ITHEA

International Journal INFORMATION THEORIES & APPLICATIONS

2019 Volume 26 Number 1

International Journal

INFORMATION THEORIES & APPLICATIONS
Volume 26 / 2019, Number 1

International Journal "INFORMATION THEORIES & APPLICATIONS" (IJ ITA)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITA welcomes scientific papers connected with any information theory or its application. IJ ITA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITA are given on *www.ithea.org*.

Responsibility for papers *published in* IJ ITA belongs to authors.

International Journal "INFORMATION THEORIES & APPLICATIONS"
Vol. 26, Number 1, 2019

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with: University of Telecommunications and Posts, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politécnica de Madrid, Spain,
Hasselt University, Belgium, University of Perugia, Italy,
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia
St. Petersburg Institute of Informatics, RAS, Russia,

Printed in Bulgaria
Publisher ITHEA®
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: office@ithea.org
Technical editor: **Ina Markova**

ISSN 1310-0513 (printed) **ISSN 1313-0463 (online)**

DECISION TREE ANALYSIS TO IMPROVE E-MAIL MARKETING CAMPAIGNS

Hamzah Qabbaah, George Sammour, Koen Vanhoof

***Abstract**: The efficiency of e-mail campaigns is a big challenge for any e-commerce venture in terms of the response rate of e-mail campaigns and customer segmentation based on loyalty. Decision tree analysis are useful tools to extract customer information related to response rate from e-mail campaigns data. This study aims at predicting customer loyalty and improving the response rate of e-mail campaigns, specifically open rate and click through rate, using decision tree analysis such as CHAID , CART and QUIST.*

The methodology used in this study is Cross Industry Standard Process for Data Mining (CRISP – DM) methodology. The models are trained and tested using split sample validation. Furthermore, we used a classification measures to calculate the accuracy, precision, recall and F1 to evaluate the models. The models reported satisfactory results in predicting customer loyalty based on open rate, click through rate values and on customer demographic variables. The response rates also increase at the preferred moment at which e-mails should be send to customers in email campaigns.

***Keywords**: E-Commerce, E-mail campaigns, CRISP-DM, Decision tree, open rates, click through rates.*

1. Introduction

The rapid development and popularization of internet use, has created an extraordinary growth of e-commerce and online shopping (Liu et al. 2015). We understand e-commerce to be the buying and selling of products or services through electronic media (Carmona et al. 2012). One important element of e-commerce practices is web advertising. Companies can address customers

directly using different channels such as e-mail campaigns and contextual advertising (Xuerui Wang 2010).

The appeal of e-mail communication to talk directly to customers is double: cost effectiveness and time efficiency (Cases et al. 2010). In order to reach these objectives, companies wanting to use e-mail as a direct communication channel with their customers, need to thoroughly understand how e-mail campaigns may affect customers' attitudes and behaviour (Shan et al. 2016). This knowledge can lead to a more optimal design of their e-mail campaigns and thus be turned into a competitive advantage.

This subject has attracted much attention in recent years. E-mail marketing research studies have been conducted amongst others by online surveys, in-depth interviews, controlled experiments and by tracking customer behaviour. Click-through links and the visiting patterns have been used as descriptive variables but few of the studies have really investigated the effects of e-mail characteristics on customer attitudes and behavioural intentions (George Sammour 2009).

Loyalty and response rates have been an important focus of attention in the research on e-mail campaigns (i.e., more than one email sent by a company) (Cases et al. 2010). Building models to improve response rates by using individual preferences of customers to personalize e-mail newsletters is the major topic of this research stream. The reason of this is obvious : marketing campaigns and products can on the basis of the results of this research be customized to have a more effective appeal to groups of customers or to individuals (George Sammour 2009).

These data are in most companies abundantly available. They are a very valuable resource when efficient access to the data is created and salient information is extracted from them. Making efficient use of the information has thus become an urgent need. Data mining is an excellent tool to extract or detect hidden customer characteristics and behaviours from large databases (Ngai, Xiu, and Chau 2009; Chiu et al. 2009). It can be defined as "the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. It allows corporations to improve its marketing,

sales, and customer support operations through a better understanding of its customers" (Michael J. A. Berry 2004). For many organizations, the goals of data mining essentially include improving marketing capabilities, detecting abnormal patterns, and predicting the future based on past experiences and current trends.

Web mining is the use of data mining techniques to automatically discover and extract knowledge from data available on the use of websites (Etzioni 1996). The importance of web mining is considering the behaviour and preferences of the user of websites(Cooley, Mobasher, and Srivastava 1999). Several authors subdivide web mining in several stages: (1) Finding resources, (2) Selecting information and pre-processing of the data, (3) Discovering knowledge and finally (4) Analysing the obtained patterns(Liu 2006; Kosala and Blockeel 2000). This research analysis uses web usage mining, which is defined as "The process of applying data mining techniques to the discovery of usage patterns from Web data" (Srivastava et al. 2000).

2. Problem statement and Research questions

Customer loyalty is a complex phenomenon. Most research has focused on trying to look at it as a factor in purchase behaviour or as a final action by the customer. The link between customer loyalty and the steps in buying behaviour preceding the purchase, like e-mail behaviour have not been the focus of much attention. Our goal is to look at the link between this pair of variables, namely by researching into the impact of customer demographic and customer e-mail behaviour on customer loyalty and to see whether it can be used as a predictor of loyalty.

From this, we can derive our research questions. They are:

1. What is the effect of some descriptive variables of the e-mail behaviour of customers on customer loyalty?
2. Which model predicts customer loyalty best on the basis of both the customer demographic and customer e-mail behaviour variables ?

Next section we describe the methodology we used to answer the research questions.

3. Methodology

The data mining methodology used in this research was Cross Industry Standard Process for Data Mining (CRISP – DM) and the methodological steps were followed: business understanding, data understanding, data preparation, modeling, evaluation and deployment, as shown in figure 1.

- **Business understanding**

Email marketing for E-Commerce companies focusses on sending relevant emails to the right customers at the right time. It has thus become a "widely-used marketing tool for companies to communicate with customers, handle customer complaints, and cross-sell and up-sell products". (Zhang 2015) The goal of an email marketing is consequently to generate profit by making customers become more active in purchases, taken into consideration that the customer is always entirely free to opt-out.

E-mail marketing is thus very closely linked to increasing customer loyalty. Loyalty has been used in a business context, to describe a customer's willingness to continue buying from a company over the long term, preferably on an exclusive basis, and recommending the company's products to friends and associates (Lovelock and Wirtz 2011). Reichheld and Jr suggest that "customer repeat purchase loyalty must be the yardstick of success" (F. Reichheld and Jr. W. E 1990). Customer loyalty is indeed widely seen as a key determinant of a firm's profitability (Reinartz and Kumar 2002). The increased profit from loyalty comes from reduced marketing costs, increased sales, and marketing cost (T. and Shiang-Lih 2001). Loyalty reduces the need to incur customer acquisition costs (Reichheld and Teal 1996).

The common important efficiency metrics of e-mail marketing are thus : Delivery rate, Open rate and Click-through rate. Increased rates lead to an improve customer buying process, increased loyalty and increased firms profits.(Stokes 2011)

Figure 1. CRISP –DM life cycle.

- **Data understanding and Data Preparation**

Data have been obtained using the webmaster tool of Google Analytics from an E-commerce website. Cleaning, merging and pre-processing of the data has been applied in order to obtain the final data set. As such we filtered out customers who received all 32 campaigns, resulting in a sample of 1428 customers (n=1428). For each customer we collected information such as ID, gender, age, country , total e-mails received, total e-mails opened , total e-mails clicked, loyalty segment2 and the time of the campaign mail was opened. After that we calculated the click rate, open rate and the average time of opening the campaign for each customer. Table one shows the data type of the variables that are used in our study.

The company offers a loyalty program to their customers: loyal customers receive numerous additional benefits. Customers can buy a loyalty membership or get it through some consumption formula. For privacy reasons

the company did not explain it, but they clarified that one of the important key factors affecting the loyalty program is collecting points by purchasing their products online using their website. Both the number of purchases and the value of the purchases play a role. Therefore, none of our independent variables are in the formula of the loyalty programme membership.

Since our goal is to predict customer loyalty, we analysed the impact of customer demographics and customer e-mail behaviour characteristics on customer loyalty. Consequently, the independent variables in our analysis are the customer demographics (determined by age, gender and country of residence) and the customer e-mail behaviour (determined by the parts of the day and the open and click through rates. The dependent variable is the customer loyalty segment.

Table 1. Study variables.

Variable	**Data type**
Customer ID	Continuous
Gender	Categorical
Age	Continuous
Country	Categorical
Open rate	Continuous
Click through rate(CTR)	Continuous
Loyalty segment2	Categorical
Part of the day(POD)	Categorical

- **Modeling**

Our research questions indicate that the main objective of this research is to find a way of predicting customer loyalty based on their response pattern to e-mail campaigns and linking that to their demographic and behavioural characteristics.

This goal can be achieved by experimenting with different campaigns and comparing the resulting customer e-mail behaviour to them. These data must then be linked to demographic characteristics of the respondents. This is however very difficult to realize since companies are reluctant to experiment with real life campaigns. They indeed fear that the response rates will drop due to the experiment itself.[6] This research is therefore based on the use of decision tree technique of the web data that allow to experiment with campaigns promising a high potential for increased response rate levels .

Decision tree analysis as a method of data mining techniques allows to achieve the major steps needed to achieve the required objectives of analysing and predicting models that measure the impact of customers demographic and behavioural characteristics to customer's loyalty .[6] Sections 4 and 5.1 will explain the decision tree algorithms used in this study and present the analysis and results.

- **Evaluation and Deployment**

The confusion matrix used in this study to evaluate the models by determining the classification measures: Accuracy, precision, recall and F1 of these models." (Carmona et al. 2012).

Table 2. shows the confusion matrix. Where TP is defined as normal behaviour that is correctly predicted, FP indicates normal behaviour wrongly assumed whereas abnormal TN specifies the normal performance that is detected as correct and FN indicates the abnormal performance that is misidentified as normal (Sumaiya Thaseen and Aswani Kumar ; Lopes and Roy 2015).

Table 2. Confusion matrix

	Observed	
Predicted	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

Based on this confusion matrix we can measure accuracy, recall , precision and F-score as defined in equations 1 to 4 (Aziz Yarahmadi 2017).

Accuracy is used to measure the effectiveness of proposed models based on the variables in the equation , it is defined as the ratio of correctly predicted observation to the total observations.

$$\text{Accuracy} = TP+TN /(TP+TN+FP+FN) \quad (1)$$

Recall is the ratio of correctly predicted positive observations to the all observations, recall used to measure whether the selected variables that are relevant.

$$\text{Recall} = TP / (TP+FN) \quad (2)$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, precision detects the fraction of all recommended variables that are relevant.

$$\text{Precision} = TP / (TP+FP) \quad (3)$$

F-measure combines precision and recall which is the harmonic mean of precision and recall.

$$F_1 = 2* ((\text{Recall}*\text{Precision})/(\text{Recall}+\text{Precision})) \quad (4)$$

These metrics are focused on the performance of the models , all the three derived metrics should be close to 1 for a good model.

Managerial Consequences and findings are shown in section 6. Next section we describe the Decision tree algorithms used in this study.

4. Decision tree analysis

Decision tree analysis is one of data mining techniques that can encompass numerous algorithms to predict a dependent variable. These predictions are determined by the influence of independent predictor variables. It is a structure that can be used to divide a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. With each successive division, the members of the resulting sets become more and more similar to one another. Decision trees thus use a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable. (Linoff and Berry 2011).

Each procedure goes as follows. After the population is split into *n* nodes, the process is repeated on each node until the data is completely partitioned out or until a stopping rule is met. This means that there are no longer enough cases to partition out, or only one case remains for the last node. The entire classification of data is then published through a graphical tree, which explains the interaction of *x* (*x1, x2, …, xm*) and *y*. An ideal decision tree is one that has no limitations; this tree would correctly classify each individual case. Since this is impossible, the classification procedures will attempt to provide as much separation as possible in regards to classifying data into correct cases (Long et al. 1993). Graphically, decision trees will produce a tree *T* which is comprised of a root node and child nodes (De'ath 2000) (Ripley and Hjort 1995). The tree is essentially a connected graphic that is inverted; however, the tree display is user specific and can be displayed horizontally (Loh and Shih 1997; L. and Jr 2012)

There are numerous decision tree classification procedures. The algorithms used in this paper are CHAID, CART and QUEST. A brief description of these techniques is added.

4.1 The Chi-Squared Automatic Interaction Detection (CHAID) algorithm

CHAID is the most popular classification procedure because it can manage both continuous and categorical variables (Díaz-Pérez and Bethencourt-Cejas

2016). The algorithm has the ability to choose different statistical techniques for splitting, determined by the level of measurement for the dependent variable. Since the algorithm can handle continuous or categorical dependents, multiple statistical techniques are available. If *the* dependant variable is continuous, an F-Test will be selected for splitting. In contrast, if *it* is either ordinal or nominal then a maximum likelihood estimate is selected. The user has the ability to select Pearson-Chi when *the variable* is nominal (L. and Jr 2012).

In regards to splitting, CHIAD is not bound to binary splits and allows for multiple level splits at each node. The algorithm searches for the best possible split for the different values of the independent variable then determines if the data should be merged to form a node or split in the data. Essentially, the best predictor is selected to begin splitting the data. During this process the CHAID is selecting the best predictor based on comparisons of the adjusted p-value for each predictor (L. and Jr 2012). This is a three step process for determining splitting. The splitting and merging procedures for CHAID are independent of one another. The algorithm uses as sequential process where splitting and merging occur at the same time. The basis behind this aspect of the algorithm is rooted in the computation time of *T*. CHAID has difficulty determining when and where to stop (Loh and Shih 1997). Stopping is also a three step process determined by the chi-square or F-Test statistic. As such, stopping occurs when the critical level of the selected test statistic (chi-square or F) fails to meet the test.

The Independent variable statistics in the tree display of CHIAD algorithm shown are *F* value (for scale dependent variables) or chi-square value (for categorical dependent variables) as well as significance value and degrees of freedom.(IBM 2016)

4.2 Classification and regression trees (CART/C&RT)

This algorithm is a decision tree program very well introduced in the statistic community (Long et al. 1993). It produces a binary decision tree by restricting the partitions at each node to two (G. T. Denison, K. Mallick, and F. M. Smith

1998). CART algorithm uses an extensive and exhaustive search of all possible univariate splits to determine how the data have to be split for the classification tree (Breiman et al. 1984). Partitioning will continue until the algorithm is unable to produce mutually exclusive and homogenous groups further (De'ath 2000; G. T. Denison, K. Mallick, and F. M. Smith 1998). At this point, the node is considered a terminal node.

CART can account for missing values by estimating a prediction based on other independent factors or "surrogates" (Breiman et al. 1984). The CART algorithm thus allows for the inclusion of partial data. Therefore missing values or cases are not removed and bias is reduced(De'ath 2000). CART handles these missing values for predictor variables in two ways. First, if the predictor variable has missing values, it can be restricted from moving farther down the tree. Second, it can send them farther down the tree (De'ath 2000). If the predictor variable is stopped, the cases with missing values will be recorded with response values as the rest of the cases in that node. If the predictor variable with missing cases is sent further down the tree, its missing values will be replaced by responses from another independent factor with non-missing cases. CART will determine which method produces the best agreement with the splitting variable. Since CART is a forward stepwise method, data that may not be influential in predicting the dependant variable can however be included in the tree leading to extensive and lengthy trees since the algorithm uses an exhaustive search to determine the splits in the data (L. and Jr 2012).

The Independent variable statistic in the tree display of CART algorithm shown is improvement value . The minimum decrease in impurity required to split a node is 0.0001. Higher values tend to produce trees with fewer nodes, this called Minimum change in improvement. To measure impurity and the minimum decrease in impurity required to split nodes. In this for categorical (nominal, ordinal) dependent variables, the impurity measure selected is: Gini. Splits are found that maximize the homogeneity of child nodes with respect to the value of the dependent variable. Gini is based on squared probabilities of membership for each category of the dependent variable. It reaches its minimum (zero) when all cases in a node fall into a single category. This is the default measure.(IBM 2016)

4.3 Quick unbiased efficient statistical trees (QUEST)

QUEST is a binary splitting algorithm for building decision trees. The algorithm is essentially a complex and technical version of CART. While the two algorithms share commonalities, the major difference is the speed in which they are produced (L. and Jr 2012).

The most significant advantage in choosing QUEST over another algorithm lies in the speed of computation, QUEST uses to reduce the processing time required for large CART Tree analyses. However, while there is a significant increase in speed, accuracy in prediction is not reduced, Sometimes QUEST is better and other times other algorithms is better (Loh and Shih 1997).The algorithm contains no identified bias in variable selection for the splitting. The lack of bias in selection is controlled by the algorithm because it handles the variation of levels in *the independent variables x*. Therefore, QUEST can control for the inclusions of $x_1, x_2, \ldots, x_m$ when variables have variations in levels.

There are two steps within the splitting aspect of the QUEST algorithm. These steps are dependent upon the significance test to split each node (Loh and Shih 1997). The first step of the algorithm examines the association of each predictor variable (X_i) with respect to the dependent variable y. Then the most significant variable is selected. After this step is finished, the algorithm conducts an exhaustive search for set *S*, which is the subset of values taken by the predictor variable in the node above. QUEST does not conduct exhaustive searches unless the variable is unordered and has a small number of values. QUEST is significantly better when searching for variables because selection bias and computational cost were not present, but the accuracy is identical to the other algorithms (Loh and Shih 1997; L. and Jr 2012).

The Independent variable statistics in the tree display of QUEST algorithm shown are F,

significance value, and degrees of freedom for scale and ordinal independent variables; and for nominal independent variables, chi-square, significance value, and degrees of freedom are shown.(IBM 2016)

The analysis and results of the use of these methodologies will be presented in section 5.

5. Analysis and Results

The results of the decision tree analysis are presented in section 5.1 which we present three models according to which we have tried to predict customer loyalty using the three decision tree algorithms , Next on section 5.2 we present the evaluation of the models using classification measures tests.

5.1 Prediction models of customer loyalty using decision tree analysis

In this section we create a three models using three decision tree algorithms (CHAID, CART and QUEST) to predict customer loyalty based on our independent variables. They can be described as follows:

Model 1: Customer loyalty based on customer demographic variables.

Model 2: Customer loyalty based on customer e-mail behaviour variables

Model 3: Customer loyalty based on customer age, gender, part of day, open and click-through rates.

For each decision tree we have used the **split sample validation**, the data were split into two sets: a training set of 80%, and a testing set of 20 %. The model trained will be tested on the test data to verify whether they give similar results. The total number of customers present in each node is based on the loyalty variable (category "Not loyal" or category "Loyal"). Then each time, the rules extracted from the tree are mentioned.

Since three decision trees were developed for three different models, in total nine analysis sets are presented.

1. Decision trees for Model 1 (predicting loyalty based on customer demographic variables)

The decision tree using the CHAID method is presented in Figure 2.

The rules explaining the nodes in Figure 2. are as follows :

Node 1:If (Age <=45) then class-loyal is (81.7%) and class-non loyal is 18.3%.

Node 3:If (Age>45) and (gender is male) then class-loyal is 20.7% and class non loyal is 79.3%.

Node 4: If (Age>45) and (gender is female) then class loyal is 15.3 % and class non loyal is 84.7%

It is clear that according to the CHAID model the country of residence does not play an important role in predicting loyalty as it does not figure in the decision rules at the different nodes. The strongest independent variable predicts loyalty segment is age since X^2=385 (df=1, p-value<0.01).

The decision tree using the CART method is presented in Figure 3.

The nodes in Figure 3 can be explained by the following rules :

Node 1: If (Age<=45.5) then class loyal is 88.6 % and class non loyal is 11.4%

Node 4:If (Age>45.5) and (gender is female) then class loyal is 16.9% and class non loyal is 83.1%

Node 6 : If (Age>63.5) and (gender is male) then class loyal is 33.3% and class non loyal is 66.7%.

Node 7: If (45.5<Age<55.5) and (gender is male) then class loyal is 33.3% and class non loyal is 66.7%.

Node 8: if (63.5>age >55.5) and (gender is male) then class loyal is 24.2% and class non loyal is 75.8%.

Again it appears as if the country of residence plays a less important role than the other demographic variables in explaining the nodes of this CART decision tree. The strongest independent variable predicts loyalty segment is also age since the improvement value is 0.163 .

