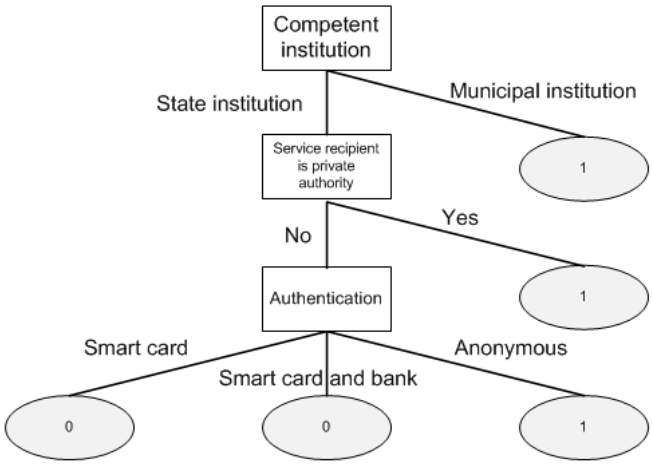


Fig. 7. Model tree using the ID3-GA algorithm

By analysing model results, these experts comments were derived: the constructed tree helps formulate service popularity to conceive and check it with a

simple rule Most popular services are those that are offered to the residents by state institutions (potential auditorium is the entire state), and those that are anonimous or available with bank authentication, which once again indicates that the potential auditorium is the whole state. That rule in essence nearly precisely represents expert's forecasts and characterises the key attributes for those new services analysis which are planned to be implemented in the near future.

Recommendations for Branch Experts

Whenever possible, e-service developers should pay more attention to anonimous or bank authenticated services. As experimental results show, in model testing with classification tree algorithms an attribute „Authentication" was chosen in both cases, which points at residents' wish to use the system anonymously or employ bank authentication. It means that currently, according to the constructed model results, this attribute is determinative and gives evidence of residents' wish not to perform, if it is not necessary, authentication or registration on the portal of state government and municipal institutions.

Analysing the cluster distribution, one can state that most popular among residents are services offered by state institutions.

Conclusions

The tree structures discussed in this paper do not provide a complete picture of resident's choice selecting e-services on the state and municipal e-service portal; they simply outline guidelines for branch experts on which the services have to be based when making forecasts regarding new service introduction. In the course of experiments two cluster groups were distinguished, which indicates e-service request number within the corresponding time period. Based on these groups, a model was constructed with whose help experiments aimed at determining e-service membership were conducted. Taking into account experimental results, one can conclude that even at a very small data amount it is possible to make forecasts related to creating new e-services.

Based on the experimental results as well as taking into account that demand statistics will enlarge in the future, it will be possible to apply other techniques implementing a detailed data classification into groups (clusters) to interpret the constructed model in as much detail as possible for e-service developers and establishers.

By performing other kind of experiments, say, comparing e-service demand count with regard to curve similarity, it would be possible to create clusters according to time-series curve similarity [2,3,6].

References

1. Data Mining: Fundamentals / A. Sukovs, L. Aleksejeva, K. Makejeva et al. – Riga: Riga Technical University, 2007, p.130. (In Latvian).
2. Thomassey S., Fiordaliso A. A hybrid sales forecasting system based on clustering and decision trees. Decision Support Systems, Volume 42, Issue 1, 2006, p. 408-421.
3. Written I.H., Frank E. Data mining:practical machine learning tools and techniques - 2nd edition. - Amsterdam etc.: Morgan Kaufman, 2005.
4. Das, G., Lin, K., Manilla, H., Renganathan, G., Smyth, P. Rule Discovery from Time Series. // In Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining, 1998, p. 16-22.
5. Kornienko Y., Borisov A. Investigation of a Hybrid Algorithm for Decision Tree Generation. IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications 8-10 September 2003, Lviv, Ukraine. Retrieved 15 May, 2009 from http://www.umcs.maine.edu/~idaacs/2003/downloads/articles/063-068_(065).pdf.
6. Toshniwal D., Joshi R. C. Similarity in Time Series Data Using Time Weighted Slopes. Informatica, An International Journal of Computing and Informaatics, Volume 29, Number 1, May 2005, p. 79-88.
7. McLachlan G., Krishnan T. The EM algorithm and extensions. Wiley Series in Probability and Statistics. John Wiley & Sons, 1997.
8. Dellaert F. The Expectation Maximization Algorithm. College of Computing, Georgia Institute of Technology, Technical Report No. GIT-GVU-02-20, February 2002. Retrieved 15 May, 2009 from http://www.cc.gatech.edu/~dellaert/em-paper.pdf.
9. 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. Retrieved 15 May, 2009 from http://www.cs.ualberta.ca/~szepesva/ICML2008Health/Devisscher.pdf.
10. Technologies of Data Analysis: Data Mining, Visual Mining, Text Mining, OLAP / A.

