Computer Science

Decision making based on data analysis methods

Miki Sirola, Mika Sulkava

Aalto University

SCIENCE + TECHNOLOGY

TECHNICAL REPORT

Aalto University publication series
SCIENCE + **TECHNOLOGY** 5/2016

Decision making based on data-analysis methods

Miki Sirola, Mika Sulkava

Aalto University
School of Science
Department of Computer Science
Applications of Machine Learning

Aalto University publication series
SCIENCE + **TECHNOLOGY** 5/2016

ISBN 978-952-60-6757-5 (printed)
ISBN 978-952-60-6758-2 (pdf)
ISSN-L 1799-4896
ISSN 1799-4896 (printed)
ISSN 1799-490X (pdf)
http://urn.fi/URN:ISBN:978-952-60-6758-2

Unigrafia Oy
Helsinki 2016

Finland

441 697
Printed matter

Author
Miki Sirola, Mika Sulkava

Name of the publication
Decision making based on data-analysis methods

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series SCIENCE + TECHNOLOGY 5/2016

Field of research Applications of Machine Learning

Abstract

This technical report is based on four our recent articles:"Data fusion of pre-election gallups and polls for improved support estimates", "Analyzing parliamentary elections based on voting advice application data", "The Finnish car rejection reasons shown in an interactive SOM visualization tool", and "Network visualization of car inspection data using graph layout". Neural methods are applied in political and technical decision making. We introduce decision support schemes based on Self-Organizing Map (SOM) combined with other methods. Visualizations based on various data-analysis methods are developed. In political decision making we have examples from one parliamentary election and one presidential election utilizing opinion data collected beforehand. In technical decision making we concentrate on rejection reasons in car inspection data. This technical report is a summary of our recent non-nuclear studies.

Keywords Decision Making, Data Analysis, Self-Organizing Map, Political, Technical

ISBN (printed) 978-952-60-6757-5	**ISBN (pdf)** 978-952-60-6758-2	
ISSN-L 1799-4896	**ISSN (printed)** 1799-4896	**ISSN (pdf)** 1799-490X
Location of publisher Helsinki	**Location of printing** Helsinki	**Year** 2016
Pages 18	**urn** http://urn.fi/URN:ISBN:978-952-60-6758-2	

Contents

Introduction

This technical report is a summary of recent non-nuclear studies of our research group. The recent studies related to nuclear power are already summarized in our previous technical report [1]. Before that two TEKES projects (NoTeS and NoTeS2) in an earlier MASI program together with our industrial partner TVO Olkiluoto were summarized in a technical report [2]. Altogether this is the third technical report summarizing our research during the last ten years.

This report includes topics of political and technical decision making. Political decision making has examples both from a parliamentary election and a presidential election. Data from pre-election gallups and voting advice applications are utilized. In technical decision making car inspection data and rejection reasons are analysed. Visualization techniques to support the decision making are an important part of these studies. Self-Organizing Map (SOM) [3] is one technique that has been utilized especially in visualization. One goal has been to generate friendly techniques to achieve better decision making on various application areas.

In this report each of these four studies are shortly presented and summarized. Scientific articles have been published, and one doctor's thesis [4]. The figures in this report have been published before also in the original papers by this research group found from the references.

Decision making supported by data analysis methods

In our research we have combined data analysis and decision making. We have developed tools to help the decision makers to understand better various phenomena in their application area. The used methods can find out significant features from the data.

The four following case studies enlighten the problem from different perspectives. Each case study gives their unique contribution to the whole toolset. Visualization is one important aspect, but the model output can be other too. The political decision making and the technical decision making have both their application oriented special features.

Data analysis can also be a tool in prediction. Both in political and in technical field this need can be recognized. The data for these studies is collected from publicly available sources. We try to show how the decision makers can get profit by these methodologies.

Political decision making

In political decision making we concentrate on data collected in preparation for elections. One study is about presidential election in Finland. Data fusion of pre-election gallups and polls are used to improve support estimates. Another study is about parliamentary elections in Finland. The analysis is based on voting advice application data. In both studies visualization is an important tool to improve the decision making.

Support estimates with pre-election gallups and polls in presidential election

Gallup results and a questionnaire in a voting advice application about Finnish presidential election are combined [5]. The main focus is on preprocessing phases where raw data is reformed to temporal data sets. Optimized parameters for a merged recursive model are found. Aggregated data from a questionnaire was stored frequently and modified by a differential equation. With this method the daily support of each candidate before the election is visualized more accurately. Forecasting the results and the success of presidential campaigns can be supported with the results.