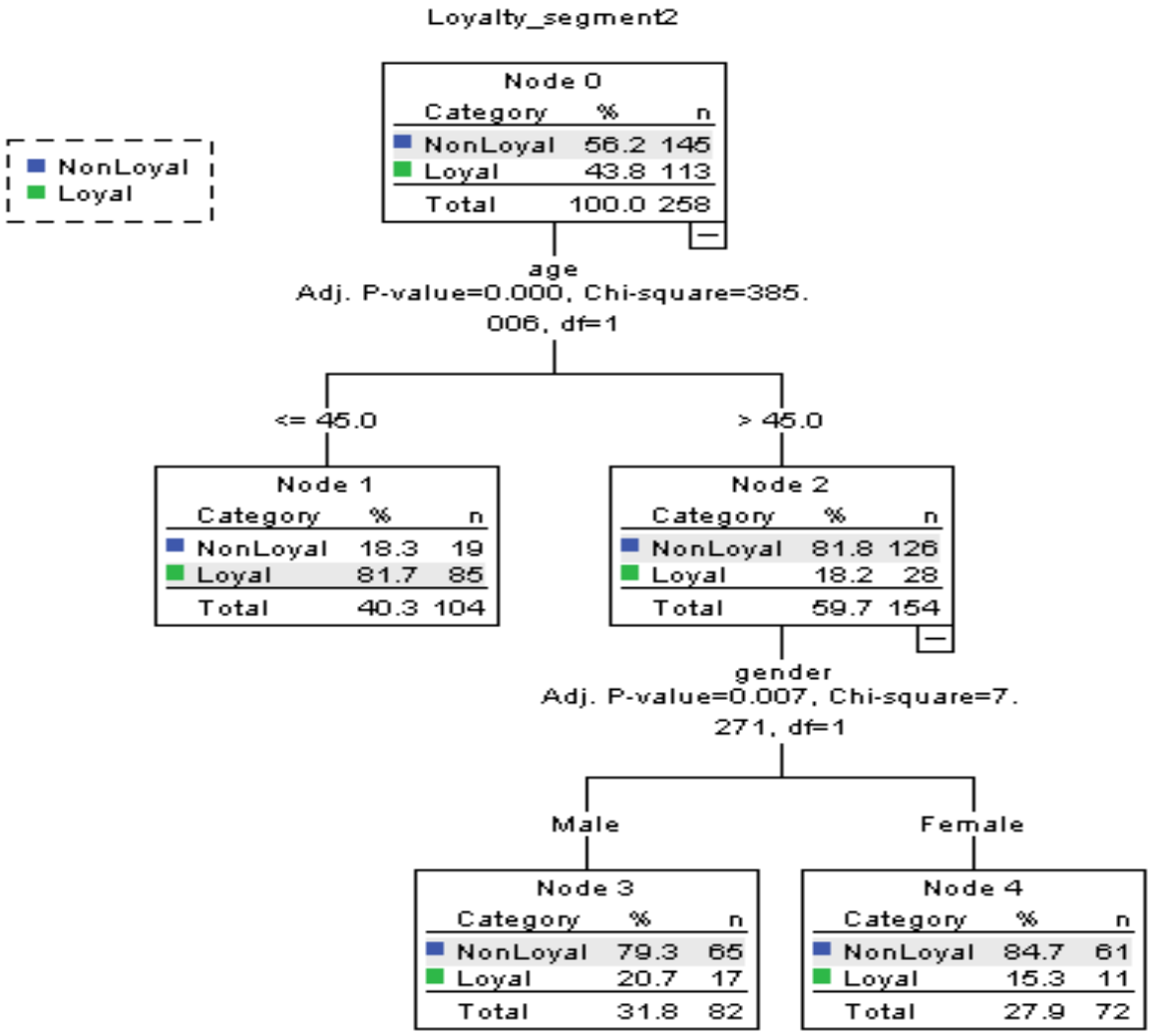

Figure 2. CHAID decision tree for Model 1 (using customer demographic variables)

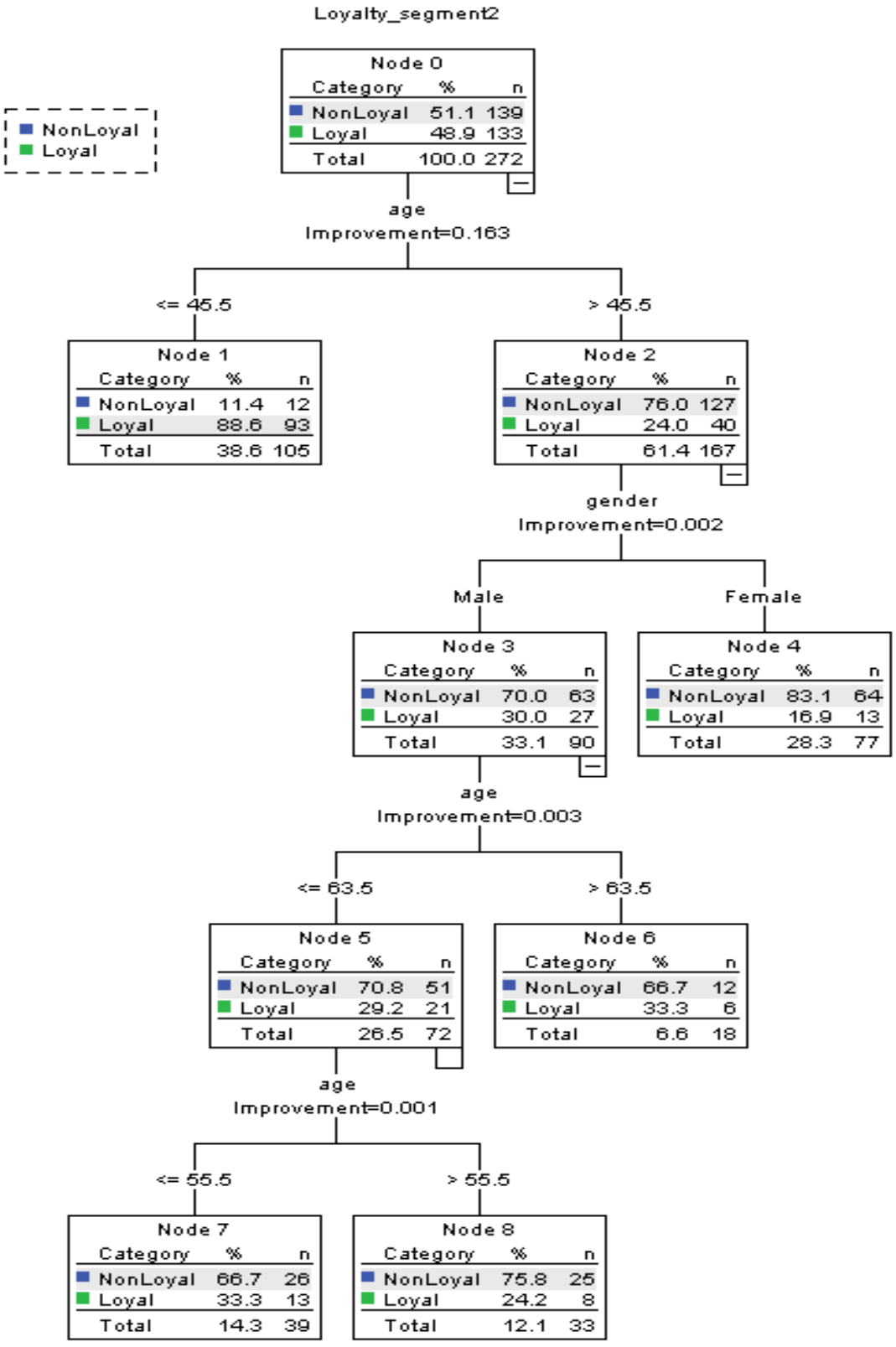

Figure 3. CART decision tree for Model 1 (using customer demographic variables)

The decision tree using the QUEST model is presented in Figure 4. The rules explaining the decision tree in Figure 4 are fairly simple. They read as follows :

Node 1: if (age<=43.5) then class loyal = 86.2% and class non loyal = 13.8%.

Node 2: if (age >43.5) then class loyal = 26.5% and class non loyal = 73.5%

Using this model both gender and country of residence do not seem to play an important role in prediction loyalty of the customers. While Age is the only independent variables effects loyalty segment based on QUEST method in this model, since F=471.2 and P-value<0.01.

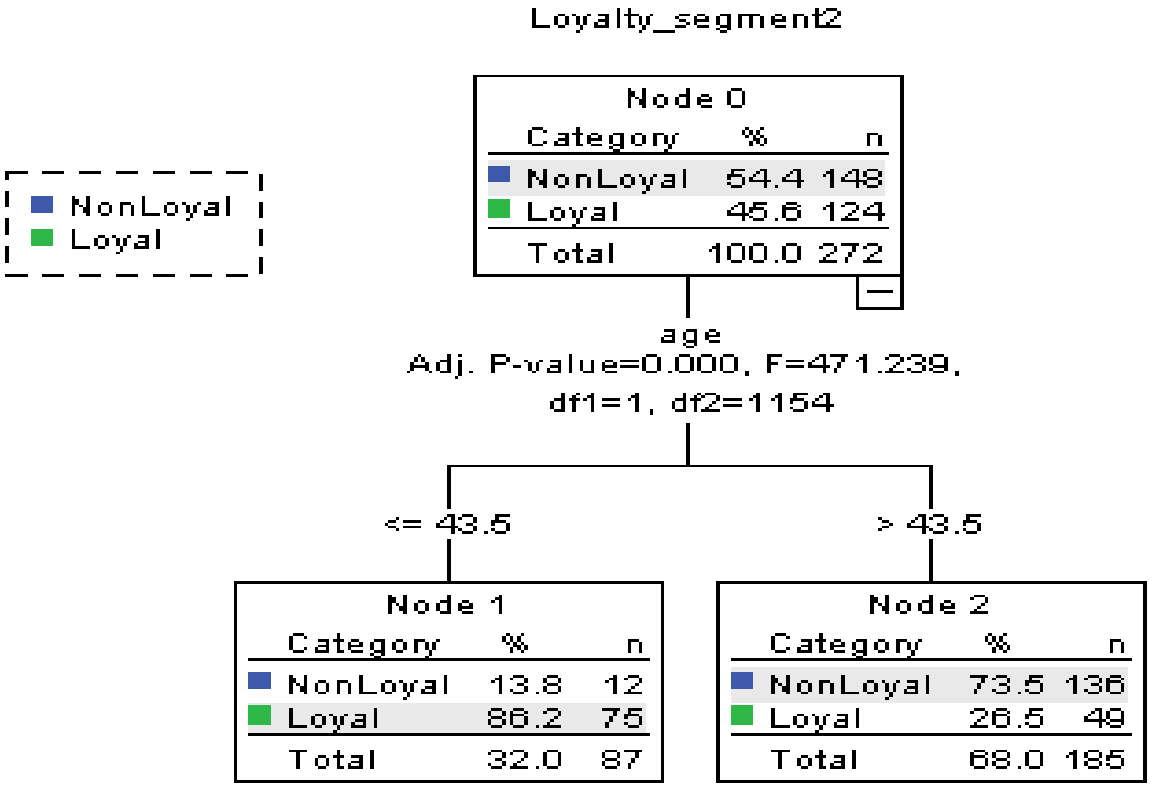

Figure 4. QUEST decision tree for Model 1 (using customer demographic variables)

2. Decision trees for Model 2 (predicting loyalty based on customer e-mail behaviour variables)

The decision tree using the CHAID method is presented in Figure 5.

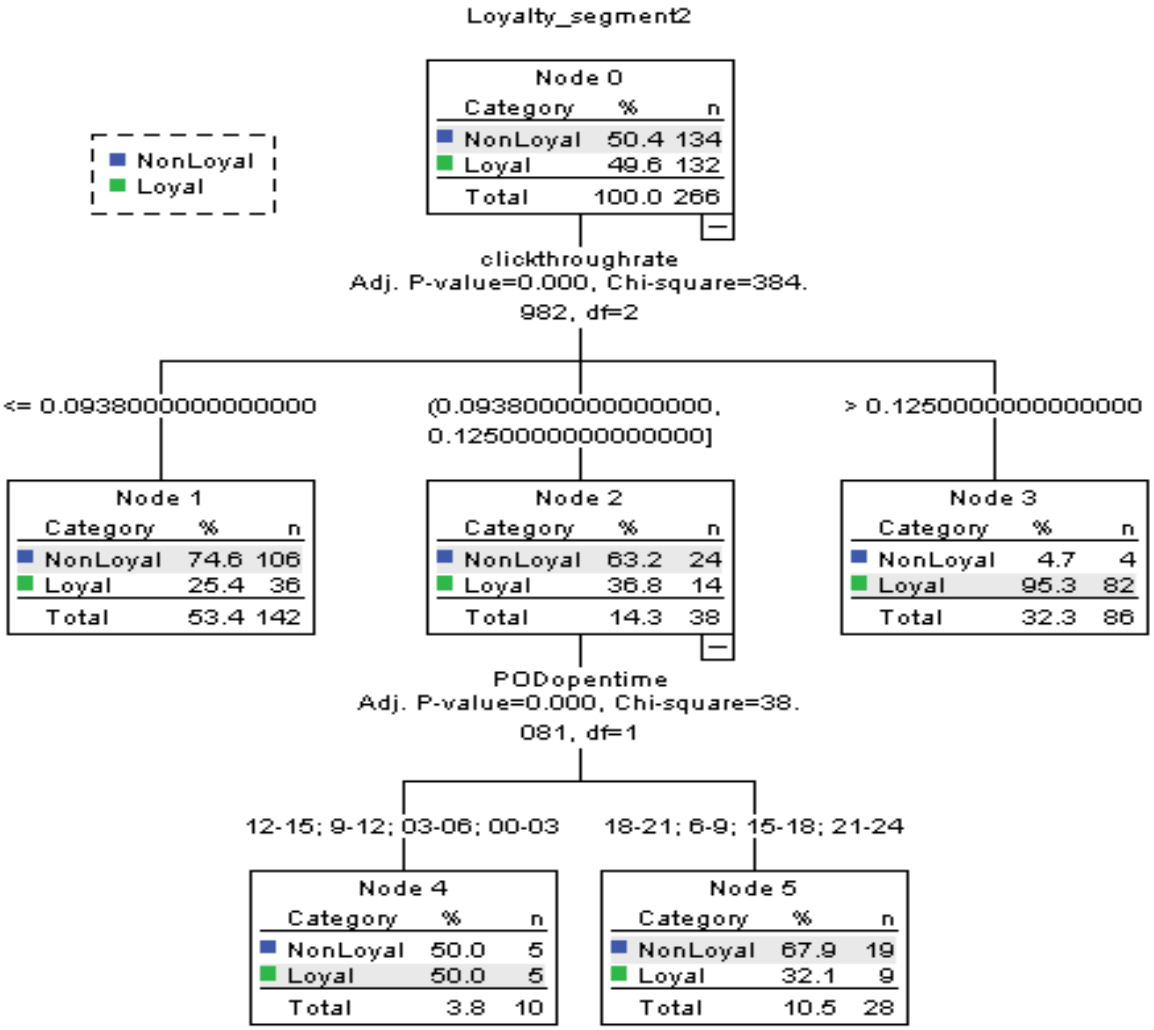

Figure 5. CHAID decision tree for Model 2 (using customer e-mail behaviour variables)

The CHAID method is somewhat more complicated for this model when the rules of the different nodes are read completely. They are :

Node 1 : if (click-through rate (CTR) < 0.094) then class loyal = 25.4% and class non loyal = 74.6%.

Node 3: if (CTR > 0.125) then class loyal = 95.3% and class non loyal = 4.7%.

Node 4: if (0.094< CTR < 0.125) And (POD open time on [12-15] OR[9-12] OR[03-06] OR [00-03]) then class loyal =50% and class non loyal = 50%.

Node 5: if (0.094< CTR < 0.125) And (POD open time on [18-21] OR[6-9] OR[15-18] OR [21-24]) then class loyal =32.1% and class non loyal = 67.9%.

According to CHAID method in this model the strongest independent variable predicts loyalty segment is CTR since X^2=384 (df=2, p-value<0.01).

The decision tree using the CART method is presented in Figure 6.

The rules that guide the decision tree presented in Figure 6 are as follows :

Node 4: If (CTR<=0.14) and (open rate>0.55) then class loyal = 55.6% and class non loyal = 44.4%.

Node 5: If (CTR>0.14) and (POD open time between [9-12] OR [18-21] OR [21-24]) then class loyal = 93.2% and class non loyal = 6.8%.

Node 6: If (CTR>0.14) and (POD open time between [12-15] OR [15-18] OR [6-9] OR [00-03] OR [03-06]) then class loyal = 87.2% and class non loyal = 12.8%.

Node 8: if (0.14>CTR> 0.11) and (open rate<= 0.55) then class loyal = 31.6% and class non loyal = 68.4%

Node 10: if (CTR< 0.11) and (open rate<= 0.55) and (POD open time between [15-18] or [21-24] or [6-9] or [00-03] or [03-06]) then class loyal = 27% and class non loyal = 73%.

Node 11: if (CTR< 0.11) and (open rate<= 0.55) and (POD open time between [12-15] or [9-12]) then class loyal = 16.7% and class non loyal = 83.3%.

Node 12: if (CTR< 0.11) and (open rate<= 0.55) and (POD open time between [18-21]) then class loyal = 25% and class non loyal = 75%. Contrary to the CHAID method and the CART method clearly also uses the CTR (click-through

rate) in developing the tree model. since the improvement value of the first split was CTR =0.157 .

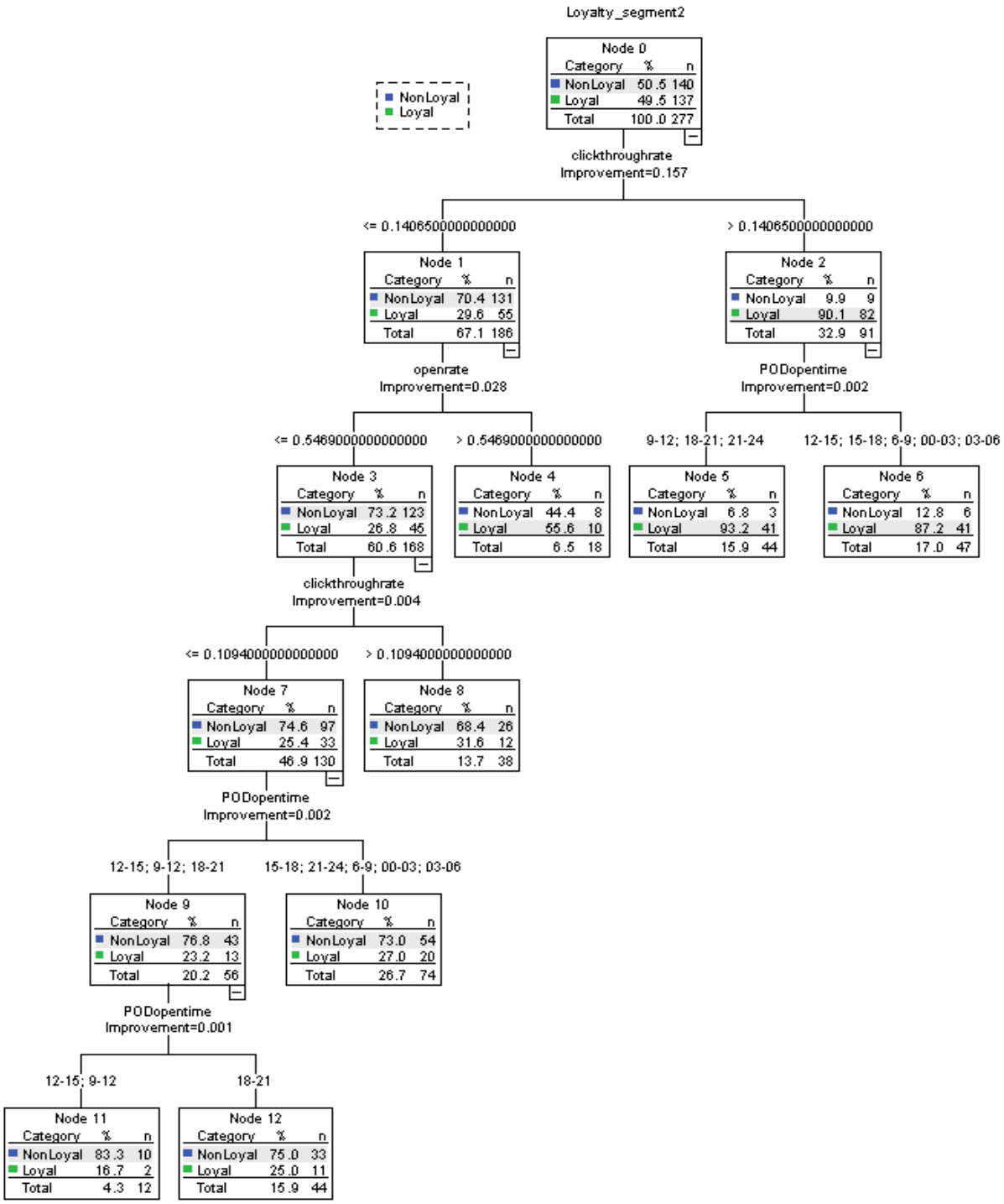

Figure 6. CART decision tree for Model 2 (using customer e-mail behaviour variables)

The decision tree using the QUEST method is presented in Figure 7.

The rules explaining the nodes of this QUEST method analysis are simple, Since only CTR effects loyalty segment based on QUEST method in this model, the F=455.8 and P-value <0.01. The rules are:

Node 1: if (CTR <= 0.155) then class loyal = 31.9% and class non loyal = 68.1%.

Node 2: if (CTR>0.155) then class loyal is 84.7% and class non loyal = 15.3%.

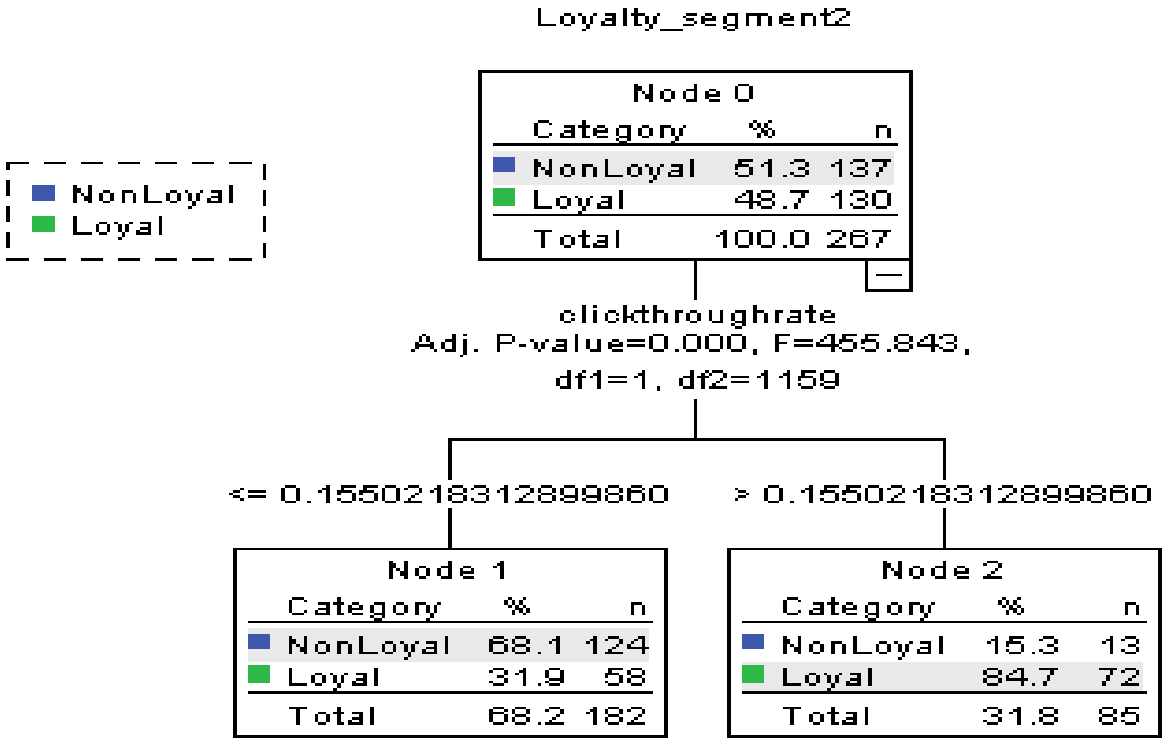

Figure 7. QUEST decision tree for Model 2 (using customer e-mail behaviour variables)

3. Decision trees for Model 3 (predicting loyalty based on age, gender, open rate, click-through rate and Parts of the day)

The decision tree using the CHAID method is presented in Figure 8.