Barsegyan, M. Kupriyanov, V. Stepanenko, I. Holod. St.Petersburg, 2007, p. 384. (In Russian).

11. Kornijenko J. E-service standard. (In Latvian). Retrieved 15 May, 2009 from https://ivis.eps.gov.lv/IVISPortal/files/folders/standarti/entry41.aspx.

Arnis Kirshners, is Master's Student and Senior Laboratory Assistant in the Department of Modelling and Simulation at Riga Technical University. He received his diploma of Bc.sc.ing. in Information Technology from Riga Technical University. His research interests include data mining and knowledge extraction, intelligent systems, programming, database and evolutionary computing.

Yuri Kornienko is currently Software Architect in the ABC Software Company (Riga). He received his Dr.sc.ing. degree from Riga Technical University in 2007. His major research interests include regression decision trees and cluster analysis.

Arnis Kiršners, Jurijs Korņijenko. Datu ieguve laika rindās E-pakalpojumu analīzei

Rakstā veikta portālā www.latvija.lv pieejamo valsts pārvaldes un pašvaldību iestāžu elektronisko pakalpojumu analīze. Pētījumos izmantotas metožu kombinācijas, kas apvieno laika rindu analīzi un datu ieguvi. Laika rindu analīze ļauj noteikt klasteru skaitu, kuri nosaka e-pakalpojumu sadalījumu pēc to pielietošanas biežuma. Meta informāciju apstrādā ar datu pirmapstrādes metodēm un iegūtās atribūtu vērtības tiek diskretizētas. Rakstā aplūkojamās metožu kombinācijas eksperimentāli pārbaudītas ar ierobežotu datu apjomu. Šie dati apraksta jau pastāvošus e-pakalpojumu pieprasījumus pa mēnešiem. Turpinājumā iegūtie klasteri tiek pakārtoti pievienoti sākotnējai meta informācijai, kas pieejama pakalpojuma plānošanas un izstrādes procesā. E-pakalpojumu piederību izveidotajai datu kopai nosaka, pielietojot induktīvo secināšanu, kuru izmanto klasifikācijas koki. Šie algoritmi zināšanas attēlo klasifikācijas koku veidā, analizējot īpašību vērtības un cikliski sašķeļ apmācības piemērus klasēs. Rezultātā, balstoties uz veikto analīzi, izstrādātas rekomendācijas e-pakalpojumu ieviesējiem un izstrādātājiem, noteikti galvenie parametri, kas veicina sekmīgu e-pakalpojumu ieviešanu un izmantošanu.

Арнис Киршнерс, Юрий Корниенко. Поиск данных временного ряда для анализа приложения Э-услуг

В статье проведен анализ использования Э-услуг, доступных на портале государственных и муниципальных электронных услуг www.latvija.lv. В исследовании использована комбинация методов анализа временных рядов и добычи данных. Анализ временных рядов позволил обосновать основные кластеры – распределение услуг по частоте использования. Метаинформацию обрабатывают с помощью методов предобработки данных, и полученное значение атрибутов дискретизируется. Рассмотренные в статье комбинации методов экспериментально проверены на ограниченном количестве данных. Эти данные описывают по месяцам уже существующий спрос на Э-услуги. В свою очередь, при помощи методов добычи знаний удалось провести анализ соответствия кластеров, полученных на предыдущем этапе, и первичной метаинформации, доступной при планировании и разработке услуг. Принадлежность Э-услуг сформированному множеству определяется с помощью индуктивных деревьев классификации. Эти алгоритмы отображают знания в виде деревьев классификации, анализируя значения признаков и циклически разделяя обучающие примеры в классы. В результате, на основе проведенного анализа, выработаны рекомендации для разработчиков Э-услуг, определены основные параметры успешного внедрения и использования Э-услуг.