A method for combining two different data sources is proposed to get more accurate estimate of election candidates' support in time. Voting advice applications (VAA) are increasingly popular in democracies worldwide, especially among young people [6]. Sampling over time by repeated gallups enables analysis of process through estimation [7]. In Figure 1 the data mining phases are described.

In left part of the Figure 1 citizens fill voting advice application and the answers are stored to database. Aggregated data is captured from the website and stored in our database every hour. In right part of the Figure 1 a collection of published gallup data sets are stored in our database. The computer analyzes the combined data sets. The results are used for election results prediction. Merged data can be used in the campaign analysis.

In Figure 2 a gallup preprocessing flow chart is presented. Support value matrix X and sample size S are inputs. Outputs are modified matrices including information of all candidate support values and sample sizes, see also Table 1. The phases in the Figure 2 are Gallup type change (F1), scaling (F2), sample resizing (F3), potential input size reduction (F4) and potential input size growth (F5). There is also a forced exit in

the loop. The rest of the phases are traditional Gallup (Q1), accurate Gallup (Q2), total number of supporters (Q3), and sum of scaled and rounded (Q4) with their varying sample sizes.

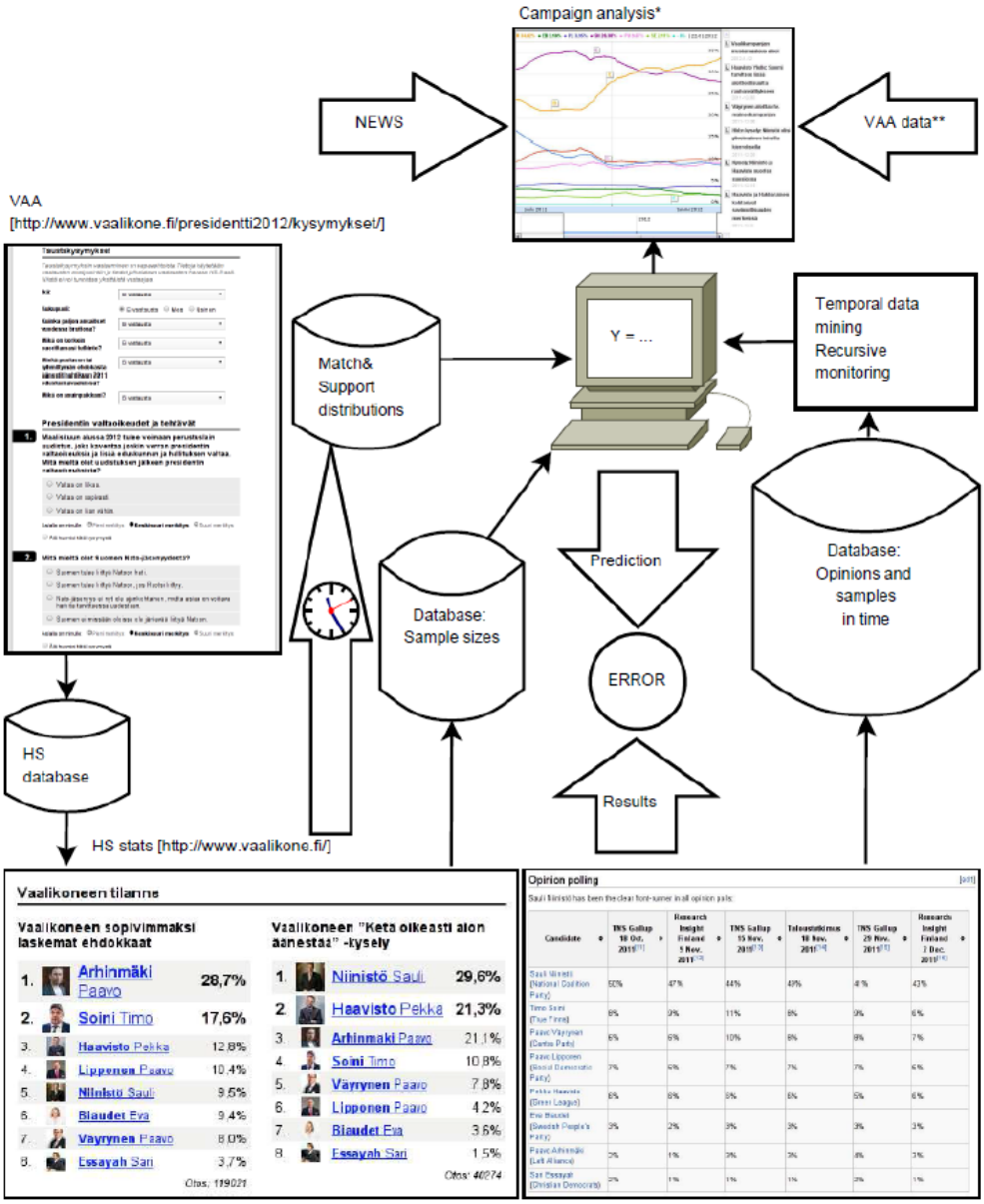

Figure 1. Data mining phases.