The rules explaining the nodes in Figure 8 are as follows :

Node 3: If (age <=44) and (CTR<=0.22) then class loyal = 78.8% and class non loyal = 21.2%.

Node 4: If (age<=44) and (CTR>0.22) then class loyal = 93.2% and class non loyal = 6.8%.

Node 7: If (44<age<=57) and (POD open time between [12-15] or [18-21] or [21-24]) then class loyal = 23.2% and class non loyal = 76.8%.

Node 8: If (age >57) and (POD open time between [12-15] or [18-21] or [21-24]) then class loyal = 24.2% and class non loyal = 75.8%.

Node 9: If (POD open time between [9-12] or [6-9] or [15-18] or [03-06] or [00-03]) and (open rate<=0.44) and (age >44) then class loyal = 25.6% and class non loyal = 74.4%.

Node 10: If (POD open time between [9-12] or [6-9] or [15-18] or [03-06] or [00-03]) and (open rate>0.44) and (age>44) then class loyal = 33.3% and class non loyal = 66.7%.

Age is the strongest independent variable effects loyalty segment in this model since the first split of the tree is based on it , the X^2 is 374 (df=1, p-value<0.01). CTR, open rate and POD open time variables also play an important role in predicting loyalty in this model.

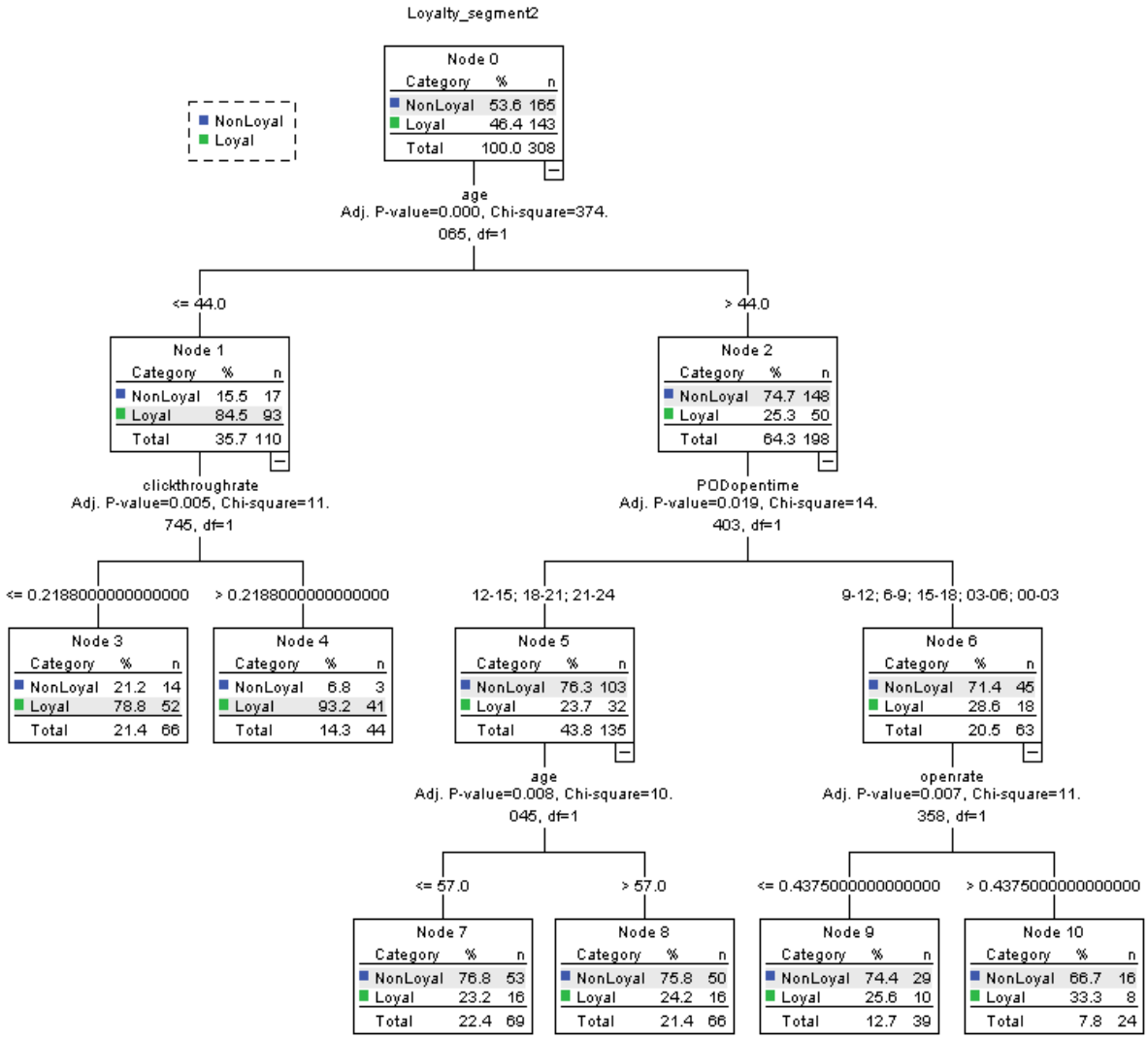

Figure 8. CHAID decision tree for Model 3

The decision tree using the CART method is presented in Figure 9.

The rules explaining the nodes of this CART method are very simple comparing to the model 1 and model 2, the nodes are based only on the age variable, and the improvement value was 0.166 .

The rules are:

Node 1: If (age<45.5) then class loyal = 86% and class non loyal = 14%.

Node 2: If (age >=45.5) then class loyal =23.4% and class non loyal = 76.6%.

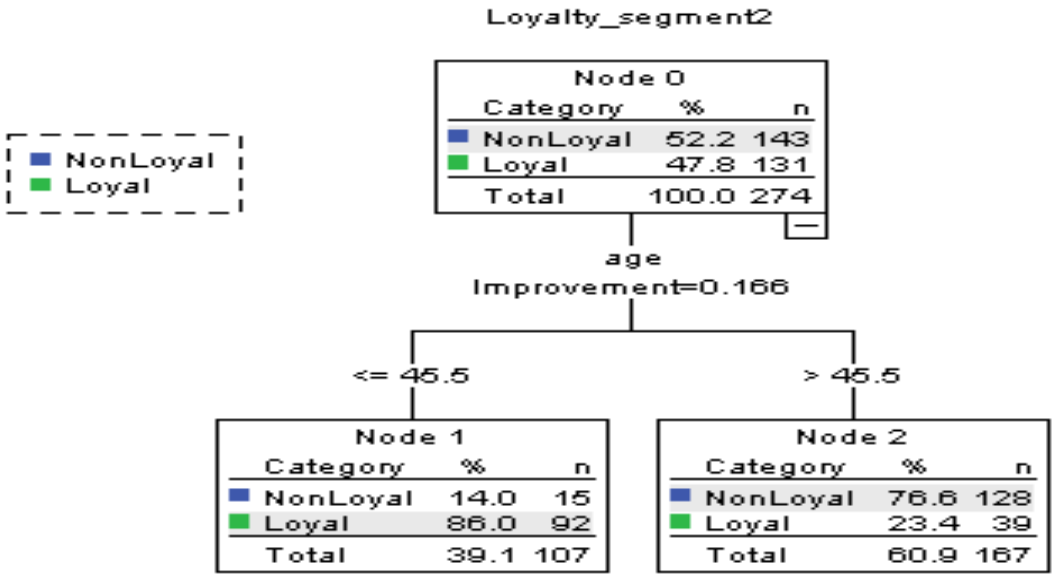

Figure 9. CART decision tree for Model 3

The decision tree using the QUEST method is presented in Figure 10.

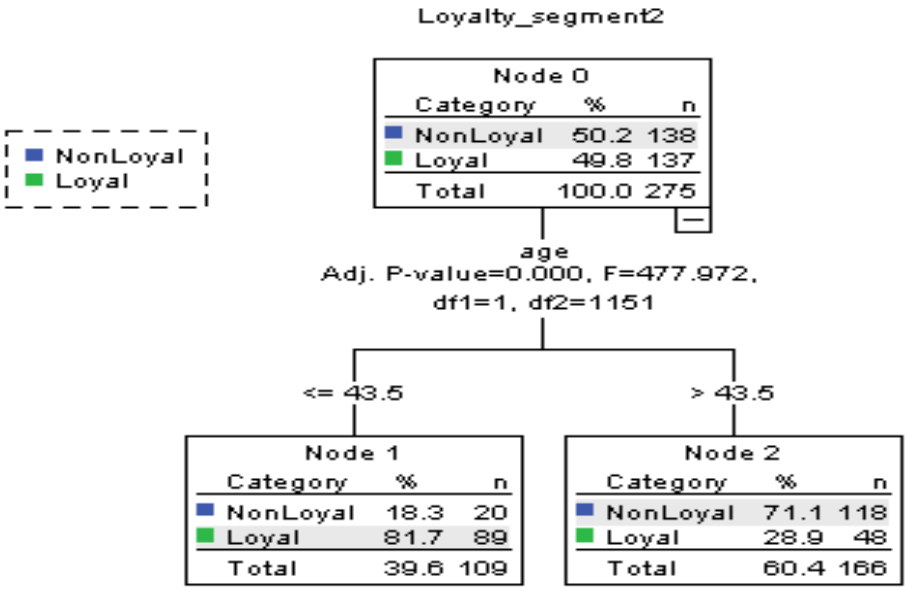

Figure 10. QUEST decision tree for Model 3

The rules explaining the nodes of this QUEST method are simple as the rules extracted from the model 1 and model 2, the nodes are based only on the age variable, since F=478, and P-value<0.01. The rules are:

Node 1: If (age<43.5) then class loyal = 81.7% and class non loyal = 18.3%.

Node 2: If (age >=43.5) then class loyal =28.9% and class non loyal = 71.1%.

Looking at all these models, we can make some initial conclusions. First, The CART method is clearly more complex than the other the two methods we have tested. The resulting decision trees have more nodes and rules of subdivision. Second, age and click-through rate play the most important role in subdividing the population according to customer loyalty, with the open rate to a lesser degree. The other variables are less important in the decision trees resulting from these methods.

5.2 Evaluation the models using classification measures tests

Using To evaluate the quality of these models four classification measures, accuracy, recall, precision and F1- score as mentioned in the methodology section have been used to see which of the three models predicts customer loyalty best.

Figure 11 Shows the accuracy of the three models according to the decision tree analysis.

Given these results we can conclude that the accuracies achieved by CHAID and CART methods are higher than the accuracies achieved by QUEST method.

The precision measures for the decision tree analysis are presented in Figure 12.

From figure 12 we can conclude that the precision of the model 2 achieved by CHAID method using only the behavioural variables is higher than the precision of the other models using any of the three decision tree models.

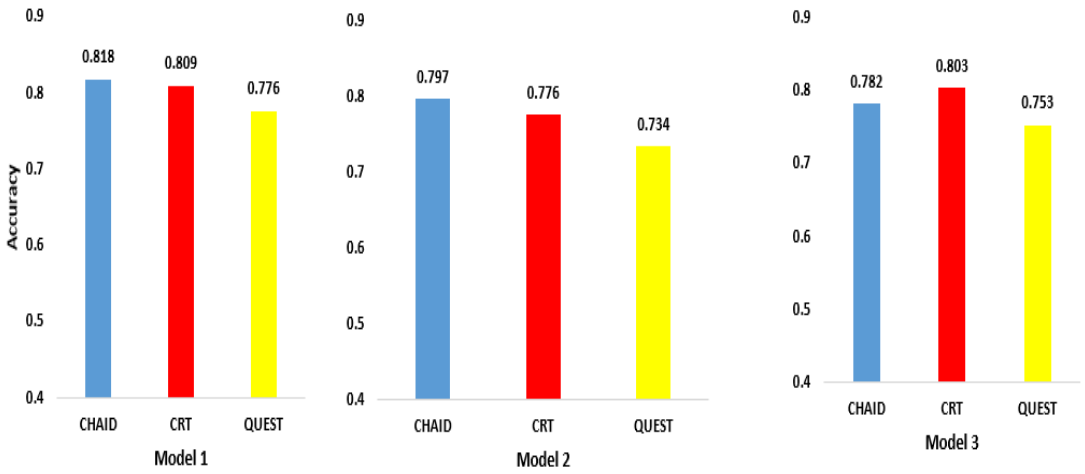

Figure 11. Accuracy test of the results of the decision tree models

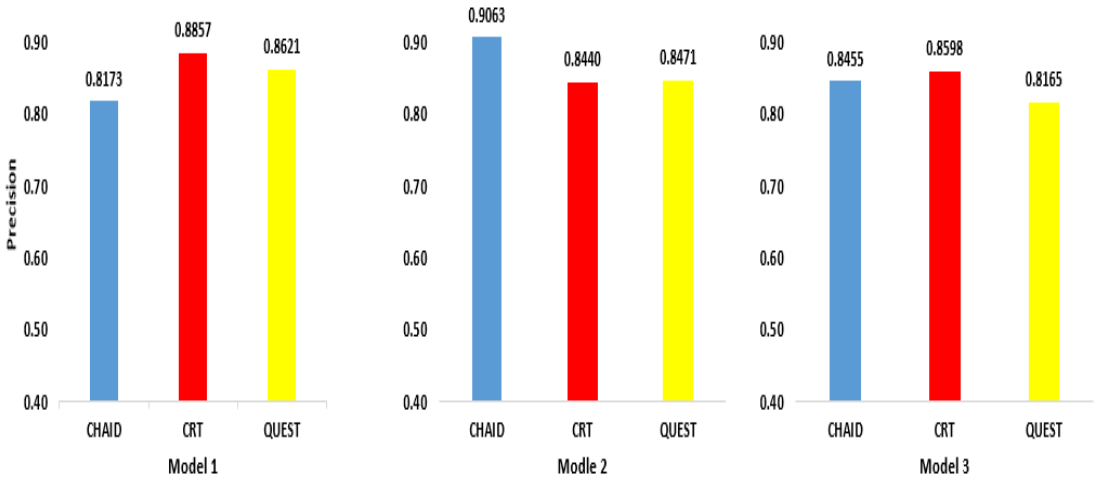

Figure 12. Precision test for the three decision tree models

Figure 13 shows the results of the recall test on the different decision tree models presented before.

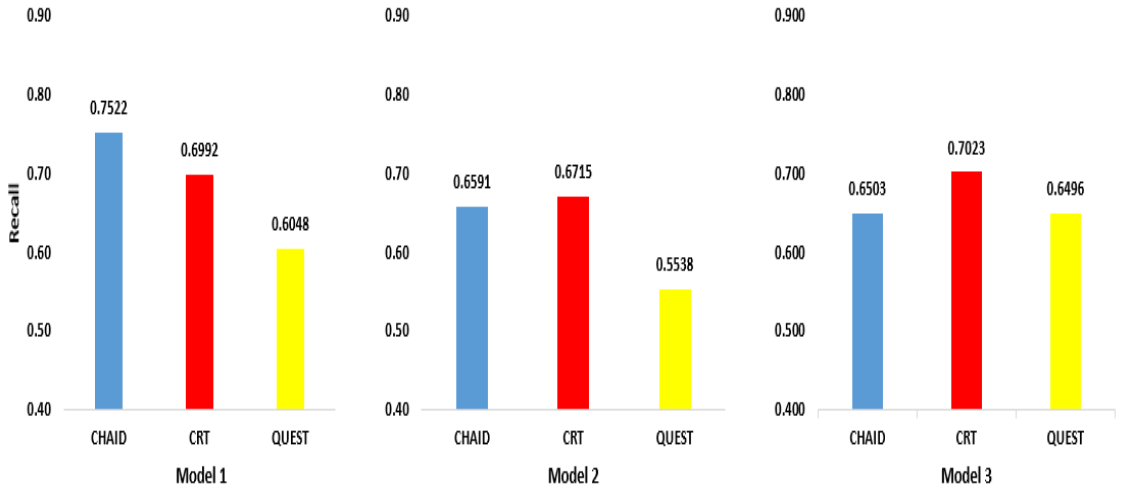

Figure 13. Recall measures of the different decision tree models.

In figure 13 we can observe that the recall measures for the first model achieved by CHAID method are higher than the recall results of the other models.

Figure 14 represents the results the F1-scores for those decision tree models.

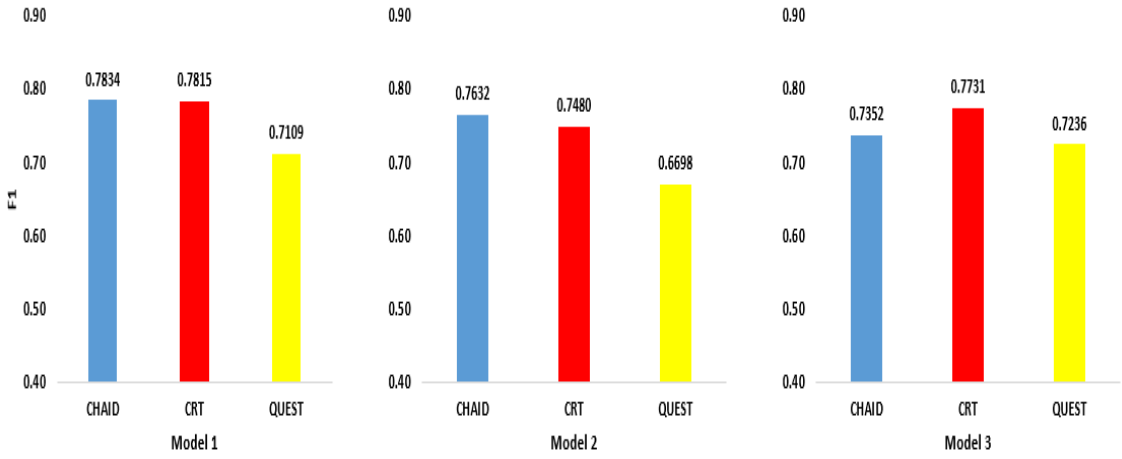

Figure 14. F1-score for the decision tree analysis models.

Figure 14 shows the F-scores for the first model are higher than for the other models.

Figure 15 shows there is no overfitting risk when we compare between the accuracy of the decision tree models for the training and test data.

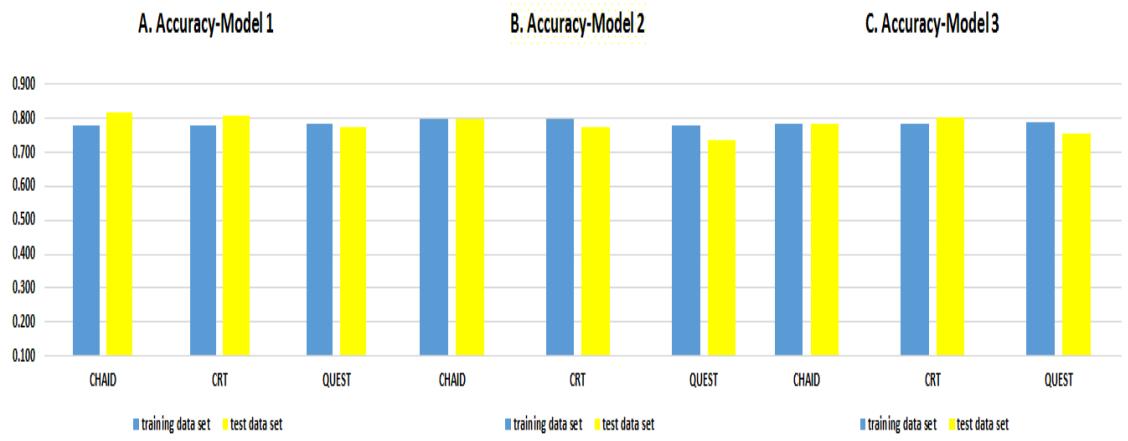

Figure 15. Accuracy of the decision tree models for the training and test data.

Section 6 discusses the conclusion , Managerial Consequences and findings.

6. Conclusion and future work

The major results of our different analyses enable us to answer the research questions that we have formulated.

The first research question read as follows :

What is the effect of some descriptive variables of the e-mail behaviour of customers on customer loyalty?

The effect of customer demographic variables and customer e-mail behavioural characteristics on customer loyalty was measured by decision tree analysis. Results indicated that nearly all those variables had a significant impact on

customer loyalty. The only variable not to have a significant impact proved to be the country of residence of the customer.

The second research question was :

Which model predicts customer loyalty best on the basis of both the customer demographic and customer e-mail behaviour variables ?

This question was answered by using decision tree analysis. The results of the different analysis algorithms and models were evaluated using the classification measures of accuracy, precision, recall and F-score.

The decision tree analysis used three different methods or procedures: CHAID, CART and QUEST. In all cases the decision tree analysis of a model based on demographic variables alone proved to be a little bit more accurate, precise and better on recall as the other models. The best results were mostly obtained using the CHAID method. Since the accuracy of all models (both separate and the combined models of variables) is mostly close to 80 % all models have a good predictive value.

Managerial Consequences and findings

In general, this paper proves that it is useful to analyse and examine the loyalty of customers based on their behaviour when receiving e-mail advertisements as a part of an e-mail marketing campaign. Results of our research based on the decision tree analysis show that it is indeed possible to predict customer loyalty based on response rates of e-mail campaigns thus allowing web master teams of an e-commerce company to group customers in different segments.

