Arcada Working Papers 6/2015
ISSN 2342-3064
ISBN 978-952-5260-65-6

Working Papers Presented in Arcada Workshop on Analytics in May 25, 2015

Göran Pulkkis[i] (Ed.)

Abstract

The Department of Business Management and Analytics in Arcada University of Applied Sciences arranged a Workshop on Analytics in May 25, 2015. Four Working Papers presented in this workshop are published in this report.

CONTENTS

[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [goran.pulkkis@arcada.fi]

Reducing False Positives in Web Content Classification

Thomas Forss[i], Ulf Hedlund-Salmenkari[ii], Shuhua Liu[iii], Kaj-Mikael Björk[iv]

Abstract

In this paper we research methods that can have help reduce false positives when classifying web content. Some practical applications in automatic web content classification are very sensitive towards false positives; this means that even one false positive classification can reduce user trust in the system. Such systems are for example different types of access filters: parental control filters, virus site filters, and malicious web site filters.

Keywords: false positive, accuracy, classification, web content, dictionary, cleansing, outgoing links, tf-idf, sentiment analysis

1 INTRODUCTION

Work has been done in reducing false positives in email spam filtering (Hershkop & Stolfo 2005) and intrusion detection (Spathoulas & Katsikas 2010). Much less research has been done on data that combines multiple types of content. Web pages contain image content, text content and other structural content such as html elements and CSS elements. Our research focuses on web content classification.

[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [forsstho@arcada.fi]
[ii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [hedlundu@arcada.fi]
[iii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [shuhua@arcada.fi]
[iv] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [Kaj-Mikael Björk <bjorkpau@arcada.fi>]

1.1 Background

Different types of classifications focus on different kinds of results. The most general thing looked at while doing classifications is accuracy. Accuracy is a general measure of how good overall a classification has performed. However, some practical applications are not necessarily good enough simply by having a good accuracy. In these applications it is more important to either reduce the amount of false positives or reduce the amount of false negative classification results than to have a high accuracy. One application where low amount of false positives is imperative is the classification of violent web pages, where already classifying one non-violent web page as violent could potentially compromise trust in the system. In most health related classification research they try to minimize the false negative results as it is worse to send a patient home that has an illness than it is to do a few extra tests on a healthy patient.

2 CLASSIFICATION

Classification is one of the first tasks that humans learn to do when we grow up. We learn to classify animals, colors, family members and while the task of classifying is easy for a human it is quite time consuming and can vary a lot between different people, different moods, and different time of day when the classification task becomes specific. On the other hand using a computer to do classification is a very complex process that requires machine learning algorithms and/or complex mathematics but will once completed always return the same results for the same information.

In classification there are a number of different approaches, these are: supervised learning, unsupervised learning, and algorithms that use a combination of supervised and unsupervised learning. We use mainly supervised classification in our experiments, although we do use unsupervised algorithms for extracting sentiment features.

2.1 Supervised Classification

In supervised classifications we have a training set with pre-labeled data that is used by the chosen machine learning algorithm to train on. The pre-labeled set normally has data that is labeled as either positive results or negative results so that the algorithm knows which data is linked to the category and which is not. This helps the algorithm to determine in which category unlabeled data should be classified. Classification results are divided into four different types: true positives, true negatives, false positives and false negatives. (Powers 2011)

2.1.1 Classification Results

A true positive is defined as a positively labeled data point being correctly labeled by the algorithm as positive. A true negative is defined as a negatively labeled data point being classified correctly as negative. A false positive is defined as a negatively labeled data point being incorrectly labeled as positive. A false negative is defined as a positively labeled data point being incorrectly labeled as negative. (Powers 2011)

Precision is the measure for how exact a classifier is. Precision is measured by taking the amount of true positives divided by the number of true positives plus false positives added together. A low precision indicates that the classification has many false positive results. Recall is the measure of sensitivity in a classifier. (Powers 2011)

2.2 Web Content Classification

Web content classification is the task of categorizing web pages into predefined categories. There are a number of different applications that uses web content for classifications. For example parental control systems want to limit the web pages minors are allowed to visit. Another example would be limiting access to web pages containing malicious software.

Web content classification is unique in the way that it is one of the only types of classifications that are done on a combination of several content types. Web content can consists of structured text, unstructured text, images, videos, audio, html elements, Cascading Style Sheets, and elements from web programming languages.

3 REDUCING FALSE POSITIVES

To be able to reduce the amount of false positives in any given set of web pages there are a number of existing methods that can be used. The two most common techniques are testing different machine learning algorithms and using threshold selection on each of the learning algorithms. We also introduce new ways of reducing the false positive count in web content classification: text-only labeling vs. mixed content labeling, manual cleansing of dictionaries, extraction and usage of sentiment features, extraction of topic modeling features, and extraction features based on outgoing links from web pages.

3.1 Labeling Data Sets

For this study we start from a training set of 1919 pages categorized as having either violent text content or containing violent images and an equal amount of pages labeled as not containing violence. Violent text content includes discussions about violent acts, violent music lyrics and violent entertainment discussions. Violent images are defined as pictures real or made-up containing gore, blood, and violent acts. We then chose to relabel the web pages to see if manipulating the training data could reduce our false positive results.

3.1.1 Text-only Labeling vs. Blended Content Labeling

To be able to complete the relabeling we needed to develop a tool that would help us separate the web content into text-only content and a blended content. The idea is that since we are using classifiers trained only on text content we wanted to see how the performance of the classifiers faired if we labeled the datasets in two different ways. In the first data set pages are labeled as violent if it contains anything that is considered violent where as in the second data set pages are only labeled as violent if the text content contains violence.

3.2 Building and Cleansing Dictionaries

Our base system uses similarity analysis of TF-IDF values extracted from web pages compared to TF-IDF values extracted from the entire set of web pages labeled as violent. Theoretically TF-IDF should perform well, but in practice many of the top ranked category words are found in both violent and non-violent pages. This means that by going through the automatically generated TF-IDF dictionary and removing non-violent words we can get a better precision.

3.3 Performance of Different Learning Algorithms

Different learning algorithms perform different depending on the data (Xhemali, Hinde, & Stone 2009). That is why it is useful to try several of the different learning algorithms available. In future work we will do a comparison of Gaussian Naïve Bayes, Neural Networks (Lee, Hui, & Fong 2002), and Support Vector Machine performance (Sun, Lim, & Ng 2002).

3.4 Threshold Selection

As all learning algorithms work based on probabilities threshold selection is an effective way of reducing either false negative or false positives. By increasing the threshold closer to 1 we can tell the algorithms to only include those data points that have a higher probability of being correct. Reducing the threshold tells the algorithm to include data points with lower probability and thus reduces false negatives. A too high threshold lowers macro-average F-measures while a too low threshold lowers both macro- and micro-average F-measures (Fan & Lin 2007).