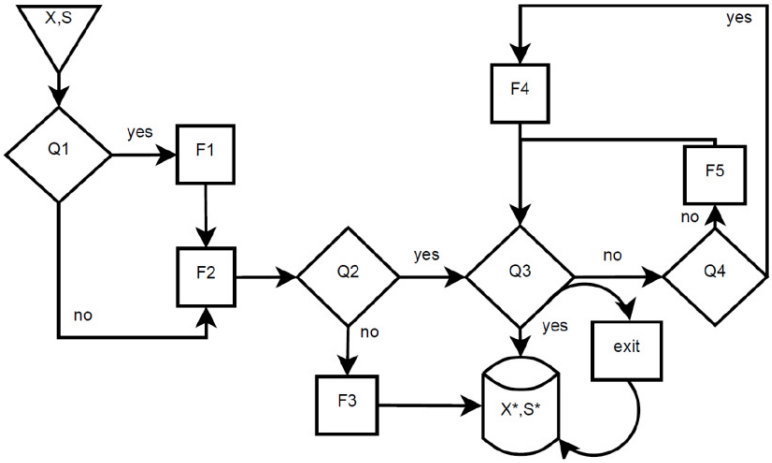

Figure 2. Gallup preprocessing.

In this presidential election in 2012 in Finland Sauli Niinstö ended up to a selected a president as a clear winner. From the second place there was a tight competition with three different candidates, see Figure 3. These three candidates were fighting for second position in the first round election, which opens the attendance on the second election round.

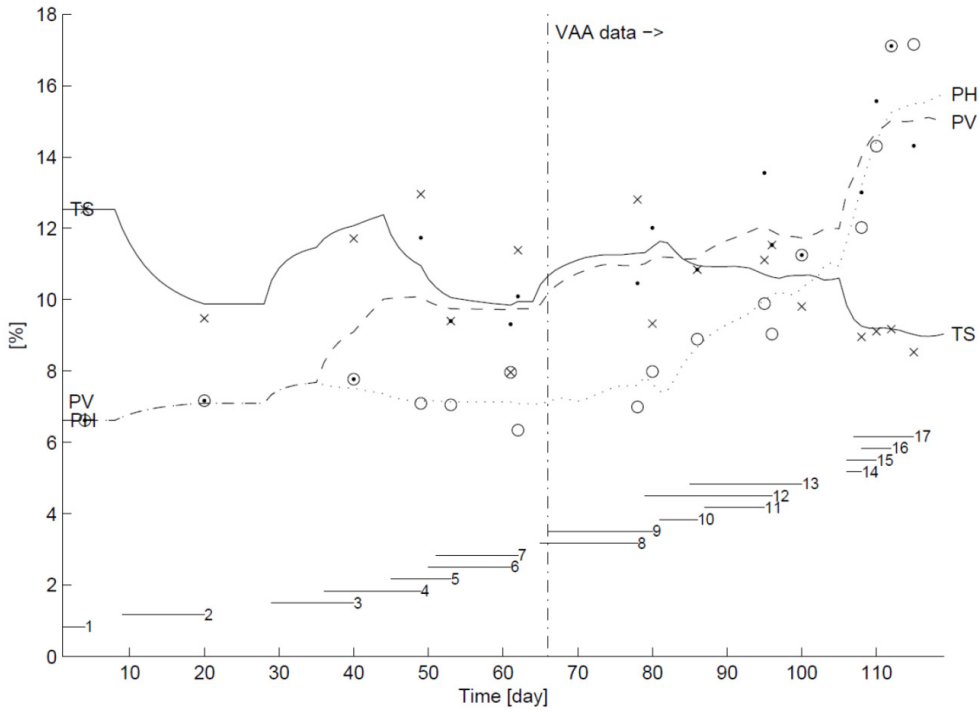

Figure 3. Competition of the second position in the first round of election.

The daily and Gallup support values for candidates are calculated using Gallup sample sizes. Timo Soini (solid line and x) was leading in the beginning. Paavo Väyrynen (dashed line and dots) and Pekka Haavisto (dotted line and circles) had both successful presidential campaigns. The horizontal numbered lines in the Figure 3 represents gallups and the line length represents the duration of the poll.

Table 1. Preprosessed Gallup results.

Table III. PREPROCESSED GALLUP RESULTS. START AND END DATES t, SAMPLE SIZES S WITH AND WITHOUT NON, AND CORRECTED SUPPORT NUMBERS. *) APPROXIMATED.