It is further important for all companies that use e-mail advertisements for their business to send these e-mails at the right time to the customers. Our study shows that the younger generation, female customers and customers who open the advertisements in the morning are more likely to be loyal customers. The Click through rate plays an important role in predicting customer loyalty, since the CTR and the high customer loyalty segment have a positive relationship .

So, If the company intends to increase the profit by increasing the response rates of the customers that have low click rates, it should send relevant advertisements to these customers based on their demographic characteristics to avoid customers unsubscribing from their website.

We recommend that the process of sending e-mails should consider the contents of the advertisement as well as the time customers usually open the e-mail campaigns in their mail. Figure 16 explains the different steps we recommend the company to use.

Figure. 16. The steps of the E-mail campaign process

The first step in this process is to know which contents of the e-mail campaign are relevant to the customer based on their demographic characteristics (age and gender) . Finally the web master team should consider several parts of day to send the e-mail to customers based on the right time at which customers normally open their e-mail advertisings.

Future work will have to address customers buying behaviour so that a company wanting to attract customers depending on their buying history by sending e-mails can be advised in more detail and depending on the products which customers like to buy.

Bibliography

Aziz Yarahmadi, Mathijs Creemers, Hamzah Qabbaah, Koen Vanhoof. 2017. 'UNRAVLING BI-LINGUAL MULTI-FEATURE BASED TEXT CLASSIFICATION: A CASE STUDY', International Journal "Information Theories and Applications, 24.

Breiman, L., J. Friedman, C.J. Stone, and R.A. Olshen. 1984. Classification and Regression Trees (Taylor & Francis).

Carmona, C. J., S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. del Jesus, and S. García. 2012. 'Web usage mining to improve the design of an e-commerce website: OrOliveSur.com', Expert Systems with Applications, 39: 11243-49.

Cases, Anne-Sophie, Christophe Fournier, Pierre-Louis Dubois, and John F. Tanner Jr. 2010. 'Web Site spill over to email campaigns: The role of privacy, trust and shoppers' attitudes', Journal of Business Research, 63: 993-99.

Chiu, Chui-Yu, Yi-Feng Chen, I. Ting Kuo, and He Chun Ku. 2009. 'An intelligent market segmentation system using k-means and particle swarm optimization', Expert Systems with Applications, 36: 4558-65.

Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava. 1999. 'Data Preparation for Mining World Wide Web Browsing Patterns', Knowledge and Information Systems, 1: 5-32.

De'ath, Glenn. 2000. 'Classification and regression trees: a powerful yet simple technique for ecological data analysis', Ecology, v. 81: pp. 3178-92-2000 v.81 no.11.

Díaz-Pérez, Flora M., and M. Bethencourt-Cejas. 2016. 'CHAID algorithm as an appropriate analytical method for tourism market segmentation', Journal of Destination Marketing & Management, 5: 275-82.

Etzioni, Oren. 1996. 'The World-Wide Web: quagmire or gold mine?', Commun. ACM, 39: 65-68.

F. Reichheld, Frederick, and Sasser Jr. W. E. 1990. Zero Defections: Quality Comes to Services.

G. T. Denison, D., B. K. Mallick, and A. F. M. Smith. 1998. A Bayesian CART algorithm.

George Sammour , Benoît Depaire , Koen Vanhoof and Geert Wets. 2009. "Identiying homogenous customer segments for risk email marketing experements." In 11th International Conference on Enterprise Information Systems, 89-94. milan , italy.

IBM. 2016. 'IBM SPSS Decision Trees 24.' in.

Kosala, Raymond, and Hendrik Blockeel. 2000. 'Web mining research: a survey', SIGKDD Explor. Newsl., 2: 1-15.

L., Steven, and Brewer Jr. 2012. 'AN EMPIRICAL COMPARISON OF LOGISTIC REGRESSION TO DECISION TREE INDUCTION IN THE PREDICTION OF INTIMATE PARTNER VIOLENCE REASSAULT', Indiana University of Pennsylvania.

Linoff, G.S., and M.J.A. Berry. 2011. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (Wiley).

Liu, Bing. 2006. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications) (Springer-Verlag New York, Inc.).

Liu, Yanbin, Ping Yuan, Wei Liu, and Xingsen Li. 2015. 'What Drives Click-Through Rates of Tourism Product Advertisements on Group Buying Websites?', Procedia Computer Science, 55: 221-30.

Loh, Wei-Yin, and Yu-Shan Shih. 1997. 'SPLIT SELECTION METHODS FOR CLASSIFICATION TREES', Statistica Sinica, 7: 815-40.

Long, William J., John L. Griffith, Harry P. Selker, and Ralph B. D'Agostino. 1993. 'A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain', Computers and Biomedical Research, 26: 74-97.

Lopes, Prajyoti, and Bidisha Roy. 2015. 'Dynamic Recommendation System Using Web Usage Mining for E-commerce Users', Procedia Computer Science, 45: 60-69.

Lovelock, C.H., and J. Wirtz. 2011. Services Marketing: People, Technology, Strategy (Prentice Hall).

Michael J. A. Berry, Gordon S. Linoff. 2004. Data mining techniques second edition – for marketing, sales, and customer relationship management (wiley: canada).

Ngai, E. W. T., Li Xiu, and D. C. K. Chau. 2009. 'Application of data mining techniques in customer relationship management: A literature review and classification', Expert Systems with Applications, 36: 2592-602.

Reichheld, F.F., and T. Teal. 1996. The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value (Harvard Business School Press).

Reinartz, Werner, and V. Kumar. 2002. The Mismanagement of Customer Loyalty.

Ripley, Brian D., and N. L. Hjort. 1995. Pattern Recognition and Neural Networks (Cambridge University Press).

Shan, Lili, Lei Lin, Chengjie Sun, and Xiaolong Wang. 2016. 'Predicting ad click-through rates via feature-based fully coupled interaction tensor factorization', Electronic Commerce Research and Applications, 16: 30-42.

Srivastava, Jaideep, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. 2000. 'Web usage mining: discovery and applications of usage patterns from Web data', SIGKDD Explor. Newsl., 1: 12-23.

Stokes, R. 2011. Emarketing: The Essential Guide to Digital Marketing (Porcupine Press).

Sumaiya Thaseen, Ikram, and Cherukuri Aswani Kumar. 'Intrusion detection model using fusion of chi-square feature selection and multi class SVM', Journal of King Saud University - Computer and Information Sciences.

T., Bowen John, and Chen Shiang-Lih. 2001. 'The relationship between customer loyalty and customer satisfaction', International Journal of Contemporary Hospitality Management, 13: 213-17.

Xuerui Wang, Wei Li, Ying Cui, Ruofei (Bruce) Zhang, Jianchang Mao. 2010. 'Click-Through Rate Estimation for Rare Events in Online Advertising, In : online multimedia advertising ', Techniques and Technologies: 1-12.

Zhang, Xi. 2015. 'Managing a Profitable Interactive Email Marketing Program: Modeling and Analysis', Georgia State University.

Authors' Information

Hamzah Qabbaah *- PhD student at Research group of business informatics, Faculty of Business Economics, Hasselt University, B2a, Campus Diepenbeek, B-3590 Diepenbeek, Limburg,Belgium*
e-mail: hamzah.qabbaah@uhasselt.be

Major Fields of Scientific Research: Data mining, digital marketing, Social Network Analysis, Knowledge Management

George Sammour *- Director of Quality Assurance and Accreditation Centre, Business Information Technology Department, Princess Sumaya University for Technology, Amman, Jordan.*

e-mail: George.Sammour@psut.edu.jo

Major Fields of Scientific Research: Data mining, Digital learning, lifelong learning, professional development

Koen Vanhoof *- Professor Dr., Head of the discipline group of Quantitative Methods, Faculty of Business Ecnomics, Universiteit Hasselt; Campus Diepenbeek; B-3590 Diepenbeek, Limburg,Belgium*
e-mail:koen.vanhoof@uhasselt.be
Major Fields of Scientific Research: data mining, knowledge retrieval

EXPERT MODELS OF VECTOR OPTIMIZATION

Albert Voronin, Yuriy Ziatdinov

Abstract. *An approach is proposed to solve the vector optimization problem of complex technical and economic systems in cases where there is insufficient (or lacking) information on the experimental and statistical data necessary for building regression models. To solve the problem under consideration, a multi-criteria optimization approach is undertaken using a non-linear compromise scheme. A model example is provided.*

Keywords: *vector optimization, regression models, expert estimation method, approximation polynomials, least squares method, nonlinear compromise scheme.*

ACM Classification Keywords: *H.1 Models and Principles – H.1.1 – Systems and Information Theory; H.4.2 – Types of Systems.*

Problem Content

When optimizing complex technical and economic systems we are often faced with the fact that for the construction of the necessary mathematical models is not enough experimental and statistical data. The situation is exacerbated in the case where the optimization is performed over several conflicting performance criteria.

In conditions of the acute shortage of experimental data, we propose to obtain necessary information ("quasi-experimental" data) from the experts – professionals who have sufficient experience in the design and operation of complex systems of this class.

A study is carried out on the example of vector optimization of space facilities objects by the generalized criterion of "reliability-value", but the results could

easily be used in other subject areas. By the term "reliability" we mean the probability of finding the defining parameters of all elements of the object within the permissible limits under the terms of operability.

Please note that in this case we are talking about designing a fundamentally new technology, for which technical and economic characteristics significantly different from those with which we have dealt earlier or developers are working today. One of the specific aspects is extreme limitation (and sometimes complete lack of it) of experimental and statistical data by which we could identify the mathematical models of reliability and cost.

In these hard formalizable conditions we have to resort to non-traditional approaches, one of which is considered in this section. Naturally, in this case, we can talk only about rough calculations, approximate determination of the main trends in the choice of factors affecting the reliability and value of developed space facilities.

Statement of the Problem

To solve the optimization problem, you must have the following starting data:

1. Mathematical models:

$$y_1^{'} = f_1(x);$$

$$y_2 = f_2(x),$$

where $y_1^{'}$ – the reliability of the space facility object (criterion to be maximized); y_2 – the cost of measures that affect the reliability (criterion to be minimized); f_1 and f_2 – some criteria functions; $x = \{x_i\}_{i=1}^{n}$ – *n*-dimensional vector of independent variables (optimization arguments).

2. Restrictions on independent variables $x \in X$, where

$$X = \{x \mid x_{i\min} \le x_i \le x_{i\max,}\, i \in [1,n]\}.$$

3. Restrictions on criteria $y \in M$, where $y = \{y_k\}_{k=1}^{s}$ – *s*-dimensional vector of the minimized nonnegative criteria (here *s*=2). We explain that as a single method of extremalization criteria in our task is selected minimization. To make the reliability criterion minimized too, we define

$$y_1 = 1 - y_1^{'}$$

(if one hundred percent reliability is expressed in the unit). Then

$$M = \{y \mid 0 \le y_k \le A_k, k \in [1;2]\}.$$

Limitations $x_{i\min}, x_{i\max}$ on the arguments $x \in X$ and A_k on criteria $y \in M$ are set for physical reasons.

If all this takes place, there are all prerequisites for the optimization of space objects on the criteria of reliability and cost, i.e., to determine a compromise-optimal parameters values $x^* = \{x^*_i\}_{i=1}^{n}$.

Method of Solution

In view of the apparent inconsistency of the criteria, it is necessary to resort to specific methods of multicriteria (vector) optimization theory. If you are using a method of scalar convolution, then a mathematical model for solving the problem of vector optimization is represented as

$$x^* = \arg\min_{x \in X} Y[(x)], \tag{12.1}$$

where $Y(y)$ – a scalar function, which has the meaning of the scalar convolution of partial criteria vector, which depends on the kind of compromise scheme. It should be ensured that its minimization leads to a Pareto optimal solution: $x^* \in X^K$. In (Voronin, A.N. at al 1999), a scalar convolution on nonlinear compromise scheme is proposed

$$Y[y(x)] = \sum_{k=1}^{s} A_k [A_k - y_k(x)]^{-1}, \tag{12.2}$$

where *s* – the dimension of the criteria vector. Convolution (12.2) makes it possible to obtain in formalized procedure the Pareto-optimal solutions, appropriate to given situations. For *s*=2 model (12.1) has the form

$$x^* = \arg\min_{x \in X} [\frac{A_1}{A_1 - y_1(x)} + \frac{A_2}{A_2 - y_2(x)}]. \tag{12.3}$$

The qualitative composition of the vector *x* is rather variegated and, consequently, the dimension *n* of the vector is generally high. Taking full account of the parameters *x* would lead to unnecessary complication of criteria functions f_1 and f_2, and to excessive difficulties in solving the optimization problem. Therefore, natural is selection only the most informative parameters *x* – coordinates of space in which optimization will be carried out of the criteria y_1

and y_2 , while the other parameters are assumed to be fixed and predetermined.

Selection will be done with the involvement of experts. They are introduced to the conditions of the problem, i.e., is called a developed particular type of space object (rocket or spacecraft), are described the terms of its design, manufacturing, testing and operation. Experts are asked to write down those activities that, in their view, may affect the reliability and cost of the space object at different stages of the product lifecycle.

This, for example, is the redundancy multipleness of control systems x_1, the values of stock on structural strength coefficient x_2 and power energy sources coefficient x_3; the relative volume of incoming inspection of materials and components x_4, the selection of production technologies and the relative volume of the control of their stability x_5, the amount of control-sample testing x_6; the volume of experimental testing of components and systems in all modes x_7, the amount of material incentives for staff x_8 and so forth. As a result of a special procedure (Voronin, A.N. at al 1999), is determined an adequate qualitative structure and dimension *n* of the vector of independent variables of criterion functions $f_1(x)$ and $f_2(x)$.

Kind of criterion functions depends on what information a researcher has to build the model. The range is wide – from a full knowledge of the mechanisms of phenomena (deterministic model) to the total uncertainty ("black box"). Between these two poles of the information is a probability level of uncertainty. Deterministic mathematical model of any characteristic of the space facility is extremely difficult to develop because of the complexity of the physical processes taking place and of the object reactions to complex of internal and external factors.

Let us consider, for example, the reliability criterion function $f_1(x)$ and approximate it on the set of arguments $x \in X$ by an approximating function $F_1(x,a)$ known to an accuracy of a vector of constants (coefficients) $a = \{a_j\}_{j=1}^{m}$.

When choosing the type of a function $F_1(x,a)$ you need to keep in mind the following. It is established (Voronin, A.N. at al 1999) that the best results are obtained if the regression model is based on some information known about the mechanisms of the studied phenomena. Then the model is called substantional. If such information is not available, then we have to work in the class of formal regression models, and to pay for the lack of information, a large amount of computation.

For a formal model are presented two contradictory requirements. On the one hand, the approximating function should be simple enough to computing processes are not proved to be too cumbersome. On the other hand, the approximating relationship must have sufficient accuracy and prognostic properties. In most practical cases, both these requirements are met in the class of the second order polynomial regression:

$$F_1(x,a) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i,j=1, i<j}^{n} a_{ij} x_i x_j, \qquad (12.4)$$

where a_0, a_1, a_{ij} – coefficients. Function (12.4) is well adapted to the topography of the objective function $f_1(x)$, it is capable of expressing such features as the ravine, and so on. In practice, are used different truncations of regression polynomial (12.4), mainly linear approximations.

Determining of coefficients *a* may be performed as by the interpolation methods and by the method of least squares (MLS). Interpolation formulas provide an exact match to approximating and objective functions in the reference points (interpolation nodes), the amount of which *N*, as well as the number of unknown constants *a*, is equal to *m*. The coefficients *a* are determined by solving a determine system of equations for reliability criterion

$$F_1(x^{(u)}, a) = f_1(x^{(u)}), u \in [1, N = m], \tag{12.5}$$

where $x^{(u)}$ – the interpolation nodes.

It is assumed that the values of the objective function in the approximation nodes $f_1(x^{(u)}), u \in [1, N]$ are known. The determination of these values with the help of experts is a key point in this study and is discussed below.

MLS includes *N* control points (approximation nodes), the number *N* may be less, greater than or equal to (as a special case) constants number *m*. The unknown coefficients approximating functions are determined from the condition

$$E(a) = \sum_{u=1}^{N} [F_1(x^{(u)}, a) - f_1(x^{(u)})]^2 = \min_a. \tag{12.6}$$

Using the necessary condition for a minimum of functions, we get called in the MLS theory a system of *normal* equations for reliability criterion

$$\frac{\partial E(a)}{\partial a_j} = 0, j \in [1, m], \tag{12.7}$$

whose solution determines the coefficients of the approximating function. Note that regardless of the number of selected reference points *N* the system of normal equations (12.7) is always determined.

For the cost criterion $f_2(x)$ are valid all the foregoing considerations, but instead of $a = \{a_j\}_{j=1}^{m}$ in expression of the approximating function $F_2(x,b)$ in the general case appears another vector of unknown constants $b = \{b_h\}_{h=1}^{p}$.

The specificity of the problem is that to get the values of the objective functions in reference points is very difficult. Indeed, even for the one single point, which corresponds to the established to date set of measures to ensure the reliability of the space object of this class, there is no a sufficient statistic for a confident assessment of reliability. It particularly relates to newly developed objects not having a long operating period. And it is quite illusory a possibility of objective evaluation of reliability for the other points of the region of existence of optimization arguments $x \in X$.

As always in cases when the task is difficult to formalize, one has to resort to the methods of expert estimates. Skilled specialist (expert), having sufficient experience in the design, manufacture and operation of the objects of this class can produce a **thought experiment** and imagine what will be the object reliability levels for various combinations of factors $x \in X$. Thus, the method is based on individual opinion (postulate), expressed by a qualified expert on the assessed value on the basis of his professional experience. The main drawback of the postulation is subjectivity and arbitrariness.

The procedure of expert assessments processing method allows reducing this disadvantage. The method consists in the fact that for quantitative assessment of some characteristic are used the postulates of not one but *several* persons competent in this matter. It is assumed that the "true" value of unknown to us quantitative characteristic is within the range of expert opinions and a "generalized" collective opinion is more reliable. In (Voronin, A.N. at al 1999) is proposed a procedure for processing the data of expert estimations, during which are obtained more accurate aggregated assessments, and (as a by-product) the coefficients of confidence to the opinion of individual experts.

Applying this method to the treatment of expert assessments of reliability and cost in each of the *N* nodal points in domain $x \in X$, we obtain two ratings vectors (quasi-experimental-data):

$$\{f_1(x^{(u)})\}_{u=1}^{N};$$
$$\{f_2(x^{(u)})\}_{u=1}^{N},$$

which serve as the basis for determining the vectors of the constants *a* and *b*, by condition (12.5), if applicable is interpolation method, or by the condition (12.6) - (12.7), if applicable is MLS. So are determined the mathematical regression models

$$y_1^{'} = f_1(x) \approx F_1(x);$$
$$y_2 = f_2(x) \approx F_2(x),$$

which are involved in the optimization procedure (12.3). Since the information about the objective functions at the nodal points of the approximation area is not obtained experimentally but by expert estimation, the models $F_1(x)$ and $F_2(x)$ are called *expert regression models*.

Let us separately discuss the problem of choosing the method of approximating criterion functions in the specified circumstances. At various truncations of a regression polynomial (12.4), the number of unknown constants, as a rule, exceeds the number of approximation nodes *N,* in which the expert can give his quite confident assessment of the value of a criterion function. Therefore, using the method of interpolation, we obtain underdetermine system of equations in which the number of equations is less than the number of the unknown constants.

To get out of this situation, it is necessary to apply the method of least squares, which in mathematics is seen as a way to solve underdetermine, overdetermine and determine (as a special case) equations systems. The solution can be both analytical (by (12.6) - (12.7)), and numerical, if the residual function $E(a)$ has a complex expression. In the latter case, a Levenberg-Marquardt algorithm is used.

Illustrative Example

The problem of optimizing the process of developing a perspective carrier rocket on the generalized criterion "reliability-cost" is considered. With the help of expert procedures (Voronin, A.N. at al 1999) are selected and fixed values of the factors that affect the reliability of the product, except for the following three (hereinafter all numeric data is arbitrary):

1. Multiplicity of reservation of missile control systems, x_1 is selected from a possible range from one to ten: $x_1 \in [1;10]$;

2. The safety factor in strength of design x_2 is in the range of one to five: $x_2 \in [1;5]$;

3. The coefficient of material incentives for developer company personnel x_3 may be selected from the range of one to six: $x_3 \in [1;6]$.

It is required, using inconsistent reliability $y_1^{'} = f_1(x)$ and cost $y_2 = f_2(x)$ criteria to evaluate the compromise-optimal values of these three factors: $x^* = \{x_i^*\}_{i=1}^3$.

A criteria reliability function is approximated by a linear approximation with a free term

$$y_1^{'} = f_1(x) \approx F_1(x) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3,$$

the number of unknown constants $m = 4$.

In determining the objective function values in the approximation nodes $f_1(x^{(u)}), u \in [1, N]$ we keep in mind that with sufficient confidence experts can estimate these values only in three (N = 3) points of the region $x \in X$:

$$x_1^{(1)} = x_{1\max} = 10; x_2^{(1)} = x_{2\max} = 5; x_3^{(1)} = x_{3\max} = 6;$$
$$x_1^{(2)} = x_{1\min} = 1; x_2^{(2)} = x_{2\min} = 1; x_3^{(2)} = x_{3\min} = 1;$$
$$x_1^{(3)} = x_{1\text{nom}} = 6; x_2^{(3)} = x_{2\text{nom}} = 3; x_3^{(3)} = x_{3\text{nom}} = 3.$$

The last point corresponds to the ideas of experts about the nominal values of reliability factors.