3.5 Extracting Sentiment Features

Previous work has shown that extracting sentiment features for any given text data can increase the classification accuracy (Liu, Forss, & Björk 2014). We chose to use unsupervised learning based on the SentiStrength program (Thelwall et al. 2010) to extract sentiment features for our two data sets. We extract 14 sentiment features for each web page and combine the features with the previous results.

3.6 Extracting Topic Modeling Features

There are both unsupervised and supervised (McAuliffe & Blei 2007) methods for topic modeling. The intuitive way to use topic modeling in our experiments is the supervised models for feature extraction since our data sets are labeled as positive and negative. We will be using supervised latent Dirichlet allocation developed by Mcauliffe & Blei (2007) to extract features from the two data sets.

3.7 Extracting Features from Outgoing Links

By calculating a set of domains for each manually labeled category we will try and gain meaningful features based solely on outgoing links in a web page. The idea is that the same kind of web pages probably link to the same kind of domains and that there is some meaning data to gain from counting number of matches in a web page against the outgoing links in a whole category.

4 CONCLUSION

We have so far done manual labeling of data sets for both text-only content and blended content. We have also extracted features for similarity and sentiment and are working on extracting features from topic modeling and outgoing links. No conclusion can yet be drawn as much work still needs to be done.

REFERENCES

Fan, R. E. & Lin, C. J. 2007, A study on threshold selection for multi-label classification, Department of Computer Science, National Taiwan University, pp. 1-23.

Hershkop, S. & Stolfo, S. J. 2005, Combining Email Models for False Positive Reduction, in: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge discovery in Data Mining*, ACM Press, pp. 98-107.

Lee, P. Y., Hui, S. C., & Fong, A. C. M. 2002, Neural Networks for Web Content Filtering. in: *Intelligent Systems,* Vol. 17. No. 5, IEEE, pp. 48-57.

Liu, S., Forss, T., & Björk, K.-M. June 2014, Web Content Classification with Topic and Sentiment Analysis, in: *Terminology and Knowledge Engineering 2014*, Berlin, Germany, 9 p.

McAuliffe, J. D. & Blei, D. M. 2007, Supervised Topic Models, in: *Advances in Neural Information Processing Systems, Vol 21,* pp. 121-128.

Powers, D. M. W. 2011, Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation, in: *Journal of Machine Learning Technologies*, Vol. 2, No. 1, pp. 37-63.

Spathoulas, G. P. & Katsikas, S. K. 2010, Reducing False Positives in Intrusion Detection Systems, in: *Computers & Security*, Vol. *29*, No. 1, pp. 35-44.

Sun, A., Lim, E. P., & Ng, W. K. 2002, Web Classification Using Support Vector Machine, in: *Proceedings of the 4th International Workshop on Web Information and Data Management*, ACM Press, pp. 96-99.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. 2010, Sentiment Strength Detection in Short Informal Text, in: *Journal of the American Society for Information Science and Technology*, Vol. 61, No.12, pp. 2544-2558.

Xhemali, D., Hinde, C. J., & Stone, R. G. 2009, Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages, in: *International Journal of Computer Science Issues*, Vol. 4, No. 1, pp. 16-23

Possibilistic clustering for crisis prediction: Systemic risk states and membership degrees

József Mezei[i,iii], Peter Sarlin[ii,iii]

Abstract

Research on understanding and predicting systemic financial risk has been of increasing importance in the recent years. A common approach is to build predictive models based on macro-financial vulnerability indicators to identify systemic risk at an early stage. In this article, we outline an approach for identifying different systemic risk states through possibilistic fuzzy clustering. Instead of directly using a supervised classification method, we aim at identifying coherent groups of vulnerability with macro-financial indicators for pre-crisis data, and determine the level of risk for a new observation based on its similarity to the identified groups. The approach allows for differentiating among different possible pre-crisis states, and using this information for estimating the possibility of systemic risk. In this work, we focus on the first stage of this research stream by comparing different fuzzy clustering methods and offering some theoretical properties and empirical observations to aid in the choice of a suitable clustering method for systemic risk identification.

Keywords: clustering, membership function, systemic risk

1 INTRODUCTION

Clustering is one of the most important tasks in machine learning, and aims at partitioning a set of data points into groups of ‘similar’ observations. Fuzzy clustering methods rely on set-theoretical notions introduced by Zadeh (1965), motivated by the imprecision present in many (if not all) real-life phenomena. The main idea behind fuzzy sets (i.e., degree of belonging to sets) naturally translates to clustering algorithms: elements can belong to several overlapping fuzzy clusters specified by a membership value. In fuzzy clustering, the fuzzy c-means (FCM) clustering algorithm (Bezdek 1981) is the best known and used method. Since the FCM memberships do not always

[i] Åbo Akademi University, Turku, Finland, IAMSR, [jmezei@abo.fi]
[ii] Hanken School of Economics, Helsinki, Finland, Department of Economics
[iii] Arcada University of Applied Sciences, Helsinki, Finland, Risklab, [peter@risklab.fi]

explain the degrees of belonging for the data well, Krishnapuram & Keller (1993) proposed a possibilistic approach to clustering to correct this weakness of FCM. However, the performance of Krishnapuram & Keller's (1993) approach depends heavily on the applied parameters. This has been pointed out several times in the literature, e.g. (Doring et al. 2006), and resulted in different modification of the original possibilistic fuzzy clustering algorithm.

In this work, we focus on objective function-based fuzzy clustering algorithms, which constitute the most widely applied variants (yet only a subset) of overall fuzzy clustering methods (Höppner et al. 1999). Following Marghescu et al. (2010), we propose to apply fuzzy clustering approaches to country-level indicators of financial crises. The rationale behind possibilistic clustering is that when applied to new data, memberships need not sum up to 1 but may be rather low in case an observation does not resemble any of the systemic risk states. In this preliminary study, we focus on outlining our approach and comparing different fuzzy clustering algorithms to identify potential candidate methods to be used in further research. The subset of fuzzy clustering methods considered in the numerical study can be identified as special cases of a general optimization problem with an objective function specifying different attributes of the resulting cluster structure. In a numerical study, the chosen methods as the basis of predicting whether a country at a given time is in pre-crisis state or not based on the membership degree of belonging to different clusters. We find the initial values for the membership values and the number of clusters specified has the most significant effect on the accuracy of the prediction.