$t_{i,start}$	$t_{i,end}$	source	$S_i \bigcup non_i$	$S_i \setminus non_i$	PA	EB	SE	PH	PL	SN	TS	PV	non
1	4	RIF	1 010	862	1.68	1.68	0.69	5.64	10.69	48.61	10.69	5.64	14.65
9	20	TNS	980	823	2.04	2.96	2.04	6.02	6.94	50	7.96	6.02	16.02
29	40	RIF	*1053	811	1.04	1.99	1.04	5.98	5.03	46.91	9.02	5.98	22.98
36	49	TNS	*952	818	3.15	3.15	1.16	6.09	7.14	44.01	11.13	10.08	14.08
45	53	TT	1 452	1 234	3.03	3.03	0.96	5.99	7.02	48.97	7.99	7.99	15.01
50	61	RIF	1 000	741	2.9	2.9	0.9	5.9	5.9	42.8	5.9	6.9	25.9
51	62	TNS	978	773	3.99	2.97	2.04	5.01	7.06	41	9	7.98	20.96
65	78	TNS	1 979	1 702	5	3.99	1.97	6.01	6.01	43	11.02	8.99	14
66	80	TT	1 685	1 264	3.02	3.02	2.02	5.99	4.98	39.98	7	9.02	24.97
81	86	MC	1 000	821	3.2	2.4	2.4	7.3	7.3	41.7	8.9	8.9	17.9
87	95	TNS	1 011	819	2.97	0.99	1.98	8.01	6.03	41.05	9	10.98	18.99
79	96	RIF	1 473	1 162	4.14	2.1	2.1	7.13	7.13	38.09	9.1	9.1	21.11
85	100	TT	1 464	1 039	4.03	1.98	1.02	7.99	4.03	37	6.96	7.99	29.01
106	108	MC	1 000	815	5.7	1.6	2.5	9.8	4.1	39.9	7.3	10.6	18.5
106	110	RIF	1 014	790	3.16	2.07	2.07	11.14	3.16	37.08	7.1	12.13	22.09
108	112	TNS	1 408	970	4.19	1.49	2.13	11.79	4.19	26.99	6.32	11.79	31.11
107	115	TT	1 457	1 020	3.98	1.99	1.99	12.01	5.01	29.03	5.97	10.02	29.99

Table 1 presents the preprosessed Gallup results. The start and end days marked with t, sample sizes (with and without non) are marked with S. In table there are the corrected approximates support numbers.

In this study the used data consisted of 17 gallup studies and Website poll including 400000 answers. A novel method for combining temporal political support data sets was composed. An “Obama effect” could be noticed when the candidate Pekka Haavisto came up to the second election round.

Analysis of voting advice application data in parliamentary elections

The aim is to model the values of Finnish citizens and the members of parliament [8]. Two databases are combined: voting advice application data and the results of the parliamentary elections in 2011. The data is converted to a high-dimension space, and then it is projected to two principal components [9]. With the projections it is possible to visualize the main differences between the parties.

The value grids are produced with a kernel estimation method [10] without explicitly using the questions of the voting advice application. Meaningful interpretations for the axes in the visualizations are found with the analyzed data. All candidate value grids are weighted by the results of the parliamentary elections. The results can be interpreted as a distribution grid for Finnish voters’ values.

Principal component analysis [11] is the basic methodology used here. 17 parties participated in the Finnish parliamentary elections in 2011 [12]: four big parties and four small parties in the parliament, and nine parties not getting any members to the selected parliament. The six stages of parliamentary data analysis are seen in Figure 4.

The values of Finnish voters are seen in the Figure 5. Similar figures can be drawn for the values of candidates and for members in the parliament. The visualization in the Figure 5 can be used to get an estimate for one party how the opinions of the supporters are related to the supporters of other parties. Each party label is a mode of its supporters and those are in smaller font, if the density of the party

supporters in the mode is less than some other party's density. The numbers in contour lines describe the density of citizens in thousands.