Let the experts gave the following their assessments of the product reliability levels in these approximation nodes (unit corresponds to 100 percent reliability):

$$f_1(x^{(1)}) = 0{,}8; f_1(x^{(2)}) = 0{,}6; f_1(x^{(3)}) = 0{,}7\,.$$

Note that in this problem $N < m$ and interpolation method can not be applied. MLS method involves minimizing on the unknown constants the function (6) of squared residuals, which has the form under specified conditions

$$E(a) = (a_0 + 10a_1 + 5a_2 + 6a_3 - 0{,}8)^2 + (a_0 + a_1 + a_2 + a_3 - 0{,}6)^2 + \\ + (a_0 + 6a_1 + 3a_2 + 3a_3 - 0{,}7)^2.$$

Application of the necessary condition of minimum for function (12.7) leads to the following determine system of normal equations:

$$\begin{aligned} &3a_0 + 17a_1 + 9a_2 + 10a_3 = 2{,}1; \\ &17a_0 + 137a_1 + 69a_2 + 79a_3 = 12{,}8; \\ &9a_0 + 69a_1 + 35a_2 + 40a_3 = 6{,}7; \\ &10a_0 + 79a_1 + 40a_2 + 46a_3 = 7{,}5. \end{aligned}$$

It is a system of linear algebraic equations (SLAE), for solution of which the standard computer software is developed. Successive elimination of variables by Gauss method is the basis of on-line program (http://www.webmath.ru/web/prog13_1.php.). Cramer's Formula for solving linear systems are used in the program (http://www.webmath.ru/web/prog12_1.php). Applying Gauss method for solving our system of normal equations, we obtain the values of the unknown constants:

$$a_0 = 0{,}46; a_1 = -0{,}06; a_2 = 0{,}25; a_3 = -0{,}06.$$

The minimized criterion on the reliability characteristic is determined by the formula

$$y_1(x) = 1 - y_1^{'}(x) \approx 0{,}54 + 0{,}06x_1 - 0{,}25x_2 + 0{,}06x_3$$

(is meant that the unit corresponds to 100 percent reliability of the product).

Similar calculations are made for the cost criterion. Criterion function of cost we approximate by linear approximation without constant term:

$$y_2(x) \approx F_2(x^{(u)}, b) = b_1x_1 + b_2x_2 + b_3x_3,$$

where b_1, b_2, b_3 – unknown constants (p=3). Evaluation levels experts gave in the same approximation nodes (N = 3), as in the case of reliability criterion. Let these estimates are as follows:

$$f_2(x^{(1)}) = 1; f_2(x^{(2)}) = 0{,}3; f_2(x^{(3)}) = 0{,}6$$

(Normalized value of cost at the maximum factors is equal to unity). Since in this case p=N, to determine the constants you can use an interpolation method. Equation (12.5) with respect to the cost criterion is converted to the form

$$F_2(x^{(u)}, b) = f_2(x^{(u)}), u \in [1, N = p],$$

and with given numerical data

$$10b_1 + 5b_2 + 6b_3 = 1;$$
$$b_1 + b_2 + b_3 = 0{,}3;$$
$$6b_1 + 3b_2 + 3b_3 = 0{,}6.$$

Solving this SLAE by Cramer (http://www.webmath.ru/web/prog13_1.php), we obtain

$$b_1 = -0{,}1; b_2 = 0{,}3; b_3 = 0{,}1$$

and the expression for the cost criterion is of the form

$$y_2(x) \approx -0{,}1x_1 + 0{,}3x_2 + 0{,}1x_3$$

We got analytical expressions for criteria of reliability and cost, allowing you to apply the formula (12.3) for the determination of a compromise-optimal values of factors *x** under the obvious limitations $A_1 = A_2 = 1$:

$$x^* = \arg\min_{x \in X}\left(\frac{1}{1 - 0.54 - 0.06x_1 + 0.25x_2 - 0{,}06x_3} + \frac{1}{1 + 0{,}1x_1 - 0{,}3x_2 - 0{,}1x_3}\right).$$

Having carried out minimization of the scalar convolution of criteria on the optimization arguments by Nelder-Mead method (Voronin, A.N. at al 1999), we obtain

$$x_1^* = 9{,}99; x_2^* = 3{,}69; x_3^* = 1{,}01.$$

This result of illustrative example can be interpreted as follows. When optimizing the design of the future launcher, special attention should be paid to the reservation of the product control systems x_1. Safety factor for structural strength to choose roughly a little more than mid range $x_2 \in [1;5]$. Do not place much hope in the possibility of material incentives for personnel x_3.

Thus, this study makes it possible to identify the main trends in the development and optimization of new space technology. The proposed procedures allowed practically solve the problem of vector optimization of compulsory insurance process for developed space facilities objects (Voronin, 2017).

We must be aware that expert models are less informative than regression models identified with the help of real experiments. However, firstly, expert assessments still reflect, albeit with a possible distortion, the reality, and, secondly, the expert model is only an initial approximation and with the accumulation of statistics is amenable to improvement.

Acknowledgements

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

[Voronin at al, 1999] Voronin A.N., Ziatdinov Ju.K., Kozlov A.I. Vector optimization of dynamic systems, Kiev, Tehnika.

[Voronin, 2017] Voronin A.N. Multi-Criteria Decision Making for the Management of Complex Systems, IGI Global.

Authors' Information

Voronin Albert Nikolaevich *– professor, Dr.Sci (Eng), professor of the department of National aviation university of Ukraine, Komarova Ave, 1, Kiev-58, 03058 Ukraine; e-mail: alnv@ukr.net*

Major Fields of Scientific Research: Multi-Criteria Decision Making; Man-Machine Complex Systems

Ziatdinov Yuriy Kashafovich – *professor, Dr.Sci (Eng), head of the department of National aviation university of Ukraine, Komarova Ave,1, Kiev-58, 03058 Ukraine; e-mail: oberst@nau.edu.ua*

DISCRETE TOMOGRAPHY MODEL FOR A DOE PROBLEM WITH LIMITED RESOURCES

Levon Aslanyan, Hasmik Sahakyan

***Abstract**: This work is aimed at understanding the applied value of the mathematical problem of discrete tomography. Tomography, in general, is about the reconstructing of objects by sets of observable properties. Theoretically this is a typical inverse problem of combinatorial analysis. In applied level, in addition to the well-known task of tests and testing from the electrical engineering, complementary tasks of biomedical nature are considered. An example of the second task which is about treating sets of biological linear specimens with combinations from a limited resource of drugs (antibiotics) is considered aiming at achieving as many different treatment courses as possible to take place.*

***Keywords**: discrete tomography, design of experiments.*

***ITHEA Keywords**: G.2.3 Applications, F2.2. Nonnumerical Algorithms and Problems*

1 Introduction

Problems where the object is given and it is necessary to calculate its characteristics - are called *direct/forward problems*, in contrast to the *inverse problems* in which the objectt is not available, and it is necessary to recover it based on a partial information, which is often given by measurements ([Heuberger, 2014], [DemangeMonnot, 2013].

Thus, inverse problems are a class of mathematical problems that arise when it is required to obtain information on internal or hidden data through external/available measurements. Inverse problems as a rule are ill-posed.

According to the Hadamar's definition [Hadamard, 1902], a problem should have the following three properties to be considered as "well-posed":

1. For all admissible data a solution exists,
2. For all admissible data the solution is unique,
3. The solution depends continuously on the data.

Problems that violate any of the three properties are "ill-posed". In the inverse problems the third condition is mainly violated, often referred to as a *stability* condition.

Tomography is a set of inverse problems about the reconstruction of an unknown object by means of partial data coming from its projections collected by means of X-rays, and taken along given directions. Typically, physical structures have a large variety of density values, and therefore a large number of X-rays are needed to recover the density distribution. In some cases, the object has a small number of density values (or the required number of directions, along which projections are taken is very limited in order to avoid physical damaging the objects) and thus a small number of X-rays is used. *Discrete tomography* is a domain to deal with these cases. In recent years, discrete tomography has been of research interest because of its mathematical formulation and the variety of applications. Discrete tomography is widely used, particularly in the processing of medical images. Applications are also related to the reconstruction of crystalline structures that are accessible only through some images provided by high-resolution transmission electron microscopy, and others ([SlumpGerbrands, 1982], [PrauseOnnasch, 1996], [IrvingJerrum, 1994], [Jinschek et al, 2004], [Kisieloski et al, 1995].

Discrete tomography, in the simplest case, considers an object T, which is a set of cells of the n-dimensional integer lattice Z^n. A projection of T in any direction calculates the number of points of T on the lines parallel to the projection direction. Given a set $\{d_1, d_2, \cdots, d_l\}$ of lattice directions and projections $P_1, P_2, \cdots, P_l$ along those directions. Consider Consistency and Reconstruction problems in Discrete Tomography.

Consistency: Does there exist a discrete set $T \in Z^n$ with given projections $P_1, P_2, \cdots, P_l$ in lattice directions $d_1, d_2, \cdots, d_l$?

Reconstruction: Construct a discrete object $T \in Z^n$ from its projections $P_1, P_2, \cdots, P_l$.

These are NP-hard problems for $n \geq 2$, and $l \geq 3$ non-parallel projections in the integer lattice Z^n ([GardGrizmPran, 1999]).

Due to the complexity of the problem, a special attention has been given to the 2-dimensional case. Subsets of Z^2 can be presented as binary images or binary matrices, where the 1s determine the cells of T. Various researches are devoted to the case of orthogonal projections: horizontal and vertical. In terms of binary matrices, the row sum corresponds to the horizontal projection of T, and the column sum corresponds to the vertical projection. In the case with only horizontal and vertical projections the problem has polynomial complexity, but the number of solutions can be large. Any prior knowledge /constraint/ about the image to be reconstructed, can reduce the search space of possible solutions. The existence problem under different geometrical constraints /convexity, connectivity/ is investigated by various authors (R.Gardner, P.Gritzmann, A.Del Lungo, E.Barcucci, M.Nivat, R.Pinzani, G.Woeginger, and others, [GardnerGritzmann, 1999], [Barcucci et al, 1996], [Woeginger, 2001]); for some cases the NP-completeness is proved, some particular cases can be solved by polynomial algorithms, but there also exist open problems in terms of complexity.

Summarizing, discrete tomography consistency and reconstruction problems can be presented through the model of weighted binary matrices (rows' weights correspond to the horizontal projection and columns' weights correspond to the vertical projections). The matrix model brings its own constraints, and one of them is the *requirement of non-repetitiveness of the matrix rows*. This constraint naturally appears in a number of applications; one of them is the design of experiments (DOE).

In this paper we consider a DOE problem with limited resources with examples from the biomedical field.

Firstly we compare the binary matrix model of this problem with a known mathematical problem of minimum test collections (MTC) on binary tables ([YablonskiiChegis, 1955], [ChegisYablonskii, 1958], [Solov'ev, 1978], [Dmitriev

et al, 1966]). MTC is one of the basic NP-complete problems ([Karp, 1972], [GareyJonson, 1979]). The basic reference to MTC is "unpublished papers" by M. Garey and D. Jonson (see [GareyJonson, 1979]) but much earlier S. Yablonskii and I. Chegis investigated the problem in detail. An effective machine learning application was the work [Zhuravlev et al, 1966] by Yu. Zhuravlev et al. The input of the problem is a (0,1) table of a given size $m \times n$ with different rows. It is necessary to find subsets and, if possible, minimal size subsets of columns by which the rows still remind different. Usually the interpretation of this problem is given in terms of the electrical engineering. There are observable characteristics of electrical equipment (qualitative or quantitative), corresponding to the columns of the matrix. Matrix rows correspond to individual equipment that are physically in various malfunctioning states. The problem examines at first stage such sets of observable characteristics, by which the given m states (rows) differ one from the other (forming the initial input table), and then, the goal is to minimize the sets of columns by removing from it one or another group of observations keeping the rows different.

What is the similarity and difference between the two considered problems – the minimum test sets and the planning of a large number of different experiments with limited resources. The first task is a real optimization problem about physical objects and their properties. The second task relates to virtual reality - it asks about the reality of the plan of experiments, which are many, and which are based on the framework of the available resources.

Consider another comparison of the two mentioned problems on matrices with different rows. In MTC the matrix is given so that all rows and all columns are given beforehand. We just try to find proper row fragments composed by minimal number of columns keeping all rows different. In DOE with resource limitation what is given is the column weight vector. The general idea is in organizing as much as possible different and informative experiments, with an effective use of given limited resource of the problem.

Summarizing, the use of the binary matrix model for the considered DOE problem with limited resources makes it possible to apply known approaches

and algorithms developed for the construction of binary matrices with given projections and given constraints.

As an example, a greedy algorithm, developed in [Sahakyan, 2010], [SahakyanAslanyan, 2011] is adjusted/modified and applied in the design of biomedical experiments with limited resources.

The paper is organized as follows: below in Chapter 2 the DOE problem is introduced and modeled in terms of binary matrices. Chapter 3 brings a brief description of the greedy approximation algorithm developed in [Sahakyan, 2010], [SahakyanAslanyan, 2011] for constructing binary matrices with different rows. Chapter 4 introduces some modifications of the algorithm.

2 Problem Definition

Design of experiments (DOE) is a research domain ([Fisher], [Bose, 1939], [Rao, 1996], which helps in assigning treatments to the experimental units in a way of optimizing several characteristics of experiments such as the evaluated output value dispersion, or the computational complexity, etc. Combinatorial block design ([BhattacharyaSinghi, 2013], [Beth et al, 1985], [ColbournDinitz, 2006], one of the constituents of the DOE theory, combines units into homogeneous groups to achieve the goal of the DOE.

Consider the following medical-biological DOE problem.

Multidrug resistance problem is well known in biomedicine. According to the WHO (World Health Organization), most pathogenic species in existence today have developed resistance to one or more antimicrobials. E.g., the multidrug resistance of Escherichia coli (an essential component of the digestive tract flora of healthy humans and even most animals) has increased from 7.2% during the 1950s to 63.6% during the 2000s. One of the ways of combating multidrug resistance is the use of antibiotic combinations (cocktail) [Hickman, 2011].

Suppose that for a biomedical experiment, performed to gain a practical knowledge of antibiotic combinations, there are n antibiotics of different type

given in limited quantities/portions. A cocktail combines portions of several of the given n antibiotics. Let, for the i-th antibiotic s_i portions only of the drug are given ($i = 1, \cdots, n$).

The problem is the following: design/plan experiments, which means creation of given number of different cocktails so that they use the whole available drug store $s_1, s_2, \cdots, s_n$.

It is possible to describe very large number of similar DOE scenarios. From technical point of view it is important to mention that the resources $s_1, s_2, \cdots, s_n$ are for single use. If they describe not drugs but bacteria, then these are not cultivated.

In addition to this it is specifically important to mention that the non-repetition of experiments is a natural and valid requirement of the experimental design in these cases.

Let us also mention that each cocktail has its experimental value and that we do not consider general or economic optimization issues.

The scope of the problem solutions can be wide, but we are interested in:

- Finding/composing one of the solutions, because it clarifies the possibility of planning experiments with given quantities of antibiotics (otherwise, the composition of available samples should be changed);
- Composing a solution, which involves as much as possible different cocktails with given quantities of antibiotics.

Thus, the problem can be formulated in the following way:

Antibiotics_Cocktail (***A_C***): Given n antibiotics of limited $s_1, s_2, \cdots, s_n$ quantities, correspondingly.

(1) Decide whether it is possible to design an experiment with the given number m of different cocktails (subset, combination) so that $s_1, s_2, \cdots, s_n$ quantities are used;
(2) Compose as much as possible different cocktails using $s_1, s_2, \cdots, s_n$ quantities of antibiotics.

Each cocktail can be presented by a binary vector of length n such that the j-th component of the vector is 1 if and only if the j-th antibiotic is used in the cocktail. In this manner an experiment E with m different cocktails corresponds to a binary matrix $A = \{a_{i,j}\}$ with m rows and n columns; the number of 1s in the j-th column of A is equal to the quantity of the j-th antibiotic used in the experiment; the number of 1s in the i-th row of A equals the number of antibiotics used in the i-th cocktail.

The problem in terms of binary matrices has the following formulation.

Existence of binary matrices with given column sum and with different rows ($\boldsymbol{CS\_D}$): Given non-negative integer vector $S = (s_1, s_2, \cdots, s_n)$ and a natural number m.

(1) Decide whether there is a binary matrix $A = \{a_{i,j}\}$ of size $m \times n$ with all different rows and with the column sum vector $S = (s_1, s_2, \cdots, s_n)$;
(2) Compose a binary matrix with the column sum $S = (s_1, s_2, \cdots, s_n)$ and with maximum possible number of different rows.

Thus $\boldsymbol{A\_C}$ is reduced to the combinatorial problem $\boldsymbol{CS\_D}$, which is known as a hard computational problem.

In the process of seeking efficient approximate solutions a greedy heuristics is investigated and an algorithm is developed in [Sahakyan, 2010], [SahakyanAslanyan, 2011] for constructing binary matrices with given column sums and with different rows. In the next sections we briefly introduce the algorithm, and also mention some peculiarities of the $\boldsymbol{Antibiotics\_Cocktail}$ problem and the algorithm itself, which makes it expedient using the algorithm for this problem.

3 Optimization Version and Approximation Algorithm

For a given non-negative integer vector $S = (s_1, s_2, \cdots, s_n)$ let $U(S)$ denote the class of binary matrices of size $m \times n$, which have the column sum vector $S = (s_1, s_2, \cdots, s_n)$.

For a matrix A of $U(S)$ let $DP(A)$ denote the number of disjoint pairs of rows of A; Obviously, $0 \leq DP(A) \leq C_m^2$, and $DP(A) = C_m^2$ if and only if the rows of A are different.

Now we consider the following optimization version of $\boldsymbol{CS\_D}$ (1) with the objective function DP.

($\boldsymbol{CS\_D^{opt}}$): find $A_{opt} \in U(S)$ such that $DP(A_{opt}) = max_{A \in U(S)} DP(A)$.

It is clear that any solution of (CS_D^{opt}) is also a solution for CS_D for the cases when $U(S)$ contains also matrices with different rows.

Now we bring a brief description of the algorithm G introduced in [SahakyanAslanyan, 2011], [Sahakyan, 2010].

Algorithm $\boldsymbol{G}$ for the problem $\boldsymbol{CS\_D^{opt}}$

Input: non-negative integer vector $S = (s_1, \cdots, s_n)$ and natural number m.

The algorithm G constructs a binary matrix $A_{G,n}$ of size $m \times n$ in the column-by-column manner, putting s_i 1s in the i-th column ($i = 1, \cdots, n$) having the goal to maximize the increase of the objective function (the number of disjoint pairs of rows) in each step.

Step 1. Construction of the first column: put 1s in the first s_1 rows and put 0s in the remaining $m - s_1$ rows. In the first column we get an interval of 1s of length s_1 (i.e. s_1 consecutive 1s) and an interval of 0s of length $m - s_1$ ($m - s_1$ consecutive 0s). Then, the number of different pairs of rows after the construction of the first column will be equal to $s_1 \cdot (m - s_1)$. We notice that any other distribution of s_1 1s and $m - s_1$ 0s in the first column would produce the same number of different pairs of rows.

In each next step the algorithm splits every interval of the previous step (column) into two parts, and puts 1s in one of them, and 0s in the other, such that the summary number of 1s equals the corresponding component of the column sum vector $S = (s_1, \cdots, s_n)$.

Suppose that the first $(k-1)$ columns of the matrix are constructed, and let the $(k-1)$-th column consists of p intervals of non-zero lengths, denoted by: $d^G_{k-1,1}, d^G_{k-1,2}, \cdots, d^G_{k-1,p}$.

Step k. Construction of the k-th column: for $i = 1, \cdots, p$, split the $d^G_{k-1,i}$-length interval into two parts, -denoting them by $d^G_{k-1,i,0}$ and $d^G_{k-1,i,1}$, - and put 0s and 1s respectively, such that:

$\sum_{i=1}^{p} d^G_{k-1,i,0} = m - s_k$, $\sum_{i=1}^{p} d^G_{k-1,i,1} = s_k$.

Then the increase of the objective function will be equal to: and $\sum_{i=1}^{p} d^G_{k-1,i,1}$ · $d^G_{k-1,i,0}$.

All rows of the matrix will be different if and only if the last column of the matrix consists of only one-length intervals.

The detailed description of the algorithm G and the proof of its local optimality (the maximal increase of the objective function is achieved in each step) is given in [Sahakyan, 2010]; and its performance guarantee is estimated in [Sahakyan, 2017]. On the other hand, experimental results given in [SahakyanAslanyan, 2011] (algorithm run for all binary matrices with $n \leq 6$ columns and with m, $m = 1,2, \cdots, 2^n$ rows) show that the constructed matrices in the last column can have at most 2-length intervals; moreover in most cases there is only one 2-length interval in the last column (and for small values of m ($m \leq 11$), all rows are different), which means that constructed matrices in most cases contain at most 1 pair of coinciding rows.

4 Modified Algorithm for the Problem "Antibiotics Cocktail"

In the experiments, which are related to the design of antibiotic combinations the general idea is in organizing as much as possible different and informative experiments (*), with an effective use (**) of the given limited resource of the problem. Combining (*) and (**) we define 2 simple strategies. The first is direct maximization of the number of experiments; and the second, focused on effective use of the problem resources, can be formulated as high column weights, i.e. every antibiotic is used at least in certain part of antibiotic combinations. In "Antibiotics Cocktail" problem we will assume that $s_i \geq \frac{m}{2}, i = 1, \cdots, n$, that is every antibiotic is used at least in the half of antibiotic combinations.

Now we apply algorithm G to $\boldsymbol{Antibiotics\_Cocktail}$ ($\boldsymbol{A\_C}$) problem:

(1) Decide whether it is possible to design an experiment with given number of different cocktails such that $s_1, s_2, \cdots, s_n$ quantities of antibiotics are used.

It is worth mentioning that with the supposition $s_i \geq \frac{m}{2}$, the algorithm G can organize the splitting of intervals (in each column of the constructed matrix $A_{G,n}$) in such a way that the resulting matrix contains a row consisting of all 1s. It follows that in the case when $A_{G,n}$ contains coinciding rows, at least two of them will be consisting of all 1s. Thus, $A_{G,n}$ will contain either:

a) at most two coinciding rows (coinciding combinations of antibiotics, where each of them contains all types of antibiotics); or
b) more than two coinciding rows/combinations.

In the Case a) if all m rows of the matrix $A_{G,n}$ are different, then $A_{G,n}$ is the required solution. Otherwise, $A_{G,n}$ contains one pair of coinciding rows consisting of all 1s (this can be either due to the non-optimality of the algorithm G, or non-possibility to design m different experiments with $s_1, s_2, \cdots, s_n$

quantities. We remove one of the coinciding rows, and output the matrix with $m-1$ rows and with the adjusted column sum vector $(s_1-1,\cdots,s_n-1)$ (one amount of every antibiotic will remain unused).