The rest of the paper is structured as follows. In Section 2, a short literature review is provided on possibilistic fuzzy clustering algorithms, focusing on objective-function based algorithms. An outline for a general approach on using possibilistic fuzzy clustering for assessing systemic risk is described in Section 3, together with some preliminary observations on the choice of clustering method. Finally, Section 4 provides conclusions and future research directions.

2 FUZZY CLUSTERING

In clustering, the general input data consists of *n* observations with every observation described by *m* measurement variables: $x_i = [x_{i,1}, x_{i,1}, ..., x_{i,m}], i = 1,...n.$ We can use the following matrix representation:

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix}$$

As the basis of clustering, the type of partition used to define the clusters can be characterized following two main approaches:

- crisp sets: every object belongs to exactly one cluster;
- fuzzy sets: every object can belong to several clusters with different degrees.

If we denote the number of clusters with *c*, a crisp partition can be described by the matrix $U = [\mu_{i,j}]$where

$$\mu_{i,j} \in \{0,1\}, 1 \le i \le c, 1 \le j \le n,$$

$$\sum_{i=1}^{c} \mu_{i,j} = 1, 1 \le j \le n$$

$$0 < \sum_{j=1}^{n} \mu_{i,j} < n, 1 \le i \le c$$

A classical, usually termed as probabilistic, fuzzy partition can be described with the same constraints with extending the possible values for the cluster memberships, $\mu_{i,j}$, to the [0,1] interval. Furthermore, a possibilistic fuzzy partition releases the normalization constraint on the membership values and it is replaced by $\exists \mu_{i,j} > 0, 1 \le j \le n$ which ensures that every observation belongs to at least one cluster to some degree.

There are a large number of different variations of fuzzy clustering methods that can be classified based on different properties. The most commonly used distinction divides the algorithms into two main groups: (i) methods that try to find a fuzzy partition using global criteria for optimality in form of an objective function and (ii) methods generalizing the previous approaches by allowing the user to choose among multiple update equations for the prototypes and membership degrees without considering a particular criterion function. Although there are several approaches belonging to the second group (Runkler & Bezdek 1999), objective function-based approaches dominate the literature. According to these approaches, one group of parameters (e.g., the membership degrees) is optimized holding the other group (e.g., the prototypes) fixed and vice versa following an iterative updating scheme. A general objective function can be formulated as

$$J(X,U,B) = \sum_{i=1}^{c}\sum_{j=1}^{n}\left(a\mu_{i,j}^{m} + bt_{i,j}^{\lambda}\right)d^{2}\left(x_{i}, c_{j}\right)$$

$$+\sum_{i=1}^{c}\eta_{i}\sum_{j=1}^{n}\left(1-t_{i,j}\right)^{\lambda} + \sum_{i=1}^{c}\gamma_{i}\sum_{j=1, j\neq i}^{n}\frac{1}{\varsigma d^{2}(c_{i}, c_{j})}$$

where $\mu_{i,j}$ is the membership degree that is normalized for every observation, $t_{i,j}$ is the typicality degree that corresponds to non-normalized membership degrees, B is the matrix of cluster prototypes, while the other parameters can specify the distance measure in the model and the shape of the resulting clusters.

The fuzzy clustering algorithms used in the next section are all based on different special cases of this general objective function:

- the fuzzy c-means (Bezdek 1981): a =1, b=0, $\eta_i = 0, \gamma_i = 0$;
- the possibilistic fuzzy clustering (Krishnapuram & Keller 1993): a =0, b=1, $\gamma_i = 0$;
- the mixed c-means clustering model (Pal et al. 1997): a=1, b=1, $\eta_i = 0, \gamma_i = 0$;
- the possibilistic c-means clustering model (Pal et al. 2005): $\gamma_i = 0$;
- the extended possibilistic clustering model (Timm et al. 2004): a=0.

In applying clustering approaches, a crucial issue is to determine the appropriate number of clusters. For this purpose, there are several cluster validity measures proposed to assess cluster configurations, with many of them specific to fuzzy clustering as traditional measures evaluating hard clustering are not applicable. As was pointed out by Wang & Zhang (2007), and applies generally to clustering, there is no one single validity measure for fuzzy clustering that assures good performance in every situation. Hence, it is advisable to consider several measures as the basis of evaluating the optimal number and structure of clusters. The numerical study makes use of two measures: the partition coefficient (Bezdek 1981) and the separation index (Hoppner et al. 1999).

3 CRISIS PREDICTION WITH POSSIBILISTIC CLUSTERING

3.1 Data

The dataset used in this paper covers as many European economies as possible on a quarterly frequency and spans from 1976Q1 to 2014Q3. The sample is an unbalanced panel with 15 European Union countries: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, Sweden, and the United Kingdom. The sample includes 15 events of systemic banking crises. The dataset consists of two parts: crisis events and vulnerability indicators. The crisis events cover country-level distress in the financial sector with systemic implications and rely on the IMF's crisis event initiative by Laeven & Valencia (2013). The second part of the dataset consists of country-level macro-financial vulnerability indicators. We include measures covering asset prices (e.g., house and stock prices), leverage (e.g., mortgages, private loans and household loans), business cycle indicators (GDP and inflation), measures from the EU Macroeconomic Imbalance Procedure (e.g., current account deficits and government debt), and the banking sector (e.g., loans to deposits). In most cases, we have relied on the most commonly used transformations, such as ratios to GDP or income, growth rates, and absolute and relative deviations from a trend.

3.2 Experiment design

As the objective of the model in this paper is to support systemic risk analysis, we have a set-up that strictly follows the use of data in a real-time manner. The ultimate objective is two-fold: to identify systemic risk states (clusters) and their possibilistic likelihood (membership degree). To reach this, we make use of the following procedure:

- Identify optimal number of clusters and apply possibilistic clustering to pre-crisis data (i.e., use solely observations representing systemic risk states).
- Compute membership degrees for all data to the clusters, in which data distant to all cluster centroids exhibits low membership values to all clusters.
- Identify optimal thresholds on the possibilistic likelihood with the Usefulness measure (i.e., a preference weighted average of type I and II errors, see (Sarlin 2013) for further information).

We apply the above in multiple ways. The first differentiation is in-sample vs out-of-sample analysis. In this work, we target out-of-sample analysis in two ways: (i) one split into train and test data at a specific year and (ii) as a recursive real-time analysis by using all available information at each quarter for deriving models.