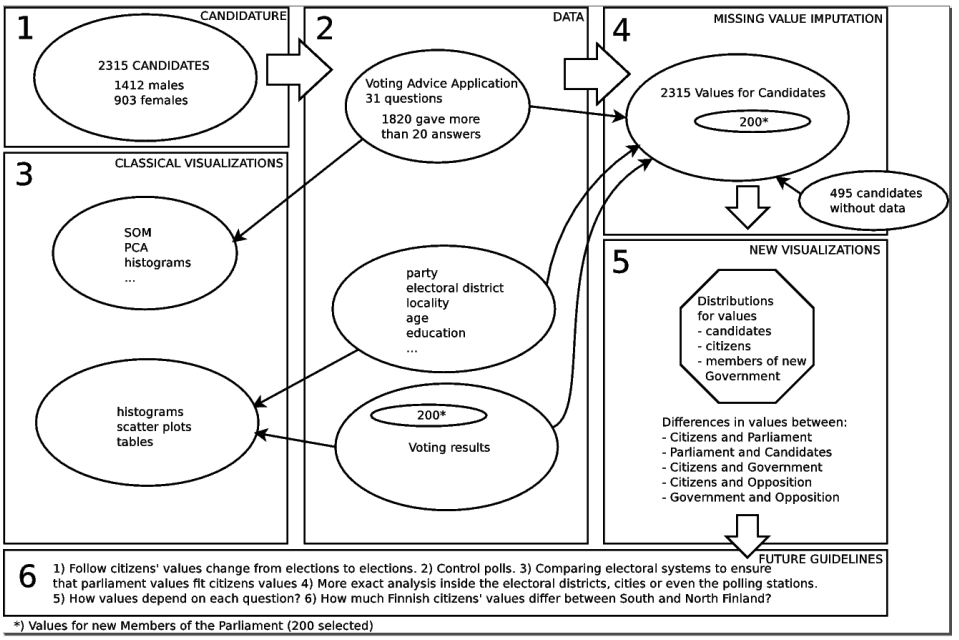

Figure 4. Parliamentary data analysis in six stages.

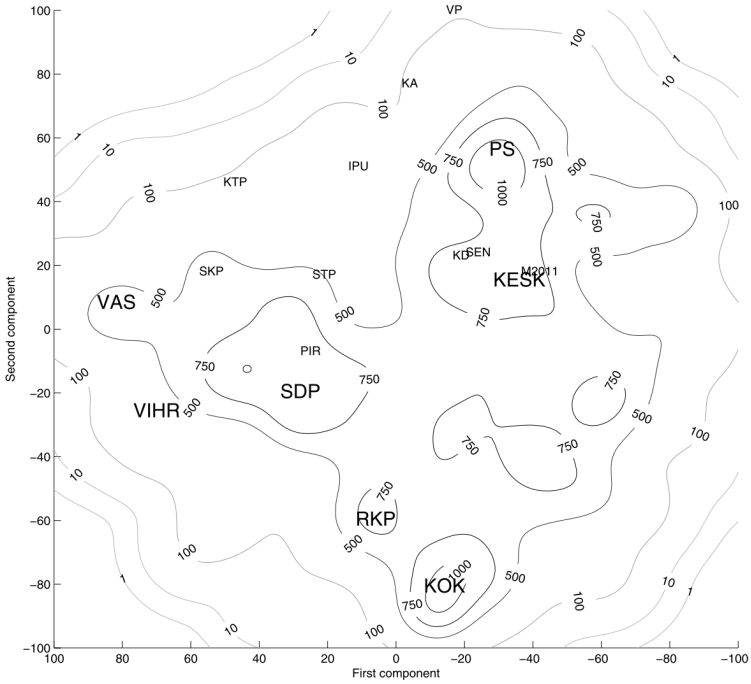

Figure 5. Finnish voters values.

Technical decision making

In technical decision making our case examples are utilizing the car inspection data. One of the studies concentrate on network visualization, and the other study is about rejection reasons shown with a special visualization tool. Self-Organizing Map (SOM) [3] method is used. Visualization of the data is again the key issue here.

Visualization of car inspection data

A network visualization based on the rejection reasons on car inspections is described [13]. A comparison with a visualization based on principal component analysis [9] is made. A-katsastus, which is the largest private provider of vehicle inspections in northern Europe, published rejection statistics in Finland for the fourth time. The statistics is published in numerous tables on the basis of the year introduction into use, make or model.

We aim to visualize all this information in one network. The car inspection data is aggregated with the produced visualization. The dependencies between the different rejection reasons and cars can be efficiently studied by exploring our network visualization.

A desktop application Gephi [14] is used in this study. It is an interactive visualization and exploration platform. The same tool is often used in social networks [15]. In Figure 6 PCA results are seen in a motion chart. Volkswagen Bora (2004) in the statistics of the year 2011 is projected to the far right. The first component loading for exhaust problem is positive.