In the Case b) we may assume that it is not possible to design m different experiments with $s_1,s_2,\cdots,s_n$ quantities; however the constructed matrix $A_{G,n}$ provides a guaranteed number of pairs of different combinations of antibiotics (related to the optimal number).

Before considering part 2 of the problem, we formulate and prove the following lemma.

Lemma

Let A be a binary matrix of size $m \times n$ with all different rows and with the column sum vector $S=(s_1,s_2,\cdots,s_n)$, where $s_i \geq m/2$ for $i=1,\cdots,n$, and $s_j > m/2$ for some j, $1 \leq j \leq n$. Then, there exists a binary matrix of size $(m+1)\times n$ with all different rows and with the same column sum vector $S=(s_1,s_2,\cdots,s_n)$.

Proof.

$s_j > m/2$ implies that A contains a row (let it be the i-th row) with 1 in the j-th position such that A does not contain the row differing from the i-th only by the j-th position, i.e. $(a_{i,1},\cdots,a_{i,j-1},1,a_{i,j+1},\cdots,a_{i,n}) \in A$, and $(a_{i,1},\cdots,a_{i,j-1},0,a_{i,j+1},\cdots,a_{i,n}) \notin A$. We append $(a_{i,1},\cdots,a_{i,j-1},0,a_{i,j+1},\cdots,a_{i,n})$ to the matrix A (it will not cause row repetitions). The resulting matrix will have $m+1$ different rows and a column sum vector $(s'_1,s'_2,\cdots,s'_{j-1},s_j,s'_{j+1},\cdots,s'_n)$, where $s'_k \geq s_k$ for $k=1,\cdots,n, k \neq j$.

Now we will modify A into a binary matrix A' of size $(m+1)\times n$ with all different rows and with the column sum vector $S=(s_1,s_2,\cdots,s_n)$.

Suppose that $s'_k > s_k$ for some k (in fact, $s'_k = s_k+1$), then $s'_k > m/2$. It follows that there exists a row in A (let it be the t-th row) with 1 in the j-th position:

$(a_{t,1},\cdots,a_{t,k-1},1,a_{t,k+1},\cdots,a_{t,n}) \in A$,

such that $(a_{t,1}, \cdots, a_{t,k-1}, 0, a_{t,k+1}, \cdots, a_{t,n}) \notin A$. We replace $(a_{t,1}, \cdots, a_{t,k-1}, 1, a_{t,k+1}, \cdots, a_{t,n})$ with $(a_{t,1}, \cdots, a_{t,k-1}, 0, a_{t,k+1}, \cdots, a_{t,n})$; this will decrease s'_k by 1 and will not cause row repetitions.

By the same reasoning we make relevant row replacements for all $s'_k > s_k$. □

Antibiotics_Cocktail (***A_C***) problem (part 2):

(2) Suppose that it is possible to design m different cocktails with $s_1, s_2, \cdots, s_n$ quantities of antibiotics; compose maximum possible number of different cocktails using the same quantities of antibiotics.

Firstly, we apply algorithm G and get as a result a binary matrix A (possibly with a small number repeated rows, which are further removed, and the column sum vector and the number of rows are adjusted) of size $m \times n$ with all different rows and with the column sum vector $S = (s_1, s_2, \cdots, s_n)$.

Now we introduce an algorithm M that increases the number of rows keeping the same column sum vector S.

Algorithm *M*.

Input: matrix A of size $m \times n$ with all different rows and with the column sum vector $S = (s_1, s_2, \cdots, s_n)$;

$A' := A$; $m' := m$; $S' := S$;

While (A' satisfies the Lemma conditions (all $s'_i \geq m'/2$ and $s'_j > m'/2$ for some j))

{

find a row (according to the Lemma), and append it to A'; $m' := m' + 1$;

calculate new column sum vector $S' = (s'_1, s'_2, \cdots, s_n)$ of A';

While ($s'_k > s_k$ for some k)

{

find and make appropriate row replacements in A' (according to the proof of the Lemma),

calculate new column sum vector $S' = (s'_1, s'_2, \cdots, s_n)$ of A';

}

}.

Output: the matrix A' of size $m' \times n$ with all different rows and with the column sum vector $S = (s_1, s_2, \cdots, s_n)$, where $m' > m$.

Conclusion

A DOE problem with limited resources from the biomedical field is modeled by binary matrices as a discrete tomography problem with the constraint of non-repetitive rows. The problem is hard computationally, and in the process of seeking efficient approximate solutions, a known greedy algorithm developed for the construction of binary matrices with given projections and with all different rows, is adjusted/modified and applied to solve the problem.

Acknowledgements

This work is partially supported by the grants № 18RF-144, and № 18T-1B407 of the Science Committee of the Ministry of Education and Science of Armenia.

Bibliography

[Heuberger, 2004] Heuberger C., Inverse combinatorial optimization: a survey on problems, methods and results, Journal of Combinatorial Optimization, vol. 8, pp. 329–361, 2004.

[DemangeMonnot, 2013] Demange M. and Monnot J., An introduction to inverse combinatorial problems, Paradigms of Combinatorial Optimization:

Problems and New Approaches, John Wiley & Sons Inc., Hoboken, NJ, USA, vol. 2, 2013. DOI: 10.1002/9781118600207.ch17.

[Hadamard, 1902] Jaques Hadamard (1902), Sur les problèmes aux dérivées partielles et leur signification physique. Princeton University Bulletin, 49-52.

[SlumpGerbrands, 1982] Slump, C.H., Gerbrands, J.J., A network flow approach to reconstruction of the left ventricle from two projections, Comput. Gr. Im. Proc., 18, 18.36 (1982).

[PrauseOnnasch, 1996] Prause G.P.M. and Onnasch D.G.W., Binary reconstruction of the heart chambers from biplane angiographic image sequence, IEEE Transactions Medical Imaging, 15 (1996) 532-559.

[IrvingJerrum, 1994] Irving R., Jerrum M., Three-dimensional statistical data security problem, SIAM J. Comput. 23 (1994), pp.170-84.

[Jinschek et al, 2004] Jinschek J., Batenburg K., Calderon H., Van Dyck D., Chen F., Kisielowski C., Prospects for bright field and dark field electron tomography on a discrete grid, Microscopy Microanal. 10 (Suppl. 3) (2004), pp.44-45. Cambridge Journals Online.

[Kisieloski et al, 1995] Kisieloski C., Schwander P., Baumann F., Seibt M., Kim Y., and Ourmazd A., An approach to quantitative high-resolution transmission electron microscopy of crystalline materials, Ultramicroscopy 58 (1995), pp.131-155.

[GardnerGritzmann, 1999] Gardner R. J., Gritzmann P., Prangenberg D., On the computational complexity of reconstructing lattice sets from their X-rays, Discrete Mathematics 202 (1999) 45-71.

[Barcucci et al, 1996] Barcucci E., Del Lungo A., Nivat M., and Pinzani R.: Reconstructing convex polyominoes from horizontal and vertical projections. Theor. Comput. Sci. 155, 321–347 (1996).

[Woeginger, 2001] Woeginger G.J., The reconstruction of polyominoes from their orthogonal projections, Inform. Process. Lett., 77, pp. 225-229, 2001.

[ColbournDinitz, 2006] Colbourn C. J., Dinitz J. H., Handbook of Combinatorial Designs, second ed., Chapman and Hall/CRC, 2006.

[Sahakyan, 2010] Sahakyan H., Approximation greedy algorithm for reconstructing of (0.1)-matrices with different rows, Information Theories and Applications, Volume 17, Number 2, 2010, pp. 124-137.

[SahakyanAslanyan, 2011] Sahakyan H., Aslanyan L., Evaluation of Greedy algorithm of constructing (0,1)-matrices with different rows, Information Technologies and Knowledge, Volume 5, Number 1, 2011, pp. 55-66.

[Sahakyan, 2017] Sahakyan H.A., Hypergraph degree sequence approximation, "NAS RA Reports", Volume 117, Number 1, 2017, pp.26-34.

[Fisher] Fisher, Ronald A., The Design of Experiments (9th ed.). Macmillan. ISBN 0-02-844690-9.

[Bose, 1939] Bose R. C., On the construction of balanced incomplete block designs, Annals of Eugenics. 9 (1939), 358–399.

[Rao, 1996] Rao C. R. Handbook of Statistics 13: Design and Analysis of Experiments, 1996 (Ed. with S. Ghosh) North Holland.

[BhattacharyaSinghi, 2013] Bhattacharya A., Singhi N. M., Some approaches for solving the general (t,k)-design existence problem and other related problems, Discrete Applied Mathematics, 161, pp. 1180–1186, 2013.

[Beth et al, 1985] Beth T., Jungnickel D., Lenz H., Design Theory, Wissenshiaftoverlag Mahheim, Wien/Zurich, 1985.

[Hickman, 2011] Hickman R., Could antibiotic cocktails be the answer? An in vitro study into antibiotic combinations with enhanced activity against Escherichia coli, Biology Education Centre and Department of Cell and Molecular Biology, Uppsala University, 2011.

[YablonskiiChegis, 1955] Yablonskii S.V., Chegis I.A., On tests for electric circuits, Uspekhi Mat. Nauk , 10 : 4 (1955) pp. 182–184 (In Russian).

[ChegisYablonskii, 1958] Chegis I.A., Yablonskii S.V., Logical methods of control of work of electric schemes, Trudy Steklov Math. Inst., 51, pp. 270–360, 1958 (in Russian).

[Solov'ev, 1978] Solov'ev N.A., Tests, Novosibirsk (1978) (In Russian).

[Zhuravlev et al, 1966] Dmitriev A.N., Zhuravlev Yu.I., Krendelev F.P., On mathematical principles of classification of objects and events, Diskretn. Anal., 7 (1966) pp. 3–15 (In Russian).

[Karp, 1972] Karp R., Reducibility among Combinatorial Problems, Complexity of Computer Computations, pp. 85–103, 1972.

[GareyJonson, 1979] D.,Garey M., Jonson D., Computers and Intractability, A Guide to the Theory of NP Completeness, W. H. Freeman and company, 1st edition, New York, 340 pages, 1979.

Authors' Information

Levon Aslanyan *– Institute for Informatics and Automation Problems of the National Academy of Sciences of Armenia, head of department; 1 P.Sevak str., Yerevan 0014, Armenia; e-mail: lasl@sci.am*

Major Fields of Scientific Research: Discrete analysis – algorithms and optimization, pattern recognition theory, information technologies.

Hasmik Sahakyan *– Institute for Informatics and Automation Problems of the National Academy of Sciences of Armenia, scientific secretary; 1 P.Sevak str., Yerevan 0014, Armenia; e-mail: hsahakyan@sci.am*

Major Fields of Scientific Research: Combinatorics, Discrete Tomography, Data Mining.

INTENSIVE CARE UNIT (ICU) DATA ANALYTICS USING MACHINE LEARNING TECHNIQUES

Zeina Rayan, Marco Alfonse, Abdel-Badeeh M. Salem

***Abstract**: The huge amount of clinical signals and measurements in Intensive Care Units (ICU) will simply overwhelm the person responsible for healthcare support and may cause treatment delays, or clinical errors. This paper analyzes the recent Machine Learning (ML) and Computational Intelligence (CI) techniques which are applied to ICU equipment data in the area of smart health during the period from 2011 till 2018. The results show that ML and CI techniques are used in many ICU applications such as mortality prediction, ICU readmission prediction, prediction of sepsis, prediction of complications in ICU, and processing and monitoring vital signs in ICU patients.*

***Keywords**: Smart Healthcare, Computational Intelligence, Intensive Care Unit, Medical Informatics, Machine Learning.*

***ACM Classification Keywords**: 1.2 Artificial intelligence, F1.1 Models of computation, H.4.2 Types of systems Decision Support.*

Introduction

Towards smart health [Solanas et al, 2014], hospitals infrastructure is now being rebuild, this infrastructure should make use of Information and Communication Technology (ICT). The ICU is designed to care for patients who are seriously injured, have a critical or life-threatening illness, or have undergone a major surgical procedure, therefore requiring 24-hour care and monitoring. The ICUs have been evolved by the evolution of ICT, where an ICU equipment has many devices and sensitive connections between the patient and these devices which leads to better patient monitoring. The ICU equipment includes many devices responsible for various tasks such as patient monitoring,

respiratory and cardiac support, emergency resuscitative, pain management, and other life support equipments for people with serious injury. The ICU equipment provides data such as electrical activity of heart, respiratory rate (breathing), blood pressure, body temperature, amount of oxygen and carbon dioxide in the blood, intracranial pressure (pressure of fluid in the brain) … etc. [Johnson et al, 2016].

The continuous monitoring of ICU patients has resulted in enormous amount of data, which leads to many opportunities and challenges [Johnson et al, 2016]. This huge amount of data can be utilized to develop optimal machine learning models that can be used in treatment, diagnosis, and discharging of ICU patients. However, the storage, analysis, privacy, integration, and harmony of these data are considered challenges for handling the data (data analysis and processing).

This paper is constructed as follows: section 2 presents the different machine learning techniques which were applied to ICU data, section 3 presents the comparative study of these techniques and finally section 4 includes the conclusions and future work.

Analysis of recent research ML and CI used in ICU data analytics

This section presents a comparative study between many different machine learning and computational intelligence techniques that are applied to ICU data.

Che, et. al. [Che et al, 2018] proposed a model (GRU_D) which is based on Gated Recurrent Unit (GRU) for mortality prediction using the Multi-Parameter Intelligent Monitoring for Intensive Care, version 3 (MIMIC-III) which is a public research archive collected from different ICU patents [Saeed et al, 2011]. The Proposed model achieved Area Under Curve (AUC) of 0.8527±0.003. In addition, they used the PhysioNet Dataset to validate the proposed model and achieved AUC of 0.8424±0.012.

Nemati, et. al. [Nemati et al, 2018] proposed a model for the prediction of sepsis based on a modified Weibull-Cox proportional hazards model and their dataset

was collected from two hospitals within the Emory healthcare system, and other publicly available ICU database. The proposed model achieved 70% accuracy using 4-Hours data of monitoring, 68% accuracy using 6-Hours data of monitoring, 67% accuracy using 8-Hours data of monitoring and 65% accuracy using 12-Hours data of monitoring.

Meyer, et. al. [Meyer el al, 2018] proposed a deep learning model (recurrent deep neural network) for real time prediction of complications in ICU, they used dataset collected from German tertiary care center (German Heart Center Berlin) for cardiovascular diseases. The proposed model achieved accuracy of 86%, then this work was validated using MIMIC-III dataset where the proposed model achieved accuracy of 75%.

Anand, et. al. [Anand et al, 2018] proposed a model for prediction of mortality in diabetic ICU Patients, they used the diabetic inpatients of the MIMIC-III dataset. The proposed model achieved accuracy of 94%.

Viegas, et. al. [Viegas et al, 2017] proposed an ensemble model to predict readmissions of ICU utilizing feature selection and fuzzy modeling approaches. They used MIMIC-II database. The proposed model gives AUC of 0.79.

Desautels, et. al. [Desautels et al, 2016] proposed a model (insight classifier) for prediction of Sepsis in ICU utilizing the MIMIC-III as a dataset. The proposed model achieved accuracy of 80%.

Pirracchio, et. al. [Pirracchio et al, 2015] proposed a model, Super ICU Learner Algorithm (SICULA), for mortality prediction. They trained the model using the MIMIC-III dataset, and The cross-validated area under the receiver operating characteristic curve (AUROC) for the proposed model was 0.85

Leite, et. al. [Leite et al, 2011] proposed a fuzzy model for processing and monitoring vital signs of ICU patients, and used MIMIC as a dataset, where the proposed model achieved accuracy of 96%.

Ramon, et. al. [Ramon et al, 2007] proposed a model for patient survival and length of stay prediction. The proposed model achieved 88% accuracy. They also proposed a model for prediction of development of endangering states where the achieved accuracy is 95%. In addition, they proposed a model for

prediction of recovery from endangering states where the achieved accuracy is 87%. The dataset used for these models was collected from various available data sources.

Results and Discussion

This section discusses the different machine learning techniques that are used for the intensive care unit data. These techniques are applied to many tasks such as ICU mortality prediction, ICU sepsis prediction, and ICU monitoring.

Table 1 shows a comparison between different machine learning approaches used in mortality prediction of ICU patients.

Table 1. ML approaches in ICU Mortality Prediction

Authors	Dataset	Preprocessing & Feature Extraction	Machine Learning Technique	Evaluation
Che, et. al., 2018 [Che et al, 2018]	MIMIC-III		Linear Regression	AUC = 0.7715±0.015
		Principle Component Analysis (PCA)	Linear Regression	AUC = 0.7246±0.014
			Support Vector Machine (SVM)	AUC = 0.8146±0.008
		PCA	SVM	AUC = 0.7235±0.012

Authors	Dataset	Preprocessing & Feature Extraction	Machine Learning Technique	Evaluation
			Random Forest (RF)	AUC = 0.8294±0.007
		PCA	RF	AUC = 0.7747±0.009
			GRU	AUC = 0.8380±0.008
			Proposed GRU_D	AUC = 0.8527±0.003
	PhysioNet Dataset		Linear Regression	AUC = 0.7625±0.004
		PCA	Linear Regression	AUC = 0.6890±0.019
			SVM	AUC = 0.8277±0.012
		PCA	SVM	AUC = 0.7741±0.014
			RF	AUC = 0.8157±0.014
		PCA	RF	AUC = 0.7561±0.025

Authors	Dataset	Preprocessing & Feature Extraction	Machine Learning Technique	Evaluation
			GRU	AUC = 0.8226±0.010
			Proposed GRU_D	AUC = 0.8424±0.012
Anand, et. al., 2017 [Anand et al, 2018]	MIMIC-III (The Diabetes inpatients)		RF with threshold 0.04	Accuracy = 61%
			RF with threshold 0.06	Accuracy = 74%
			RF with threshold 0.1	Accuracy = 85%
			RF with threshold 0.12	Accuracy = 89%
			RF with threshold 0.14	Accuracy = 92%
			RF with threshold 0.16	Accuracy = 94%

Authors	Dataset	Preprocessing & Feature Extraction	Machine Learning Technique	Evaluation
Pirracchio, et. al., 2015 [Pirracchio et al, 2015]	MIMIC-II		SAPS II	AUROC = 0.78
			SOFA	AUROC = 0.71
			SICULA	AUROC = 0.85
Ramon, et. al., 2007 [Ramon et al, 2007]	Different available data sources		Decision trees	Accuracy = 79%
			First Order RF	Accuracy = 82%
			Naïve Bayes (NB)	Accuracy = 88%
			Tree Augmented NB	Accuracy = 86%

From table 1, there are many machine learning techniques that can be used for mortality prediction of an ICU patient, such as linear regression, support vector machine, naïve Bayes, …etc. The Random Forest algorithm is the most common one used for mortality prediction.

Table 2 shows a comparison between different machine learning approaches used in prediction of sepsis for patients in ICU.

Table 2. ML approaches in ICU sepsis prediction

Authors	Dataset	Preprocessing & Feature Extraction	Machine Learning Technique	Accuracy
Nemati, et. al., 2018 [Nemati et al, 2018]	Data was collected from two hospitals within the Emory Healthcare system, as well as an external publicly available ICU database	4 hour data	An algorithm is based on a modified Weibull-Cox proportional hazards model	70%
		6 hour data		68%
		8 hour data		67%
		12 hour data		65%
Desautels, et. al., 2016 [Desautels et al, 2016]	MIMIC-III	The most recent bin value is carried forward to fill subsequent empty bins	Insight Classifier	80%
			SIRS	47%
			Quick SOFA	80%
			MEWS	76%
			SAPS II	55%
			SOFA	52%

From table 2, different ML techniques are used for proposing models for sepsis prediction of ICU patients. These models are applied to many patients' data with different intervals of time. The accuracy achieved for sepsis prediction varies from 47% to 80%.

Table 3 shows a comparison between different machine learning approaches used in ICU patient monitoring for different tasks.

Table 3. ML approaches in ICU monotoring

Authors	Task	Dataset	Preprocessing & Feature Extraction	Machine Learning Technique	Accuracy / AUC
Meyer, et. al., 2018 [Meyer el al, 2018]	Real time prediction of complications in ICU	Dataset collected from German tertiary care center for cardiovascular diseases from German Heart Center Berlin	The Postoperative bleeding, postoperative renal failure requiring renal replacement therapy, and postoperative in hospital mortality were labeled as "complication occurred" or "complication did not occur"	Recurrent deep neural network	Accuracy = 86%
		MIMIC-III			Accuracy = 75%
Viegas, et. al., 2017 [Viegas et al, 2017]	Prediction of ICU readmissions	MIMIC-II	exactly define the scope of allowable qualities as indicated by the expert	Average ensemble decision criteria	AUC = 0.77±0.02
				Linear SVM	Accuracy = 50.6%
				Polynomial SVM	Accuracy = 52.7%
				Radial basis function based SVM	Accuracy = 69.5%

Authors	Task	Dataset	Preprocessing & Feature Extraction	Machine Learning Technique	Accuracy / AUC
Leite, et. al., 2011 [Leite et al, 2011]	Processing and monitoring vital signs in ICU patients	MIMIC	Extraction of the major physiologic signals that interfere directly in the clinical condition of patients with a stroke diagnosis	Fuzzy Model	Accuracy = 96%
Ramon, et. al., 2007 [Ramon et al, 2007]	Prediction of development of endangering states	Different available data sources		Decision trees	Accuracy = 89%
				First Order RF	Accuracy = 95%
				NB	Accuracy = 95%
				Tree Augmented NB	Accuracy = 92%
	Prediction of recovery from endangering states			Decision trees	Accuracy = 82%
				First Order RF	Accuracy = 87%
				NB	Accuracy = 80%
				Tree Augmented NB	Accuracy = 85%

From table 3, The random forest, naïve Bayes, decision trees, SVM, and deep learning are used for many tasks regarding the ICU data such as complications prediction in ICU, ICU readmissions prediction, prediction of endangering states development, prediction of recovery from endangering states.