3.3 Preliminary observations

In this work, our focus is on comparing the performance of the five described fuzzy clustering methods; according to this, we only present some observations based on the performed analysis without discussing the results regarding the underlying systemic risk problem. To compare the fuzzy clustering methods, we have collected the necessary data and implemented the setup (including algorithms and validity measures). The first and foremost observation regarding the use of fuzzy clustering approaches relates to the initialization of membership functions prior to running the algorithms. As was pointed out, possibilistic clustering algorithms are highly sensitive to initializations, while this is not a crucial issue for the probabilistic FCM algorithm. In our case, the possibilistic clustering algorithms (except the one by Timm et al. (2004)) converged into identical clusters independently from the optimal number of clusters. This problem can be tackled by initializing the possibilistic variants with the resulted membership values from a FCM algorithm.

With proper initialization, the predictive performance of the four possibilistic variants is very similar to each other. The more complex models result in better performance, but even in the case of optimal parameters, they perform qualitatively equally well than the traditional FCM algorithm. As we discussed previously, the motivation behind using possibilistic models lies in the interpretation and use of the membership functions, which is not necessarily measurable directly based upon this simple numerical comparison. Additionally, we found that while the traditional FCM is very sensitive to the initialization of the exponent value *m* with regard to the accuracy of the prediction, in case of the possibilistic approaches, the choice of the optimal cluster number can affect the results more significantly than the choice of the parameters. For this purpose, we plan to implement additional validity measures to improve the performance of the models.

4 CONCLUSIONS

In the literature, there are numerous proposals to predicting systemic risk. In this short paper, we outlined a possible approach using different vulnerability measures as a basis of clustering. We propose to use possibilistic fuzzy clustering as it has obvious benefits compared to traditional clustering approaches that better allow for describing the (dis)similarity to systemic risk states. We looked at different objective function-based approaches from the literature and formulated them using a common general objective function. We outlined the approach for systemic risk analysis, and offer some preliminary observations based on a numerical comparison of the different clustering models. This study extends previous work by Marghescu et al. (2010) along several directions. First, we tackle European systemic banking crises in the past decades, in contrast to their focus on excessive exchange-rate pressure in the 1990s in Asia. Additionally, rather than identifying overall financial stability states, we aim at identifying natural clusters in pre-crisis data in order to provide more descriptive representations through different states and membership degrees. Beyond only a crisis probability provided by standard early-warning models, our approach provides more descriptive output that allows coupling the characterization of and membership to different systemic risk states with concrete policy actions.

REFERENCES

Bezdek, James 1981, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers.

Döring, C., Lesot, M.-J., & Kruse, R. 2006, Data Analysis with Fuzzy Clustering Methods, in: *Computational Statistics & Data Analysis*, Vol. 51, No. 1, pp. 192-214, ISSN: 0167-9473.

Höppner, F., Klawonn, F., Kruse, R., & Runkler, T. 1999, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*, John Wiley & Sons.

Krishnapuram, R. & Keller, J. M. 1993, A Possibilistic Approach to Clustering, in: *IEEE Transactions on Fuzzy Systems*, Vol. 1, No. 2, pp. 98-110, ISSN: 1063-6706.

Laeven, L. & Valencia, F. 2013, Banking Crisis Database: An Update, in*: IMF Economic Review*.

Marghescu, D., Sarlin, P., & Liu, S. 2010, Early-Warning Analysis for Currency Crises in Emerging Markets: A Revisit with Fuzzy Clustering, in*: Intelligent Systems in Accounting, Finance and Management*, Vol. 17, No. 3-4, pp. 143-165, ISSN: 1099-1174.

Pal, N. R., Pal, K., & Bezdek, J. C. 1997, A Mixed C-Means Clustering Model, in: *Proceedings of the Sixth IEEE International Conference on Fuzzy Systems*, Vol. 1, pp. 11 -21.

Pal, N.R. & Pal, K. & Keller, J. M. & Bezdek, J. 2005, A Possibilistic Fuzzy C-Means Clustering Algorithm, in: *IEEE Transactions on Fuzzy Systems*, Vol 13, No 4, pp. 517-530. ISSN: 1063-6706.

Runkler, T. A. & Bezdek, J. C. 1999, Alternating Cluster Estimation: A New Tool for Clustering and Function Approximation, in: *IEEE Transactions on Fuzzy Systems*, Vol 7, No 4, pp. 377-393. ISSN: 1063-6706.

Sarlin, P. 2013, On Policymakers' Loss Functions and the Evaluation of Early Warning Systems, in: *Economics Letters*, Vol 119, No 1, pp. 1-7. ISSN: 0165-1765.

Timm, H. & Borgelt, C. & Döring, C. & Kruse, R. 2004, An Extension to Possibilistic Fuzzy Cluster Analysis, in: *Fuzzy Sets and Systems*, Vol 147, No 1, pp. 3-16. ISSN: 0165-0114.

Wang, W. & Zhang, Y. 2007, On Fuzzy Cluster Validity Indices, in: *Fuzzy Sets and Systems*, Vol 158, No 19, pp. 2095-2117. ISSN: 0165-0114.

Zadeh, L. A. 1965, Fuzzy Sets, in: *Information and Control*, Vol 8, No 3, pp. 338-353. ISSN: 0890-5401.

Data Driven Decision-Making in eRetailing – A Customer Centric Funnel Approach

Niklas Eriksson[i], Mikael Forsström[ii], Carl-Johan Rosenbröijer[iii]

Abstract

The aim of this study is to use the customer centric funnel approach to assess if Finnish small eRetailers make data driven decisions to improve their performance. The research is based on an edited version of Chaffey and Smith´s (2013) PRACE framework, where we focus on the concepts of reach, act, convert and engage (RACE). We then combine this with the business analytics framework by Delen and Demirkan (2013a) divided in descriptive, predictive and prescriptive analytics. By combining these two frameworks we are able to create a research model to assess if Finnish small eRetailers make data driven decisions to improve their performance. The empirical research will be a survey sent to Finnish eRetailers. The expected results are an assessment of the readiness to make data driven decisions. The questioner will make it possible to see if the eRetailers collect data, analyze the data and finally if they use this data for decision-making and corrective actions.

Keywords: business analytics, big data, data driven decision making, eRetailing

1 INTRODUCTION

Big Data and analytics is seen as a very critical area in business when it comes to future decision-making (see e.g. (Brynjolfsson & McAfee 2012) and (Davenport 2014)). The trend seems to be that data driven decision-making is not only more trustworthy than intuition and experience based decision-making but also resulting in better performance

[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [eriksson@arcada.fi]
[ii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [forsstrm@arcada.fi]
[iii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [rosenbrc@arcada.fi]

for companies. The digitalization of business and especially the Internet infrastructure has led to enormous creation of data that can be accessed openly, in real-time and collected for advanced analytics.