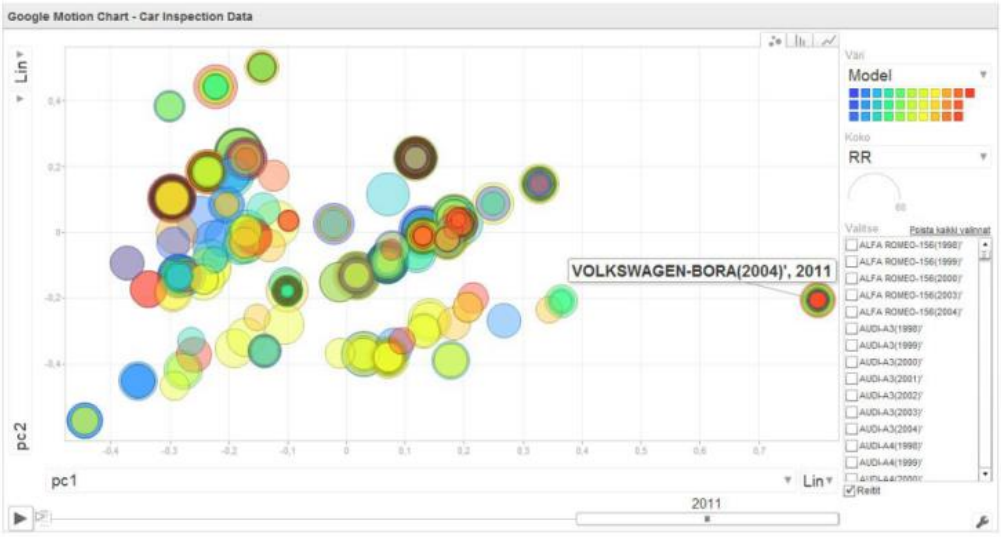

Figure 6. PCA results represented in a motion chart.

In Figure 7 the labels of cars with the parking brake rejection reason are highlighted. This is visualization for the rejection reason classes. Old cars are represented by blue balls and the newest by green balls. This is a kind of network of rejection reasons and cars, where the faults mentioned in the cars are connected to the

corresponding faults. The visualizations help in the exploration process. Somewhat similar visualization issues are discussed also in [16].

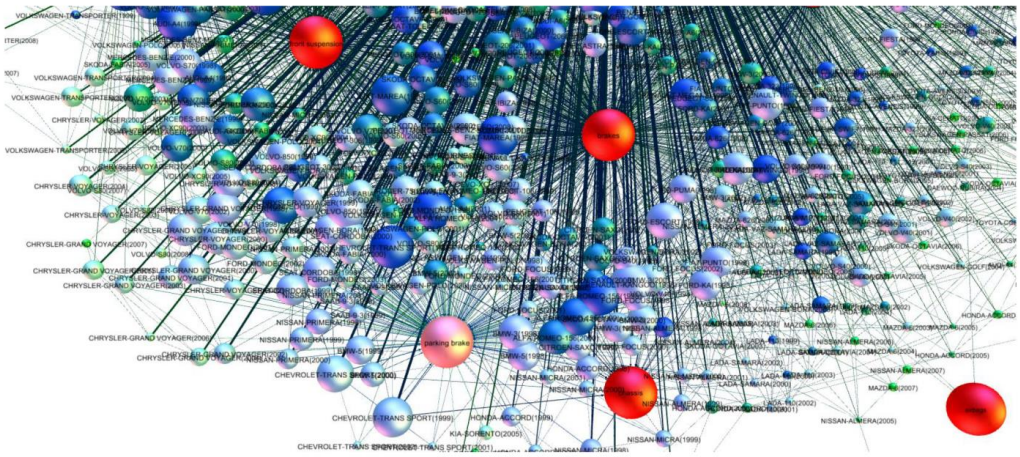

Figure 7. Rejection reason classes.

Visualization of car rejection reasons

A new SOM visualization tool is introduced [17]. Self-Organizing Map (SOM) method is used [3]. Collaborative filtering is used in preprocessing before the SOM training. The goal is to provide for the user a possibility to analyze car differences by component planes, which is not possible by the original published tables. It is also possible to explore how different flaws are related in time or with other variables.

It is possible to filter out the effects of driver dependent components, such as tires, from rejection probability using the component plane code books. The interactive SOM visualization is very convenient when a large number of labels are present. Here a function to generate the needed files for a processing language based tool was developed. Our tool can be used simultaneously with the SOM Toolbox [18].

In our experiments we could not find a proper tool for our visualization problem, although several SOM software packages exit [19]. There was an obvious need to develop a new one. The SOM Toolbox, which is flexible, general-purpose software library created by the SOM programming team of Helsinki University of Technology [20], is used in parallel with our new tool.

In Figure 8 processed car inspection data is shown in the SOM visualization tool. The labels can be shown or hidden by a mouse click. The car labels which have either small or large mask variable value are shown in this derived component plane. It is the proposition of the rejection reasons and average kilometers using codebook values. The best cars based on this feature are situated near to the center and on the top left part of the map. Vector components have some influence in the mapping. Grey hex borders allocate labels in the cells.