Conclusion and Future Work

This work presents an overview of ICU applications, by providing a structured analysis and a comparative study of numerous machine learning techniques which are used for ICU data analysis. The machine learning techniques proven to be a pioneer method for ICU data analysis. The MIMIC dataset is the most commonly used as a source of ICU data. In future work, we are going to apply the machine learning algorithms to anomaly detection over ICU data.

Bibliography

[Anand et al, 2018] Anand, R.S., Stey, P., Jain, S., Biron, D.R., Bhatt, H., Monteiro, K., Feller, E., Ranney, M.L., Sarkar, I.N. and Chen, E.S., Predicting Mortality in Diabetic ICU Patients Using Machine Learning and Severity Indices. AMIA Summits on Translational Science Proceedings, p.310, 2018.

[Che et al, 2018] Che, Z., Purushotham, S., Cho, K., Sontag, D. and Liu, Y. Recurrent neural networks for multivariate time series with missing values. Nature scientific reports, 8(1), p.6085, 2018.

[Desautels et al, 2016] Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., Shimabukuro, D., Chettipally, U., Feldman, M.D., Barton, C. and Wales, D.J. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR medical informatics, 4(3), 2016.

[Johnson et al, 2016] Johnson, A.E., Ghassemi, M.M., Nemati, S., Niehaus, K.E., Clifton, D.A. and Clifford, G.D. Machine learning and decision support in critical care. Proceedings of the IEEE. Institute of Electrical and Electronics Engineers, 104(2), pp.444-466, 2016.

[Leite et al, 2011] Leite, C.R., Sizilio, G.R., Neto, A.D., Valentim, R.A. and Guerreiro, A.M. A fuzzy model for processing and monitoring vital signs in ICU patients. Biomedical engineering online, 10(1), p.68, 2011.

[Meyer el al, 2018] Meyer, A., Zverinski, D., Pfahringer, B., Kempfert, J., Kuehne, T., Sündermann, S.H., Stamm, C., Hofmann, T., Falk, V. and Eickhoff, C. Machine learning for real-time prediction of complications in critical care: a retrospective study. The Lancet Respiratory Medicine, 6(12), pp.905-914, 2018.

[Nemati et al, 2018] Nemati, S., Holder, A., Razmi, F., Stanley, M.D., Clifford, G.D. and Buchman, T.G. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. Critical care medicine, 46(4), pp.547-553, 2018.

[Pirracchio et al, 2015] Pirracchio, R., Petersen, M.L., Carone, M., Rigon, M.R., Chevret, S. and van der Laan, M.J. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. The Lancet Respiratory Medicine, 3(1), pp.42-52, 2015.

[Ramon et al, 2007] Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M. and Van Den Berghe, G. Mining data from intensive care patients. Advanced Engineering Informatics, 21(3), pp.243-256, 2007.

[Saeed et al, 2011] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L. Lehman, G.B. Moody, T. Heldt, T.H. Kyaw, B.E. Moody, R.G. Mark. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access ICU database. Critical Care Medicine 39(5):952-960; doi: 10.1097/CCM.0b013e31820a92c6, 2011.

[Solanas et al, 2014] Solanas, A., Patsakis, C., Conti, M., Vlachos, I.S., Ramos, V., Falcone, F., Postolache, O., Pérez-Martínez, P.A., Di Pietro, R., Perrea, D.N. and Martinez-Balleste, A. Smart health: a context-aware health paradigm within smart cities. IEEE Communications Magazine, 52(8), pp.74-81, 2014.

[Viegas et al, 2017] Viegas, R., Salgado, C.M., Curto, S., Carvalho, J.P., Vieira, S.M. and Finkelstein, S.N., Daily prediction of ICU readmissions using feature engineering and ensemble fuzzy modeling. Expert Systems with Applications, 79, pp.244-253, 2017.

Authors' Information

Zeina Rayan *– T.A. & Msc Student, Computer Science Department, Faculty of Computer and Information Scinces, Ain Shams University, Cairo, Egypt;*

e-mail*: zeinarayan@ cis.asu.edu.eg*

Major Fields of Scientific Research*: Artificial Intelligence, Machine Learning, Smart Health, Medical Informatics*

Marco Alfonse *– Lecturer at faculty of computer and information sciences, Ain Shams University, Abbasia, Cairo, Egypt,*

e-mail*: marco@fcis.asu.edu.eg*

Major Fields of Scientific Research*: Semantic Web, Ontological Engineering, Medical Informatics, Artificial Intelligence, Expert Systems, Image Processing, Knowledge Engineering*

Abdel-Badeeh M. Salem *– Professor at faculty of computer and information sciences, Ain Shams University, Abbasia, Cairo, Egypt,*

e-mail*: absalem@cis.asu.edu.eg*

Major Fields of Scientific Research*: Artificial intelligence(AI), Computational intelligence, Machine learning, Knowledge engineering, Big data analytics, Intelligent biomedical informatics and Healthcare systems, and AI education*

THE USE OF COGNITIVE GRAPHICS IN THE DIAGNOSIS OF COMPLEX VISION PATHOLOGIES

Aleksandr P. Eremeev, Irina V. Tcapenko

***Abstract**: Discusses the use of cognitive graphics tools for the diagnosis of complex vision pathologies. Research and development are carried out jointly by the Department of applied mathematics of the National Research University "Moscow Power Engineering Institute" ("MPEI") and the Department of clinical physiology of vision of the Moscow Helmholtz Research Institute of eye diseases. The obtained results are used in the development of an intelligent (expert) decision support system for the diagnosis of complex vision pathologies.*

***Keywords**: Artificial Intelligence, Decision Support, Cognitive Graphic, Anomaly Detection, Vision pathology.*

Introduction

Methods of *cognitive graphics* (CG) are widely used, including in intelligent systems for various purposes for imaginative representation of the situation and a better perception of the processes [Pinker, 1984; Zenkin A.A., 1991]. It is known that the image (picture) is more convenient for human perception than text information, which is especially important in intelligent real-time systems. Typical representatives of such systems are intelligent decision support systems of real-time. At the Department of Applied mathematics of the National research University "Moscow power engineering Institute" together with the Department of clinical physiology of vision of the Moscow Helmholtz Research Institute of eye diseases conducted complex researches to creating an intelligent (expert) decision support system for the differential diagnosis of retinal diseases in complex clinical cases [Eremeev et al., 2014; Eremeev et

al., 2018]. For data analysis and diagnosis, an approach based on the integration of probabilistic methods, artificial intelligence methods (based on neural networks and fuzzy sets) and cognitive graphics tools for imaginative representation of the problem situation and the dynamics of its change is used. The main source of information for the diagnosis of complex visual pathologies such as diabetic retinopathy are the data obtained by *electroretinography* (ERG) - a method for assessing the functional state of the retina, based on the registration of biopotentials arising in it with light irritation [Shamshinova, 2009]. The use of means of CG can facilitate the process of perception of information and provide significant assistance to an ophthalmologist in the diagnosis, as well as be used to train novice specialists.

The structure of ERG

There are 3 main components in ERG: initial a-wave, b-wave and late c-wave (Figure 1). Depending on the type of ERG, the c-wave can be positive, negative or absent (in whole or in part).

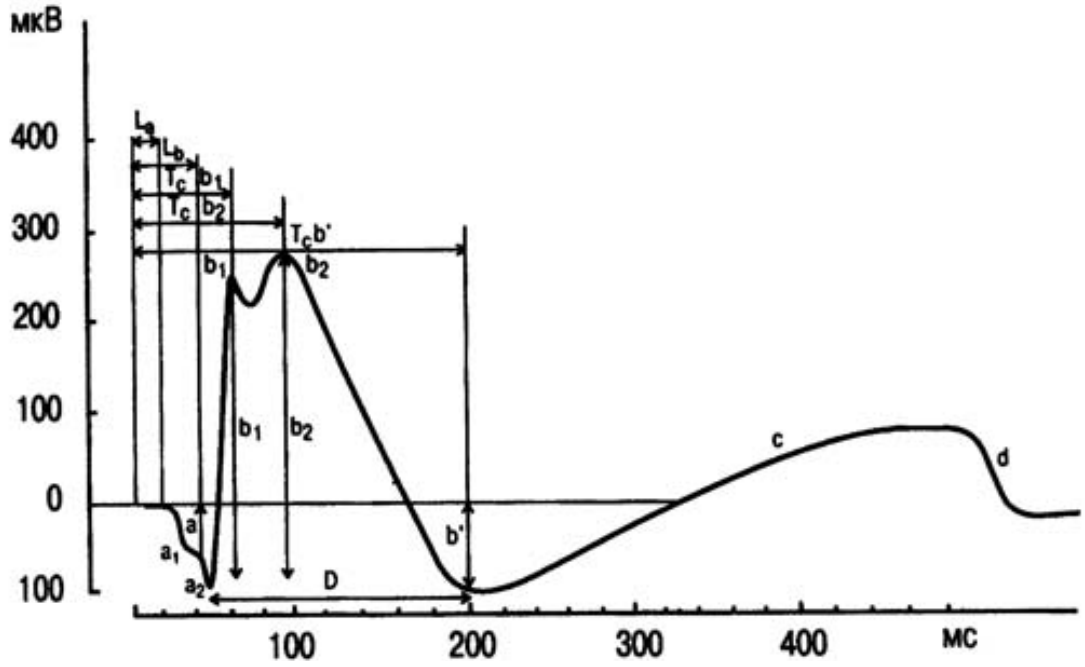

Figure 1. Schematic representation of maximum ERG: a1 and a2-amplitude of a-wave; b1 and b2 – amplitude of b-wave; D – duration of b-wave; L – latent

period; Tb – culmination time. On the ordinate axis-the amplitude of ERG waves, on the abscissa axis-the duration of ERG waves

Maximum ERG reflects the electrical activity of most cellular elements of the retina and the dependence on the number of healthy functioning cells. Different types of ERG reflect the diversity of retinal structure, and the evaluation of their components associated with different cellular elements of the retina, allows for early and differential diagnosis of retinal diseases. Figure 2 shows examples of ERG (Rod ERG, Rod-Cone ERG, Oscillatory ERG) used to diagnose retinal pathologies.

Negative *a*-wave ERG reflects the function of photoreceptors as the initial part of the late receptor potential. This wave has a dual origin, respectively, two types of photoreceptors: the earlier *a1*-wave is associated with the activity of the photopic system of the retina, *a*2-wave — with the scotopic system; *a*-wave passes into a positive b-wave, reflecting the electrical activity of bipolar and Muller cells with a possible contribution of horizontal and amacrine cells.

Wave *b* (on-effect) in the standard conditions of ERG registration reflects bioelectric activity depending on the adaptation conditions and the function of the photopic and scotopic retinal system, which are represented in the positive component by waves *b*1 and *b*2. On the ascending part of the *b*-wave there are 5-7 waveforms, called oscillatory potentials, which reflect the interaction of cellular elements in the inner layers of the retina, as well as the activity of amacrine.

The knowledge of the nature of retinal biopotential generation allows us to assess the nature and localization of retinal changes. In Figure 3 shows an example of ERG with localization of retinal changes.

In order to be able to perform a comparative assessment of the results of electroretinographphy studies conducted in different clinics around the world, the international society of clinical electrophysiologists of vision (ISCEV) proposed standards for the registration of ERG, recommended for the study of visual functions in patients with various pathological changes in the retina [Types, 2016].

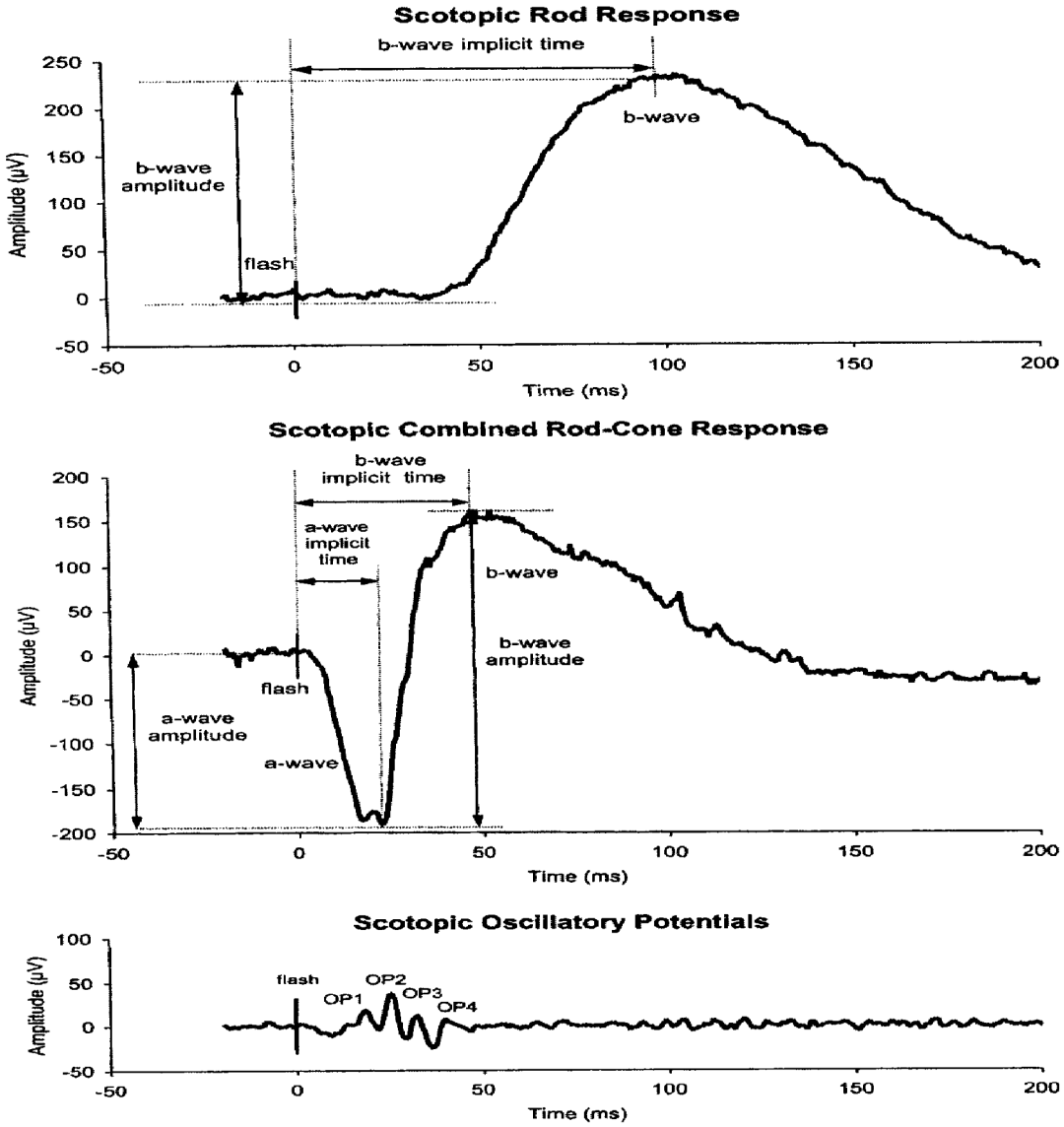

Figure 2. The examples of ERG used to diagnose retinal pathologies

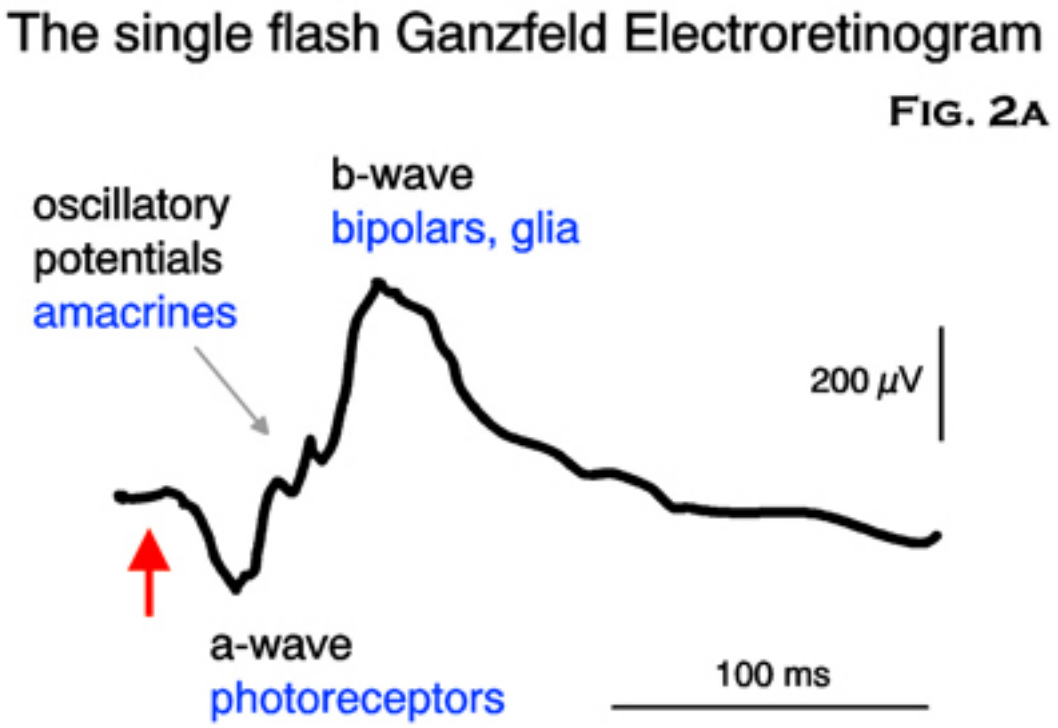

Figure 3. An example of ERG with localization of retinal changes

In clinical electroretinography using multiple ways of recording ERGS in order to distinguish the photopic, scotopic and mixed bioelectric responses of the retina. For this purpose, appropriate adaptation and stimulation conditions are used, in which the rod or cone system of the retina dominates (ERG registration for different flash intensity in special conditions of light and dark adaptation):

- maximum ERG under dark adaptation
- rod ERG in conditions of dark adaptation
- cone ERGS to single flash under light adaptation
- oscillatory potentials.

Additional information on retinal function can also be obtained using other electroretinographic techniques (which are not included in the "Standards"), such as: registration of chromatic ERG, macular ERG, multifocal ERG, ERG for long-term stimulus (on/off-response), early receptor potentials, registration of

scotopic response threshold, ERG for dual stimuli, *S*-cone ERG, determination of ERG dependence on the intensity of stimulating light (Naka-Rushton function). Visual representation

Visual representation of the retina

As known, in the retina there is a pre-processing of the entire flow of visual information. The conversion and transmission of signals in the retina in response to a light stimulus is a complex process that takes place in three stages [Zueva et al., 2002; Shamshinova, 2009]. The first stage involved the photoreceptors (rods and cones). Signals from photoreceptors come to bipolar cells (the second stage), and signals from bipolar cells — to ganglion cells (the third stage). Axons of ganglion cells form the optic nerve. The final output signal is the result of a complex integrative process occurring in the retina.

In the structure of the sensory retina are the following main layers, distinguishable under a microscope (Figure 4):

(1) the layer of outer segments of photoreceptors;

(2) the outer nuclear layer containing inner segments and nuclei of visual cells;

(3) the outer plexiform layer with synaptic contacts between photoreceptors, bipolar and horizontal cells;

(4) the inner nuclear layer containing nuclei of bipolar, amacrine and horizontal cells;

(5) the inner plexiform layer with a variety of contacts between bipolar, amacrine and ganglion cells

(6) the inner nuclear layer formed by the bodies of retinal ganglion cells;

(7) the layer of nerve fibers formed by axons of ganglion cells forming the optic nerve.

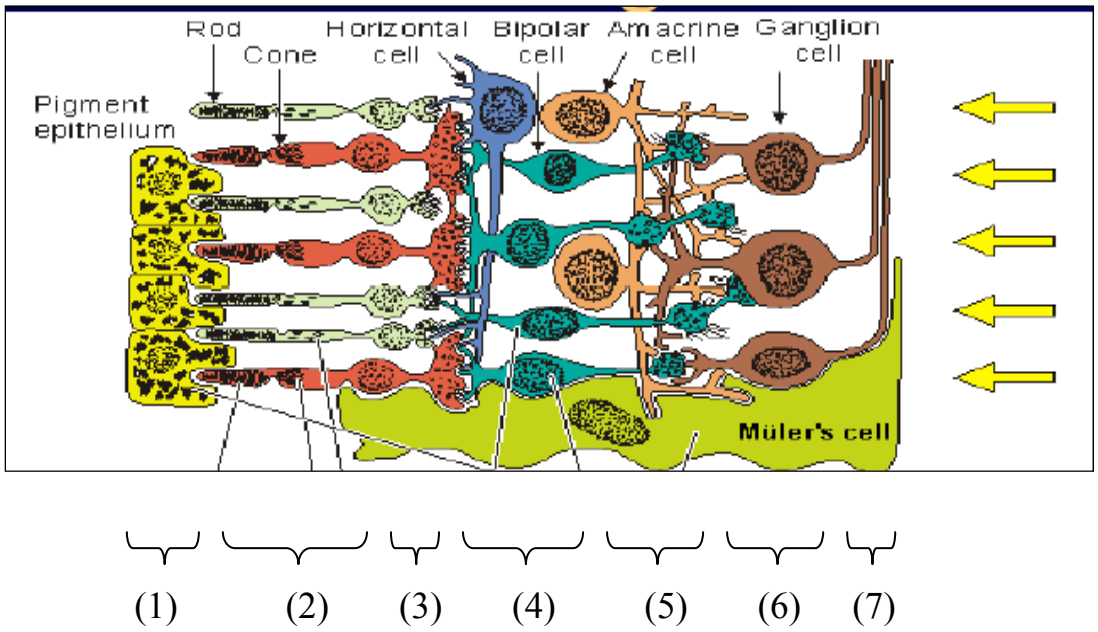

Figure 4. The main layers of the sensory retina [Webvision, 2016]

All the layers of the retina laced radial glial cells of Muller. The functions of these cells are extremely multifaceted. The intensive exchange of substances is realized trophic system of the capillary - glial cell - neuron. Muller glial cells are actively involved in the processes of repair of retinal damage, secrete various neuroactive substances. They can absorb the degenerating cells of the ganglion layer, as well as replace the nerve cells of the retina.