The eRetailing context is a totally digital context, which makes it interesting from a big data and analytics perspective. Hence, there is a continuous flow of data created as a result of consumers activities in e.g. search engines, social media, blogs, homepages and e-shops. All this data can be part of a marketing funnel that preferably ends with a purchase and customer advocacy.

In 2014 the consumers in Finland bought products and services online for a value of €10,5 billion. The proportions of services were 54 % (decrease 5 %), products 45 % (increase 7%) and digital content 1 % (increase 12 %). Over 80 % of the consumers bought products and services from an online site with Finnish language whereas 68 % bought digital content from a Finnish language site. These statistics have to be related to the ongoing economic downturn in Finland. (TNS Gallup 2015)

2 AIM

The aim of this study is to use the customer centric funnel approach to assess if Finnish small eRetailers make data driven decisions to improve their performance.

As there is constant access to data in a digital context it becomes critical to see if this data is collected, used and analyzed for decision-making. Referring to the work by Brynjolfsson, Hitt, & Kim (2011) we assume that data driven decision-making increases company performance. Hence, we therefore see a need to explore if the small Finnish eRetailers make data driven decisions or not.

As a result of the study the goal is to assess if small eRetailers in Finland use data and analytics to make data driven decisions in three different buyer stages of the customer funnel.

3 LITERATURE REVIEW AND MODEL

3.1 Data Driven decision-making

Delen & Demirkan (2013a) argue that the managers’ main job, decision-making, becomes more complex, and repeatedly making the right decisions in a timely manner becomes a matter of survival. Data is the key building block for decision-making. The quality and accuracy of data becomes crucial for successful decision-making. With the Internet infrastructure and digital revolution we have seen an explosion of sites, e.g. homepages, Ecommerce platforms, search engines and social media platforms (for more see (Chaffey & Smith 2013)). The global and local traffic on these sites create a continuing flow of what we call big data on a 24/7 basis. The institutionalized definition of big data is high volume, high velocity and high variety data (META Group/Gartner).

Another feature of big data is that it is real-time. This makes it possible to collect data and analyze it continuously. Davenport (2014) presents a development of terminology concerning using and analyzing data between 1989 – present time. Business intelligence (1989-2005) includes tools to support data driven decision-making with emphasis on reporting. Analytics (2005-2010) is based on statistical and mathematical analysis for decisions. Big data (2010-present) focuses on very large, unstructured and fast moving data. All of these eras have in common a data driven decision-making philosophy. Apart from the area of management decision support big data is very much related to the upcoming area of data-enriched service innovation (see e.g. (Davenport 2013) and (Eriksson, Westerlund, & Rosenbröijer 2014)).

General categorizations of elements needed for data driven decision-making are data, information and knowledge. Data is for example numbers, text, and/or graphics (e.g. pictures and videos) without any context. To become information data has to be placed in a certain context or system. For information to become knowledge it has to be analyzed. The knowledge level is often related to analytics, i.e. statistical and mathematical analysis of data. Delen & Demirkan (2013b) present a development of service oriented decision support systems where data-as-a-service and information-as-a-service are already established whereas analytics-as-a-service is a relatively newer concept in the business world. The analytics-as-a-service concept is based on the increased popularity of business analytics as a managerial paradigm. The authors divide business analytics into descriptive, predictive and prescriptive analytics. Descriptive analytics focuses on answering questions like what happened and what is happening, enabling data driven business reporting and dashboards. Whereas predictive analytics is future oriented, i.e. questions like what will happen and why will it happen are critical, enabling e.g. data mining, web-mining and forecasting. Finally the prescriptive analytics approach is normative in nature and aim at answering what should I do and why should I do it, enabling e.g. simulation and optimization. All three perspectives aim at outcomes that enable the decision maker to make data based business performance increasing decisions.

In an online retail setting the analytics-as-a-service concept has been referred to as web analytics. Web analytics is a technique used to assess and improve the contribution of e-marketing to a business including reviewing traffic volume, referrals, clickstreams, online reach data, customer satisfaction surveys, leads and sales (Chaffey & Smith 2013). Google Analytics is a very popular cloud based web analytics service. The service tracks all the traffic and contributes to the online consumer behavior insight concerning building an audience for your offering and the conversion to sales. It also offers descriptive analytics in the form of dashboards. The question then remains. Is this analytics (data) used for decision-making to improve the performance of for example conversion and sales?

Brynjolfsson, Hitt, & Kim (2011) argue that more companies' managerial decisions rely less on leader's "gutt instinct" and instead on data based analytics. Retailers would clearly benefit from decision-making based on customer analytics due to the fact that they make many of the same decisions repeatedly. Germann et. al. (2014) also see clear advantages with customer analytics based decision-making in retail because retailers have access to a large volume of customer data, powerful customer analytics methods tailored to retailers are available and customer analytics based methods exist for many retailing decisions that are made on a regular basis. As a result of the increased amount of data available a trend has emerged where leading edge firms have moved from passively collecting data to actively conducting customer experiments to develop and test new products and services (Brynjolfsson, Hitt, & Kim 2011). This culture of experimentation has diffused to e.g. pure online retailers, like Amazon, eBay and Google. These firms rely heavily on field experiments, utilizing high visibility and high volume of online customer interaction to validate and improve new product and pricing strategies (Brynjolfsson, Hitt, & Kim 2011).

3.2 A Customer Centric Funnel Approach

The starting point for the funnel approach is the customer behavior in an eRetail setting. This behavior creates loads of data. The data can and should then be used to guide the eRetailer to make the right data based decisions. The customer centric funnel approach used in this study is edited from Chaffey & Smith´s (2013) PRACE framework. In this piece of research we will use the model as a RACE model (R=Reach, A&C=Act and Convert, E=Engage), hence the first P=plan will not be in focus because it takes a more strategic approach where planning becomes important. Instead our approach is based on a more operational level. The funnel structure is a customer focused behavioral process that we divide in three buyer stages (originally four). Each buyer stage has a management goal, i.e. related to the RACE framework.

1) Exploration – Reach
   In the start of the funnel the consumer is in an exploration stage where he/she is exploring alternatives or only out of interest search for product information, reviews (technical or user based), tests, discussions etc. This can be done through search engines, social networks, publishers, blogs etc. Here the management goal is Reach, i.e. how can we reach the potential customers and building awareness of our brand and its products or services. The main aim of reaching the right consumers with the right message needs to lead to traffic to our own home site, social media site, blog and/or E-commerce site. To be able to assure we achieve the Reach goal and create traffic to our main content sites we need data and analytics. We also need some sort of measures, e.g. number of fans, followers, visitors, inbound links etc.