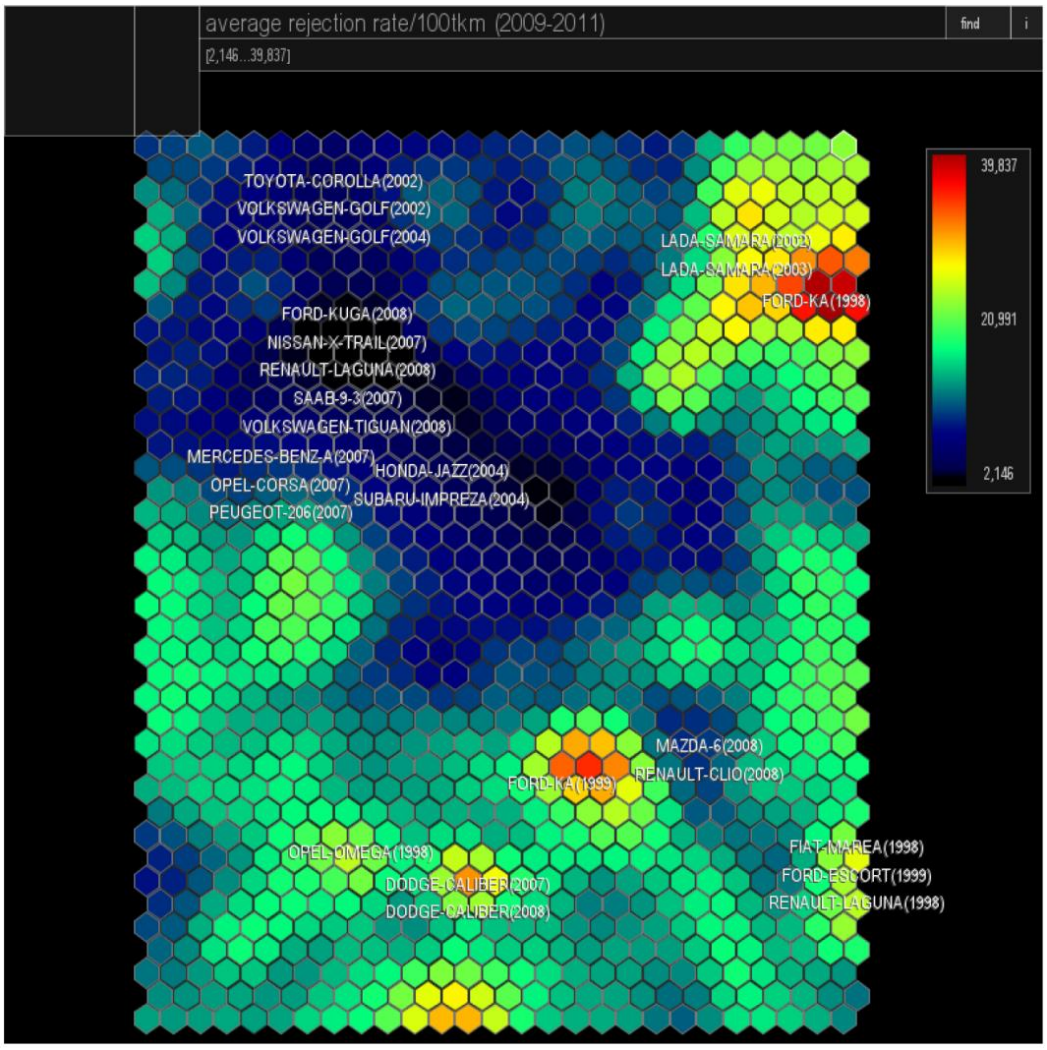

Figure 8. The car labels in a SOM visualization.

Summary

This technical report summarizes our research group's non-nuclear studies carried out during the last few years. Our two previous technical reports summarized the studies related to nuclear power in [2] and [1]. The studies about nuclear power are a follow up from earlier studies summarized in [21]. The later nuclear studies mostly deal with control room related issues. In this report we show that similar decision making concepts are relevant also in completely different application areas.

For different studies are described in detail: two studies about political decision making, and two about technical decision making. The political studies include one presidential election and one parliamentary election. Pre-election gallups and polls such as voting advice applications are utilized in prediction. In technical decision making we concentrate on car inspection data and especially in car rejection reasons.

In all studies visualization is in an important role. To give to the decision maker a visual explanation about the analyzed data is in great significance. Principal Component Analysis (PCA) and Self-Organizing Map (SOM) are the two important methodologies used. These methodologies are cleverly combined with other methods. The decision maker has got tools to base the decision on facts rather than just intuition. Both experts and ordinary folks can utilize these tools.