In retinal diseases, the normal functioning of the various layers of the retina is disturbed. Violations can go in two directions: increase or decrease the functional activity of cells. On the basis of what layers and divisions of the retina changes are observed and what character they are, it is possible to draw certain conclusions about what the pathogenesis of the disease is and correctly diagnose or clarify the diagnosis. In particular, in retinal detachment, the functioning of its layers is significantly reduced. The figurative representation of this pathology is given on Figure 5 (the result of a computed tomography of the retina obtained during examination of the patient).

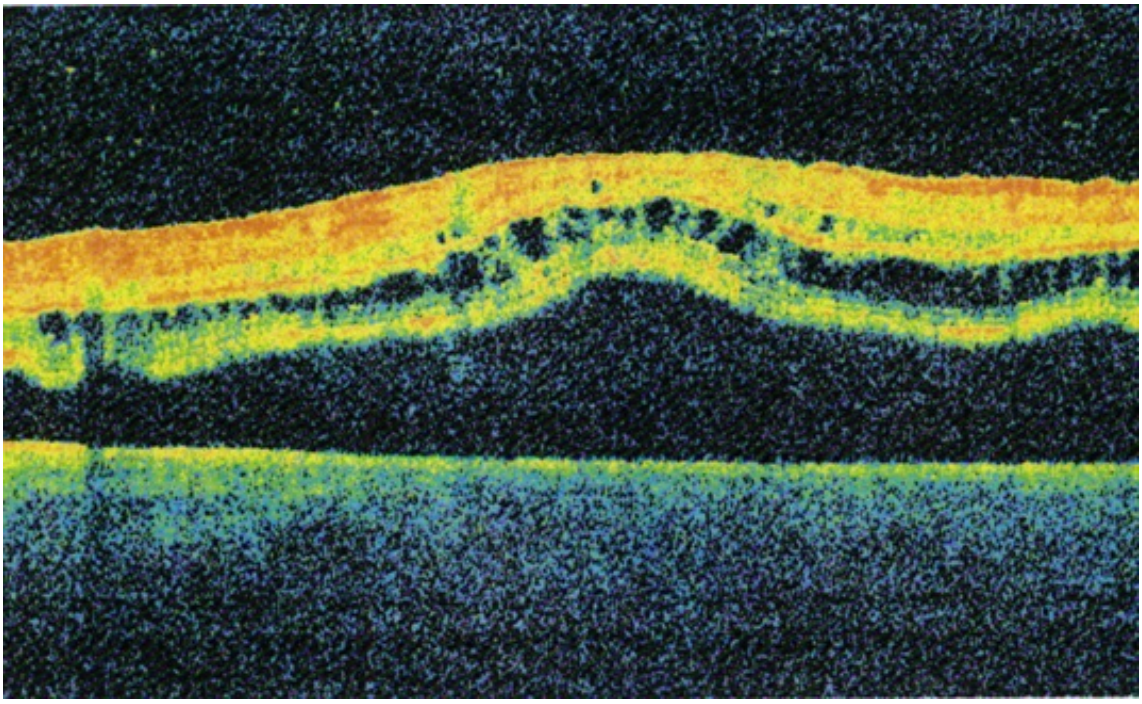

Figure 5. Retinal detachment

Formal model of the cognitive image of the retina

Formally, the model of cognitive image (CI-model) is defined by a triple: $K = (X, R, F)$, where $X = \{ C_i \mid i=1, 2, \ldots, n\}$ – the non–empty set of concepts (objects of the subject area), $R = \{ R_j \mid j=1, 2, \ldots, m \}$ – the family of relations on the set X, F – the set of interpretation functions (rules). Each concept C_i is a set: $C_i = (N_i , T_i , P_i , Ch_i , A_i)$, where N_i – the name of the concept C_i , $P_i = \{ C_i^k \mid k=1, 2, \ldots, x \}$ – the set of ancestors of concept C_i , $Ch_i = \{ C_i\ s \mid s=1, 2, \ldots, y \}$ – the set of descendants of a concept C_i , $A_i = \{ A_i^u \mid u=1, 2, \ldots, q \}$ – the list of attributes of concept C_i . Each attribute A_i^u of concept C_i is defined as follows: $A_i^u = (Na_i^u , Ta_i^u , Va_i^u)$, where Na_i^u – the name of attribute A_i^u , Ta_i^u – the type of attribute A_i^u , Va_i^u – the value of attribute A_i^u . T_i – the type of concept C_i is defined as follows: $T_i = (Nq_i , Aq_i , Mq_i , Pq_i)$, where Nq_i – the type name of concept C_i ; $Aq_i = \{ Aq_i^h \mid h=1, 2, \ldots, r \}$ – the list of attributes of concept type C_i ; $Mq_i = \{ Mq_i^d \mid d=1, 2, \ldots, f \}$ – the list of actions (methods) of concept type C_i ; Pq_i – the name of parent concept type C_i .

Then the alphabet in the CI-model of the retina is a set of geometric primitives (point, spline, ellipse, rectangle) $Aq_i = \{ Aq_i^h \mid h=1, 2, \ldots, r \}$, and the formal image is one layer element consisting of a sequence of alphabet characters.

The axioms describe building elements (elementary CI) each:

1) the image "Stick" consists of two ellipses and is in the first layer;
2) the image of the "Cone" consists of three ellipses and is in the first layer;
3) the image "Bipolar" consists of a rectangle and an ellipse, occupying 0.3 parts of the entire element, and is in the second layer;
4) the image of the "Ganglion cell" consists of a rectangle and an ellipse, occupying 0.8 parts of the entire element, and is in the third layer;
5) the image of the "Horizontal cell" consists of an ellipse and a spline and is in the second layer;
6) the image of "Amacrine cell" consists of an ellipse and is located in the second layer;
7) the image of" Muller's Cell " consists of a rectangle and a set of ellipses of two types and is located on the side of all other layers.

Reasoning (output) rules define relationships between elements in different layers:

1) images of "Stick" and "Cone" can be associated with elements of "Bipolar" and " Horizontal cell»;
2) images of "Bipolar" and "Amacrine cell" may be associated with the element "Ganglion cell»;
3) the image of the "Ganglion cell" has a downward connection called "Axon".

The set of concepts is defined as follows:

C1= (Stick, 0, Bipolar (stick), A1);

C2= (Cone, 0, Bipolar (cone), A2);

C3= (Bipolar (rod), Wand, Ganglion cell, A3);

C4= (Bipolar (cone), Cone, Ganglion cell, A4);

C5= (Ganglion cell, Bipolar (rod), Axon, A5);

C6= (Ganglion cell, Bipolar (cone), Axon, A6);

C7= (Axon, Ganglion cell, 0, A7);

C8= (Horizontal cell, Stick, Bipolar(stick), A8);

C9= (Horizontal cell, Cone, Bipolar(cone), A9);

C10= (Amacrine cell, Bipolar, Horizontal cell, A10).

The list of three attributes is defined for each concept $A_i = \{ A_i^u \mid u=1,2,3\}$:

A_i^1 *= (Color, 3 digits, 0-255);*

A_i^2 *= (Width, number, Wi);*

A_i^3 *= (Height, number, Hi).*

Similarly, multiple sets of indicators are introduced $P = \{ P_i^h \mid h=1, 2, \ldots, r \}$:

P_i^h *= (type of study, indicator, value);*

We give examples of indicators (the indicators are presented in full in the Appendix):

P_i^1 *= (Maximum ERG, AA, 70-220) - shows the change in photoreceptors (rods and cones);*

P_i^5 *= (Scotopic ERG, al, 50-150) - shows changes in rod bipolar cells;*

P_i^8 *= (Photopic ERG, Ta, 7-21)) - shows the change in the bonds between the cones and their Bipolars;*

P_i^{11} *= (FERG 30 Hz, APIC-dead end, 30-100) - shows change in cone bipolar without Mueller;*

P_i^{12} *= (Oscillatory potentials, O1, 40-70) - shows the change in Amacrines and inverse coupling functions from ganglion cells to Amacrines;*

P_i^{18} *= (ERG pattern, N95, 30-110) - shows changes in the ganglion cells and axons of cones (more in the macula).*

Calculation of indicators is based on ISCEV recommendations [Types, 2016]:

- the amplitude of *Ab* is calculated from 0 to max negative;
- the amplitude of *Ab* is calculated from *a* to max positive;
- the climax time of *Ta* is calculated from the axis to max negative;
- the climax time of *Tb* is calculated from axis to max positive;
- the amplitude of *A* peak-to-peak is calculated from max negative to max positive;
- the amplitude of *P50* is calculated from the first negative deflection to the max *P50*;
- the amplitude of *N95* amplitude is calculated from the *N50* (or contour) to the *N95*;
- the amplitudes *O1, O2, O3, O4* are calculated from max negative to max positive;

The knowledge base is a set of production rules of the type "***If*** (condition of the rule application), ***then*** (result of the rule application (conclusion of the rule))". The condition (the left part of the rule) is a set of indicators, the result-conclusion (the right part of the rule) is a set of concepts with attributes.

Examples of production rules:

If P^1 *is (Maximum ERG, Aa, 60,* ***then*** A_1^1 *is (100.100.70) (color dim);*

If P^1 *is (Maximum ERG, Aa, 100),* ***then*** A_1^1 *is (150.79) (color moderate);*

If P^{13} *is (Photopic ERG, Aa, 15),* ***then*** A_2^2 *is (10) (width is small);*

If P^{13} *is (Photopic ERG, Aa, 30),* ***then*** A_1^1 *is 1(20) (width average);*

If P^{14} *is (Photopic ERG, Ab, 150),* ***then*** A_3^1 *is (200.220.130) (color bright).*

If the condition of the rule is executed, then the result-conclusion of the rule is used to build the CI of the retina. Then, to assess the state of each type, a fragment of the retina, a combined visualization method can be used, which consists in changing the size of the image and its color depending on the selected parameters.

Software implementation

In the software system of visualization of complex pathologies of vision based on CI, the main users are the expert (electrophysiologist) and the user – the decision-maker personal (DMP), which is an ophthalmologist.

From the point of view of experts and DMP, the system should provide the following main functions:

- changes in indicators;
- getting display results;
- saving results;
- view help;
- program setting.

Diagram of use cases of the developed system is shown in the Figure 6.

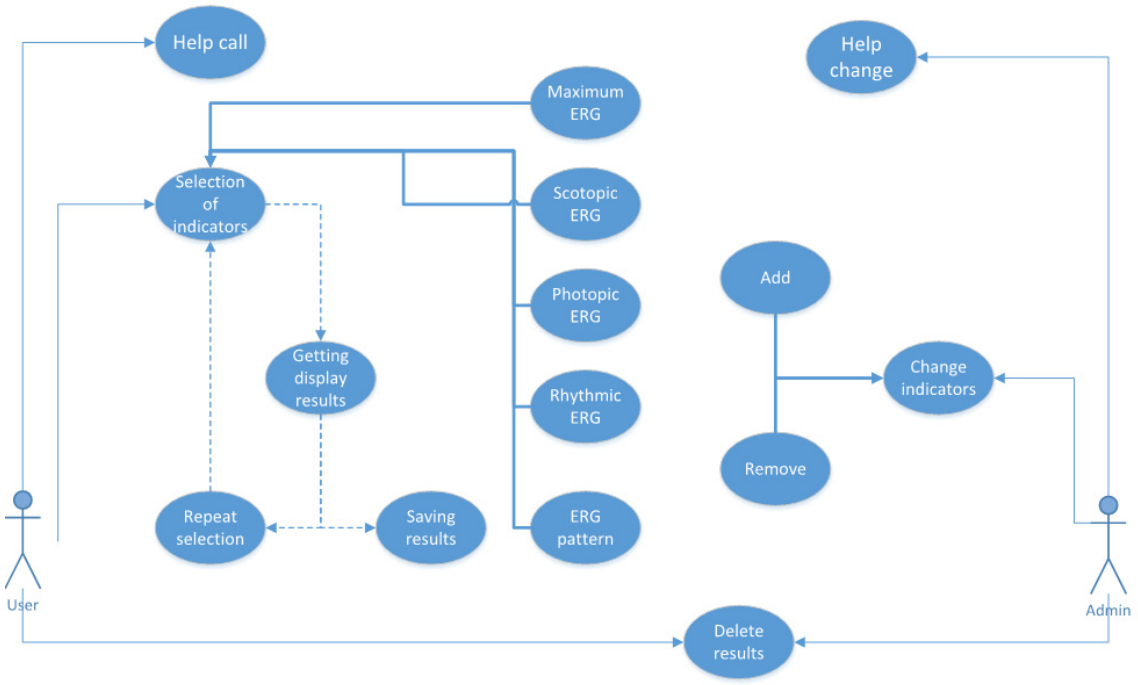

Figure 6. Diagram of use cases of the developed system

When working with the system, the user (expert, DMP or trainee) should be able to set values for all possible indicators. If the user has problems with the program, he can use the help by calling it in the "Menu". After determining all the indicators, the "Build" button is used to create a cognitive image.

Examples of cognitive images constructed by the system for the normal state and various pathologies of the retina are shown in Figure 7.

When developing the system, an object-oriented technology and C# implementation language were used, which allowed to present the main elements of the CI with the necessary expressiveness.

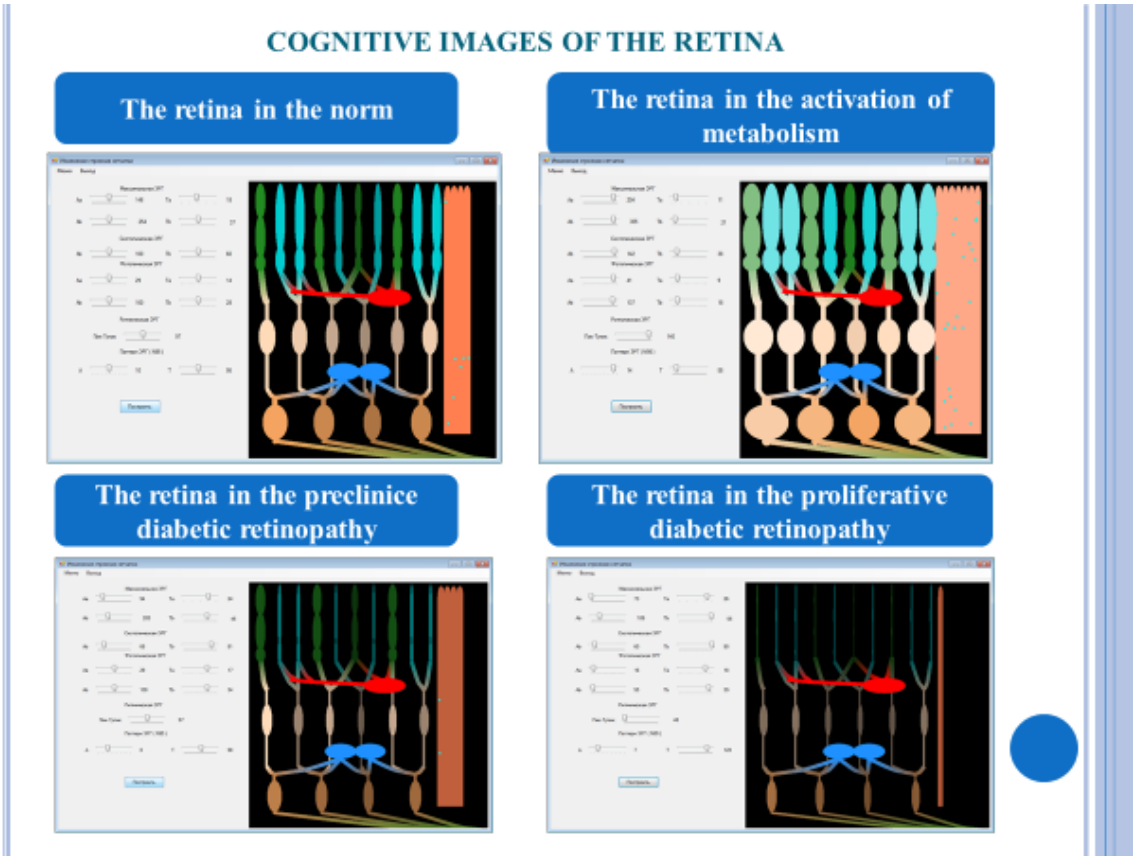

Figure 7. Examples of cognitive images

Result

The possibilities of using cognitive graphics (cognitive images) and the corresponding software system for the presentation of complex pathologies of vision according to the results of ERG are considered. A formal model of a cognitive image (CI-model) is proposed, including a set of basic elements for constructing CI and a set of production rules (reasoning rules) defining relations between elements of different levels of CI. Note that the results of the ERG can only be interpreted by physiologists. The developed means of cognitive graphics allow us to give a figurative representation of changes in the layers of the retina, which is also understandable to the ophthalmologist. The developed tools are intended to be included in the intelligent decision support system for diagnosing complex vision pathologies developed jointly by specialists from the National Research University "MPEI" and expert-physiologists from the Moscow Helmholtz Research Institute of eye diseases, a prototype system has been developed that can be used to train students and novice doctors.

Further work: An important factor is the further expansion of the library (database) of cognitive images to other pathologies of vision, as well as the implementation of the dynamics of changes in images, which will allow you to display the process of the disease in a given time mode (accelerated, normal or slow).

Acknowledgements

The work is executed at financial support of RFBR (projects 17-07-00553, 18-51-00007, 18-01-00201).

Appendix

The set of indicators:

P_i^1 = (Maximum ERG, AA, 70-220) - shows the change in photoreceptors(rods and cones);

P_i^2 = (Maximum ERG, TA, 10-30) - shows the change in the connections between photoreceptors and their Bipolars;

P_i^3 = (Maximum ERG, Ab, 125-380) - shows change in sticks and their bipolar cells, also in Muller cell;

P_i^4 = (Maximum ERG, Tb, 18-56) - shows the change in the bonds between the sticks and their Bipolars;

P_i^5 = (Scotopic ERG, al, 50-150) - shows changes in rod bipolar cells;

P_i^6 = (Scotopic ERG, Tb, 30-90) - shows the change in the bonds between the sticks and their Bipolars;

P_i^7 = (Photopic ERG, Aa, 14-44) - shows the change in cones;

P_i^8 = (Photopic ERG, Ta, 7-21)) - shows the change in the bonds between the cones and their Bipolars;

P_i^9 = (Photopic ERG, Ab , 50-150) shows change in cone Bipolars and Muller cells;

P_i^{10} = (Photopic ERG, Tb, 14-42) - shows the change in the bonds between the cones and their Bipolars;

P_i^{11} = (FERG 30 Hz, APIC-dead end, 30-100) - shows change in cone bipolar without Mueller;

P_i^{12} = (Oscillatory potentials, O1, 40-70) - shows the change in Amacrines and inverse coupling functions from ganglion cells to Amacrines;

P_i^{13} = (Oscillatory potentials, O2, 40-70) (basic index) -shows the change in Amacrines and inverse functions of the connection from ganglion cells to Amacrines;

P_i^{14} = (Oscillatory potentials, O3, 40-70) - shows the change in Amacrines and inverse coupling functions from ganglion cells to Amacrines;

P_i^{15} = (Oscillatory potentials, O4, 40-70) - shows the change in Amacrines and inverse coupling functions from ganglion cells to Amacrines;

P_i^{16} = (ERG steadystate pattern, APIC-deadlock , 50-120) - shows changes in ganglion cells and axon;

P_i^{17} = (ERG pattern, P50, 45-90) - shows change in photoreceptors and cone bipolars;

P_i^{18} = (ERG pattern, N95, 30-110) - shows changes in the ganglion cells and axons of cones (more in the macula).

Bibliography

[Pinker, 1984] Pinker, S. Visual cognition: An introduction. Cognition, 18(1-3), pp. 1-63.

[Zenkin, 1991] Zenkin A.A. Cognitive computer graphics / Ed. D. A. Pospelov. - Moscow: Science, 1991. (In Russian).

[Eremeev et al., 2014] Aleksandr Eremeev, Ruslan Khasiev, Irina Tcapenko, Marina Zueva. The Intelligent Decision Support System for Diagnostics of Difficult Diseases of Vision // International Journal "Information Content & Processing", 2014, Volume 1, Number 3, pp. 269-279. ISSN 2367-5128 (printed), ISSN 2367-5152 (online).

[Eremeev et al., 2018] Eremeev, A. P., Ivliev, S. A., & Vagin, V. N. (2018). Using Nosql Databases and Machine Learning for Implementation of Intelligent Decision System in Complex Vision Patalogies // 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC). Pp. 150-153. doi:10.1109/rpc.2018.8482230.

[Shamshinova, 2009] Shamshinova A. M. Electroretinography in ophthalmology. - M.: Medica, 2009. (In Russian).

[Types, 2016] Types of ERG recommended by ISCEV. https://zreni.ru/print:page,1,3066-metodicheskie-osnovy-sovremennoy-elektroretinografii.html

[Zueva et al., 2002] Zueva M. V., Tcapenko I. V., Neroev V. V., Zakharova G. Yu., Lysenko V. S. The role of electroretinography in the diagnosis and study of the pathogenesis of diabetic retinopathy // Collection of scientific papers "Clinical physiology of vision", dedicated to the 100th anniversary of Professor A.I. Bogoslovsky, Moscow 2002. Pp. 347-358. (In Russian).

[Webvision, 2016] Webvision. The Organization of the Retina and Visual System. Editors Helga Kolb, Ralph Nelson, Eduardo Fernandez and Bryan Jones. http://www.webvision.med.utah.edu

Authors' Information

Eremeev Aleksandr *– National Research University "Moscow Power Engineering Institute"; Head of the Applied Mathematics Department. P.O. Box: 14, Krasnokazarmennaya Str., Moscow, Russia, 111250; e-mail: eremeev@appmat.ru*

Major Fields of Scientific Research: Artificial intelligence, Decision making, Decision support system, Expert system

Tcapenko Irina *– Moscow Helmholtz Research Institute of Eye Diseases; Major Research Worker; P.O. Box: 14/19, Sadovaya-Chernogriazskaya Str., Moscow, Russia, 105062; e-mail: sunvision@mail.ru*

Major Fields of Scientific Research: Vision pathology, Data mining, Expert system

TABLE OF CONTENTS