2) Decision-making and Purchase – Act & Convert
   In this buyer stage the customer has already reached our home site, blog and/or social media community. The consumer is here in the decision-making process, i.e. should I buy or not, what should I buy, etc.. The management goal here is to get the consumer to act, hence to get the consumer to find out more of the company and its products or services. Data is here measured as e.g. time on site, shares, comments, likes, leads and conversion. With the management goal convert the main aim is to get the consumer to purchase a product and/or service. Then the management has reached its commercial objective. In this buyer stage data is crucial because it measures the commercial success of the eRetailer. The measures can then be orders, revenue, average order value, etc..

3) Advocacy – Engage
   In this last stage the consumer has purchased, hence become a customer. Here it is crucial to engage the customer to share their experiences of the process, product and/or service. The aim is to build a deeper customer relationship encouraging advocacy or recommendations through word-of-mouth/mouse. In this stage data is measured by for example repeat purchase, referral, etc.

The consulting agency Ivorio (2015) presents analytics based services in customer acquisition, online experience and sales follow up. Here the historical data of customer behavior is used to enhance searching and navigating capabilities on a personal level. Recommendations can also be launched during an online session so that the customer can be offered complementing products (cross-sell) or higher value products (up-sell). With good quality and detailed data the customer can be offered a personalized support experience.

The edited customer centric funnel approach will now be combined with the data driven decision-making approach to a model that we will operationalize in our survey questioner.

3.3 Research Model

According to our aim of the study we used the customer centric funnel approach to assess if Finnish small eRetailers make data driven decisions to improve their performance. We therefore needed to create a model that combines data driven decision-making (3.1) with the customer centric funnel approach (3.2) (see Figure 1). The data driven decision-making presented above was divided in three areas; data, analytics and decision-making. In our research model data is needed to answer the question of “What is happening?”. Here data collection and following performance measurement is key. The analytics part is answering the question of “Why is it happening?”. Through performance diagnosis the aim is to understand. Finally, the decision-making part of the process is in our model focusing on the question, “What do we do?”. Through corrective actions we can make data driven performance-increasing decisions.

The customer centric funnel approach is based on three buyer stages; exploration, decision-making & purchase and finally advocacy. From an eRetailers perspective the management goal is to Reach, Act & Convert and finally to Engage. These three goals are giving the main structure of our research model.

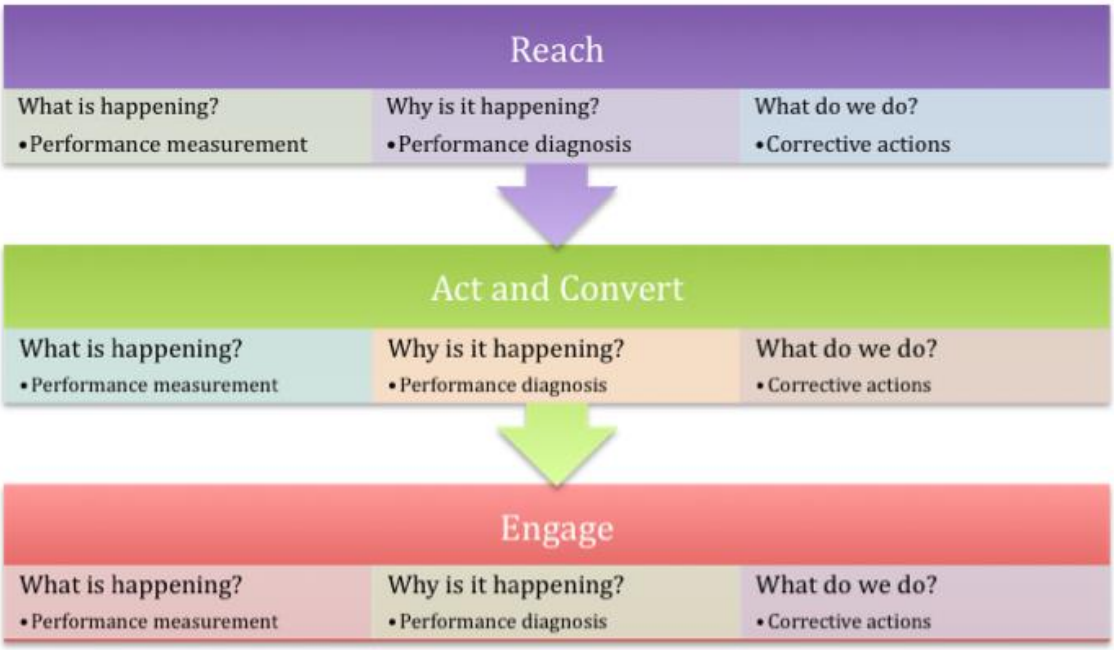

Figure 1. Research model

4 METHOD

Based on the aim of the study we conducted a survey among small eRetailers in Finland. The sample consists of small eRetailers that either operate only online or both online and offline, i.e. with one or several brick-and-mortar stores. The questioner was sent to 1300 small Finnish eRetailers. They are all customers of Vilkas Oy and use ePages, a cloud based eCommerce service platform.

The questioner was created by operationalizing the above research model. The questioner was divided in four parts:

A. eRetailer's profile data
B. Reach – eShop visitors
C. Act & Convert – Consumer behavior in the eShop
D. Engage – Customer loyalty

In part A we collect profile data concerning the eRetailer to be able to compare results between different types of eRetailers. The B-D parts are based on the research model. In each of the three parts we have nine questions, i.e. three questions for performance measurement, three for performance diagnosis and three for corrective actions. All in all there are 15 questions in part A and 27 questions in parts B-D. The responses are measured on a five point likert scale. The questioner was created with QuestionPro and sent to the respondents in May/June and August/September 2015.

5 EXPECTED RESULTS

The aim of this study is to use the customer centric funnel approach to assess if Finnish small eRetailers make data driven decisions to improve their performance. Based on this aim we intend to conduct a survey. The expected results are an assessment of the readiness to make data driven decisions. The questioner will make it possible to see if the eRetailers collect data, analyze the data and finally if they use this data for decision-making and corrective actions. The results will further be analyzed according to the edited three-phase customer centric funnel model developed by Chaffey & Smith (2013). Apart from this we will analyze the collected data in relation to the eRetailers' company profile data, thus giving us the possibility to make comparisons between different types of eRetailers in Finland. Finally we expect the study to create an impact in the small and medium sized eRetailer business context in Finland by indicating their readiness for data driven decision-making.

REFERENCES

Brynjolfsson, E., Hitt, M. L., & Kim, H. H. 2011, *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?*, Social Science Research Network.

Chaffey, D. and Smith, P. R. 2013, Emarketing Excellence – Planning and Optimizing you Digital Marketing, 4th edition, Routledge, New York

Davenport, T. H. 2013, Analytics 3.0, in: *Harvard Business Review*, December 2013, pp. 64-72.