One our basic idea has been to combine data analysis and decision making. This has turned out to be very promising approach. The industrial processes as well as other applications struggle with somewhat similar problems. Our approach has many advantages in the information process. Visually supported decision making is a basis for successful operation and outcome.

References

[1] Sirola, Miki. Neural methods in process monitoring, visualization and early fault detection. Aalto University Science + Technology Technical Report 7/2014. Espoo, 2014. 20 p.

[2] Sirola, Miki; Talonen, Jaakko; Parviainen, Jukka; Lampi, Golan. Decision support with data-analysis methods in a nuclear power plant. TKK Reports in Information and Computer Science (TKK-ICS-R29). Espoo, 2010. 23 p.

[3] Kohonen, Teuvo. The self-organizing map. Springer, 1995.

[4] Talonen, Jaakko. Advances in Methods of Anomaly Detection and Visualization of Multivariate Data. Doctor's thesis, Aalto university in Finland, 2015.

[5] Talonen, Jaakko; Sirola, Miki; Sulkava, Mika. Data fusion of pre-election gallups and polls for improved support estimates. IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2015). Warsaw, Poland, 2015.

[6] Hirzalla, F.; Van Zoonen, L.; De Ridder, J. Internet use and political participation: reflections on the mobilization/normalization controversy. The Information Society, 27(1):1 - 15, 2010.

[7] Scott, A.J.; Smith, T.M.F. Analysis of repeated surveys using time series methods. Journal of the American Statistical Association, 69(347):674 - 678, 1974.

[8] Talonen, Jaakko; Sulkava, Mika. Analyzing parliamentary elections based on voting advice application data. International Symposium on Intelligent Data Analysis (IDA 2011). Porto, Portugal, 2011.

[9] Hair, J.; Anderson, R.; Tatham, R.; Black, W. Multivariate data analysis. Prentice Hall, 5th edition, 1998.

[10] Scott, D.W. Multivariate density estimation. 5th edition, Wiley Online Library, 1992.

[11] Haykin, S. Neural networks: a comprehensive foundation. Prentice Hall PTR, 1994.

[12] Manow, P.; Döring, H. Electoral and mechanical causes of divided government in the European Union. Comparative Political Studies 41(10), 1349, 2008.

[13] Talonen, Jaakko; Sirola, Miki; Sulkava, Mika. Network visualization of car inspection data using graph layout. International Conference on Data Analytics. Barcelona, Spain, 2012.

[14] Bastian, M.; Heymann, S.; Jacomy, M. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. 2009.

[15] Ellison, N.B. et.al. Social network sites: definition, history, and scholarship. Journal of Computer-Mediated Communiaction. Wiley Online Library, Vol. 13, pp. 210 – 230, 2007.

[16] Sirola, Miki. Perspectives in science - interdisciplinary examples in data-analysis visualization and political decision making. Scientific papers of Silesian University of Technology. Organization and Management Series, Nr. 82(1940), pp. 213 – 226, 2015.

[17] Talonen, Jaakko; Sulkava, Mika; Sirola, Miki. The Finnish car rejection reasons shown in an interactive SOM visualization tool. Workshop on Self-Organizing Maps (WSOM 2012). Santiago, Chile, 2012.

[18] Vesanto, Juha; Himberg Johan; Alhoniemi Esa; Parhankangas Juha. Self-organizing map in Matlab: the SOM Toolbox. Proceedings of Matlab-DSP conference, 1999.

[19] Stefanovic, P.; Kurasova, O. Visual analysis of self-organizing maps. Nonlinear Analysis: Modelling and Control, Vol. 16, No. 4, pp. 488 – 504, 2011.

[20] Kohonen, Teuvo; Honkela, Timo. Kohonen network. Scholarpedia, 2(1):1568, revision #122029, 2007.

[21] Sirola, Miki. Computerized decision support systems in failure and maintenance management of safety critical processes. VTT Publications 397. Espoo, VTT, 1999. 123 p. + app. 24 p.

Neural methods are applied in political and technical decision making. We introduce decision support schemes based on Self-Organizing Map (SOM) and Principal Component Analysis (PCA) combined with other methods. Visualizations based on various data-analysis methods are developed. In political decision making we have examples from one parliamentary election and one presidential election utilizing opinion data collected beforehand. In technical decision making we concentrate on rejection reasons in car inspection data.

Aalto-ST 5/2016

ISBN 978-952-60-6757-5 (printed)
ISBN 978-952-60-6758-2 (pdf)
ISSN-L 1799-4896
ISSN 1799-4896 (printed)
ISSN 1799-490X (pdf)

Aalto University
School of Science
Computer Science
www.aalto.fi