Davenport, T. H. 2014, *Big Data @ Work – Dispelling the Myths, Uncovering the Opportunities*, Harvard Business School Publishing Corporation, USA

Delen, D. & Demirkan, H. 2013a, Data, Information and Analytics as Services, in: *Decision Support Systems*, Vol. 55, No. 1, pp. 359-363.

Delen, D. & Demirkan, H. 2013b, Leveraging the Capabilities of Service-Oriented Decision Support Systems: Putting Analytics and Big Data in Cloud, in: *Decision Support Systems*, Vol. 55, No. 1, pp. 412-421.

Eriksson, E., Westerlund, M., & Rosenbröijer, C.-J. 2014, Big Data Analytics – What is it, Arcada Working Papers No. 2, Arcada University of Applied Sciences

German, F., Lilien, G. L. , Fiedler, L., & Kraus, M. 2014, Do Retailers Benefit from Deploying Customer Analytics? in: *Journal of Retailing,* No. 4, pp. 587-593

Ivorio Oy 2015, Maturity Assessment Model of Online Retail and Analytics, Helsinki

McAfee, A. & Brynjolfsson, E. 2012, Big Data: The Management Revolution, in: *Harvard Business Review,* October 2012.

TNS Gallup, 2015. Accessed 11.8.2015. Published 2015. https://www.tns-gallup.fi/mita-teemme/digitaaliset-tutkimukset

Advanced Analytics

Kaj Grahn[i]

Abstract

Advanced analytics and business intelligence are about the development of technologies, systems, practices, and applications to analyze critical business data. Thus, new insights about business and markets are gained. The information can be used to improve products and services, achieve better operational efficiency, and create good customer relationships. In this paper, a brief introduction to advanced analytics is given. The differences between advanced analytics and business intelligence are discussed. The importance of big data in this context is also underlined.

Keywords: advanced analytics, business intelligence, big data, information management

1 INTRODUCTION

Traditional information management is currently in the middle of a rapid transformation as the amount of available structured and unstructured information is increasing. Therefore, utilizing unstructured data from social media in business intelligence (BI) and business analytics (BA) has become a strategic topic in many organizations (Tasala 2014). The massive amount of collected, stored and analyzed data is commonly referred to as “big data”. Collected big data creates little value; it must be analyzed in order to generate value and the key factor here is the use of analytics. According to (The Current 2012), about 97 percent of large enterprises are utilizing BA.

In the 1970s, the first decision support systems (DSS) appeared. DSS is a computer program application that supports business and organizational decision-making activities. Then in the 1990s the BI term was popularized. In (Watson 2009) BI is defined as “a broad category of applications, technologies, and processes for gathering, storing, accessing, and analyzing data to help business users make better decisions”. Thus BI can be viewed as an umbrella term for all applications that support decision making. BI evolved from DSS, and one could argue that analytics evolved from BI (at least in terms of terminology).Thus, analytics is an umbrella term for data analysis

[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [kaj.grahn@arcada.fi]

applications (Watson 2013). Today, more sophisticated analytics developed by different vendors are usually named advanced analytics.

In the following, a brief introduction to business intelligence, big data and advanced analytics is given. Differences between business intelligence and advanced analytics are underlined.

2 BUSINESS INTELLIGENCE

There has not been agreed upon any common definition of BI (Vitt et al. 2012). In the academic world and among vendors many definitions are in use. A typical definition is given by *"Business Intelligence (BI) is an umbrella term that includes the applications, products, infrastructure and tools and best practices that enable access to and analysis of information to improve and optimize decisions and performance"* (Business 2013).

Three main categories of definitions capture the essence of BI aspects – management, technology and product -, are identified in (Chee et al. 2009).The approach includes (Iankoulova 2012):

- **management** - process of gathering data from internal and external sources and of analyzing them in order to generate relevant information for improved decision making
- **technology** - tools and technologies that allow the recording, recovery, manipulation and analysis of information
- **product** - the emerging result/product of in -depth analysis of detailed business data as well as analysis practices using BI tools.

Typical BI tasks, goals, functions and applications are (Business 2015):

- to gather, store, analyze, and provide access to data to support enterprise users in making better informed business decisions
- to include decision support systems, query and reporting, online analytical processing (OLAP), data management, statistical analysis, forecasting, data analytics and data mining
- to transform raw data into meaningful and useful information for business analysis purposes
- to provide historical, current and predictive views of business operations
- to include reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

Historical, current and predictive views of business operations can be provided by BI. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics. (Business 2015)

Traditional BI forms a one-way street from the data warehouse to the business user and provides means for information consumption. User collaboration by providing information, comments and feedback is enabled by a business-technology ecosystem, i.e. a virtual unification of big data and traditional business information. BI is capable of handling large amounts of unstructured data to help identifying new opportunities, implementing strategies and providing businesses with a competitive market advantage. A wide range of business decisions ranging from operational to strategic is supported. Combining external and internal data gives a more complete picture, “intelligence” is provided.

3 BIG DATA

Big data can be characterized as having high volume, high velocity, and high variety. Many consider 10 terabytes to be big data. This numerical definition is likely to change over time. Organizations collect, store, and analyze more and more data. Ten terabytes can hold the printed collection of the Library of Congress and one petabyte about 500 million floppy disks. (Watson 2014)

Miller et al. (2012) define big data as:
Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes.

Big data analysis is a challenge. Obstacles are the sheer volume of data, different formats of data (structured/unstructured), collection across the entire organization and combining, contrasting and analyzing the data. Many problems are caused by the uncontrolled and unmanaged spreadsheet reporting culture. An integrated information platform enables handling unstructured data as easily as structured data.

The data sets are so large or complex that traditional data processing applications are inadequate. Typical challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics. Accuracy in big data may lead to more confident decision making and better decisions can mean greater operational efficiency, cost reductions and reduced risk. An important part in integrating the data from heterogeneous sources is an ecosystem which is formed by business and technology together.

4 ADVANCED ANALYTICS

The term *analytics* is not used consistently. In (Watson 2013), three related definitions of the term are given:

- analytics is an umbrella term for data analysis applications
- analyzing the data that is stored by BI
- use of “rocket science” algorithms (e.g., machine learning, neural networks) to analyze data.

Basic analytics provide a general summary of data. Deeper data knowledge and granular data analysis is delivered by advanced analytics. Furthermore, data-driven decision making is a powerful way to boost business outcomes. For example, clinical diagnosis applications plus advanced analytics recommend optimal treatment plans for individual patients (Gartner 2014)

In 2013, advanced analytics was the fastest growing segment of the BI and analytics software market (Gartner 2014). New technologies and new types of data affect the evolving of analytical processing solutions. The term, advanced analytics, is often associated with more sophisticated capabilities of the solutions. The term advanced analytics is very popular among vendors.

Advanced analytics is defined as (rapidminder 2015):
"*The analysis of all kinds of data using sophisticated quantitative methods (for example, statistics, descriptive and predictive data mining, simulation and optimization) to produce insights that traditional approaches to business intelligence (BI) – such as query and reporting – are unlikely to discover".*
If IT and business organizations use BI that addresses descriptive analysis (what happened) to advanced analytics, which complements by answering the "why," the "what will happen," and "how we can address it", the following procedure can be shown, see Figure 1.

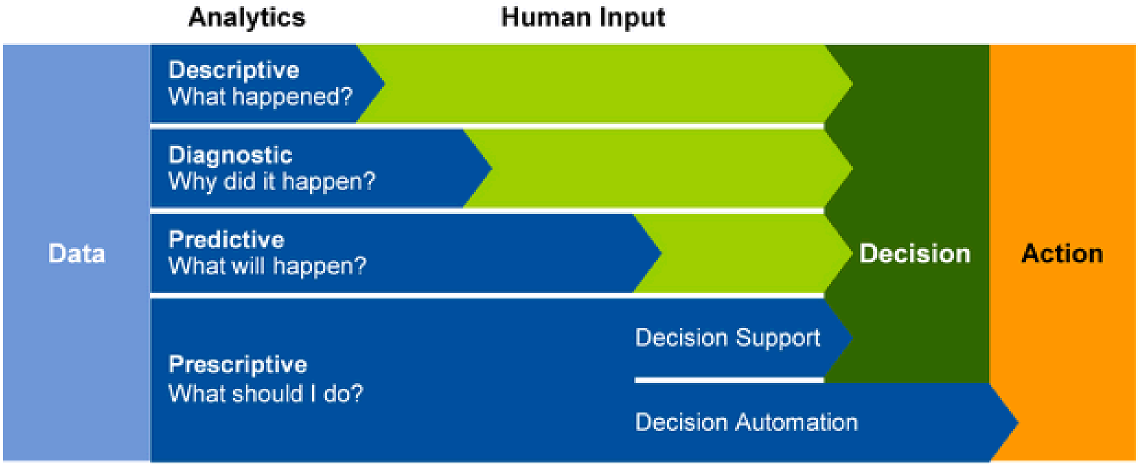

Figure 1. Four types of analytics capability (Gartner 2014)

BI uses data of the past. The user finds out what he believes is true and manually defines actions. Advanced analytics use the data of the past and automatically find hidden complex patterns and allow automatic identification and performance of the optimal actions. See Table 1.

5 CONCLUSIONS

Business intelligence/advanced analytics have attracted attention from enterprises and the research community due to the availability of big data and new business needs. Today, data-driven decision making via advanced analytics is a very fast growing field. In this paper a brief introduction to business intelligence, advanced analytics and big data has been given. Emphasis has been on advanced analytics. In order to find the right context, traditional business intelligence has been compared to advanced analytics.

Table 1. Business intelligence versus advanced analytics. (An Introduction 2015)

	Business Intelligence	**Advanced Analytics**
Orientation	Review	Future
Type of question	What happened When, who, how many	What will happen? What will happen if we change this one thing? What is next?
Methods	Reporting (KPIs, metrics) Automated Monitoring/Alerting Dashboards Scoreboards OLAP (Cubes, Slice & Dice, Drilling) Ad hoc query	Predictive Modeling Data Mining Text Mining Multimedia Mining Descriptive Mining Statistic/Quantitative Analysis Simulation & Optimization
Big Data	Yes	Yes
Data types	Structured, some unstructured	Structured and Unstructured
Knowledge Generation	Manual	Automatic
Users	Business Users	Data Scientists, Business Analysts, IT, Business Users
Business Initiatives	Reactive	Proactive

REFERENCES

An Introduction to Advanced Analytics. 2015, RapidMiner Accessed 8.8.2015. Published 2015. https://rapidminer.com/resource/introduction-advanced-analytics

Business Intelligence. 2015, Wikipedia, Accessed 8.8.2015. Published August 5, 2015 https://en.wikipedia.org/wiki/Business_intelligence

Business Intelligence (BI) 2013. Gartner IT Glossay. Accessed 8.8.2015. Published 2013. http://www.gartner.com/it-glossary/business-intelligence-bi

Chee, T., Chan, L. K., Chuah, M. H., Tan, C. S., Wong, S. F., & Yeoh, W. 2009, Business Intelligence Systems: State-of-the-Art Review and Contemporary Applications, in: *Symposium on Progress in Information & Communication Technology*, Vol. 2:4, pp. 16-30.

Gartner Says Advanced Analytics Is a Top Business Priority. 2014, Gartner. Accessed 8.8.2015. Published October 21, 2014. http://www.gartner.com/newsroom/id/2881218

Iankoulova, I. 2012, *Business intelligence for horizontal cooperation*, Master Thesis, University of Twente, The Netherlands.

Miller, S., Lucas, S., Irakliotis, L., Ruppa, M., Carlson, T., & Perlowitz, B. 2012, *Demystifying Big Data: A Practical Guide to Transforming the Business of Government*, TechAmerica Foundation, Washington, USA. Accessed 8.8.2015. Published 2012. http://breakinggov.com/documents/demystifying-big-data-a-practical-guide-to-transforming-the-bus/

Tasala, P. 2014, *Holistic Execution of Corporate Business Intelligence Strategy in a Heterogeneous Information Management Environment,* Master Thesis, Lapland University of Applied Sciences, Finland.

The Current State of Business Analytics: Where Do We Go from Here? 2012, Bloomberg Business Week Research Services. Accessed 8.8.2015. Published 2012. http://www.sas.com/resources/asset/busanalyticsstudy_wp_08232011.pdf

Vitt, E., Luckevich, M., & Misner, S. 2002, *Business Intelligence: Making Better Decisions Faster*, Microsoft Press.

Watson, H. J. 2009, Tutorial: Business Intelligence – Past, Present, and Future, in: *Communications of the Association for Information Systems*, Vol. 25, Article 39.

Watson, H. J. 2013, All about Analytics, in: *International Journal of Business Intelligence Research*, Vol. 4, Iss. 1, pp.13-28.

Watson, H. J. 2014, Tutorial: Big Data Analytics: Concepts, Technologies, and Applications, in: *Communications of the Association for Information Systems*, Vol. 34, Article 65.