2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing

Data mining in on-line social network for marketing response analysis

Jerzy Surma
Faculty of Business Administration
Warsaw School of Economics
Warsaw, Poland
jerzy.surma@gmail.com

Anna Furmanek
Faculty of Economic Analyses
Warsaw School of Economics
Warsaw, Poland
a.k.furmanek@gmail.com

Abstract— **Business usage of online social networks is a natural result of their intense development in last years. The information about members of a given community can be treated as a basis of correct identification of their needs and as a result adjusting personalized marketing messages. In this study, we will discuss the classification and regression trees (C&RT) model for identifying users of on-line social network likely to respond to a marketing campaign. This model is aimed at using the advanced data mining methods to enable business usage of social networks and related study problems concerning the importance of relational attributes in customer behavior analysis. The research presented in this paper confirms the usage of data mining techniques in marketing campaign optimization. This was justified by significant improvement in response rate. We also showed that extension of the user description by relational attributes did not improve the classical approach based on the individual attributes**

Keywords: On-line social networks analysis, data mining, marketing response analysis.

I. INTRODUCTION

We can divide data that is contemporarily registered in the context of social network into data related to a given user (e.g. age, sex, place of residence etc.) – individual data – and data describing relations between the members (e.g. frequency of contacts with other users, number of invitations, number of comments on the blog etc.) – relational data. The described data types have declarative or behavioral character, i.e. resulting from real user behaviors. This behavioral dimension of data left by large user communities is related to their substantial volume e.g. in case of long-term and intensive activities, it allows us to conduct valuable statistical social network analysis [1]. The usage of data mining in this area seems to be a promising approach [2].

The research environment is Biznes.Net – one of many polish social networks that helps with creation and maintenance of business relationships. The membership of this virtual community gives users an opportunity to meet business partners and to present their professional experience via editing their profiles, writing blogs or attending to many specified groups. Moreover, the participation in Biznes.Net can stimulate personal development of users by providing them with valuable information about business events such as trainings or workshops

In the next part of the article we will present an analytical model for classification based on decision tree C&RT [3], to be able to predict the reaction for the marketing message [4]. This model was built with the use of individual and relational attributes.

II. RELATED WORK

The main concept in this research is the business dimension of on-line social network analysis in interactive marketing. Since over the last decade we could observe a rapid popularity growth of virtual communities, they can be perceived as a graceful and valuable environment to social research in economics [5] and in organizational context [6]. The outstanding amount of available behavioural data in online social networks gives an opportunity for a knowledge discovery due to data mining techniques. The overview and research in this area is described by Han and Kamber [7].

III. PROBLEM DESCRIPTION

It should be remarked that the social networks phenomenon has put a great impact on our ordinary life since the new century began. The more people are involved in the virtual communities, the more expectations we have with potential business value of on-line networks. According to those facts, the object of our research is to indicate one of possible paths that can bring us measurable return with on-line social network examination. In our experiment, we take advantage of the C&RT model that is an adequate approach for analysing marketing campaign based on the binary response. However, taking into account the peculiarity of the research environment, the participation of relational (network) attributes in C&RT model construction is separately discussed in this paper.

Firstly, we will introduce a characteristic of the research environment and the data set specification. Then the experiment assumptions will be presented.

978-0-7695-4578-3/11 $26.00 © 2011 IEEE
DOI

A. Social network and data set specification

The Biznes.Net specification is described by a set of metrics (see Table I) that are typical for "small world" system [8] where most of the vertices belongs to the giant component (2828 from 3025) and the average path between any two users is relatively short (3.53) [9].

TABLE I. THE GROUP OF MAIN NETWORKS METRICS

Indicator	Value
Members	19593
Total number of vertices	3025
Number of components	90
Number of vertices in giant component	2828
Density	0,002
Average friends number	5.88
Betweenness centrality	0.21
Clustering coefficient	0.23
Average path length	3.53

Initial data set, organised in 62 attributes, contains information about 19593 network members. The attributes are prepared with information available in relational database that supports Biznes.Net social network. They are both numerical and categorical. All the predictors can be perceived as declarative (e.g. gender, age) as well as behavioural. – which means that they are a representation of real actions taken by the user in the network environment. This behavioural aspect of the data set is particularly important in our research. According to this fact, we assume that the missing values of some categorical attributes can be perceived as a given value – NULL which can be explained as conscious concealment of some information by the users. All obtained data is anonymous and the users are identified only with their unique id number. As it was already mentioned all attributes can be divided into two groups: individual and relational attributes. First group of attributes corresponds only to a particular user. In the second group we find the predictors describing all existing relations among the network members.

B. Experiment assumptions

In our research the C&RT model is used to built an effective classifier of marketing campaign recipients. The predicted versus actual response can be classified into four categories: True Positive, True Negative, False Positive and False Negative [4]. The True Classes are those users properly recognized by the model as truly interested in our offer or being indifferent to received advertisement. The False Classes are those users that model classifies as positively or negatively interested in our offer and in reality it turns out that it's the opposite.

TABLE II. CATEGORIES OF CLASSIFICATION

		Classified	
		True	**False**
Observed	**True**	**TP** true positive	**FN** false negative
	False	**FP** false positive	**TN** true negative

TABLE III. COST - REVENUE - PROFIT SUMMARY

	Revenue	Cost	Profit
TP	z	0.01 z	0.99 z
TN	0	0	0
FP	0	0.01 z	-0.01 z
FN	0	Z	-z

This four classes are presented in Table II. We will take an assumption (based on the expert knowledge) that each positive response gives us z units of revenue while a cost of sending one advertisement is estimated as $0.01z$. With these assumptions and referring to classification categories, we can evaluate revenues, costs and profits of classification for each of four previously defined groups of customers. A summary is presented in Table III.

IV. EMPIRICAL EVALUATION

This paragraph presents the general concept of our empirical research. We define a classifier that ameliorates the response ratio of marketing campaign in social network environment. In further theoretical analysis we want to investigate what is the significance of the knowledge of social relationships among the network users in model construction. This can reveal whether we are able to exploit the additional source of information that gives us the social network environment.

A. Response analysis

The empirical research is based on the response from marketing campaign that receivers are randomly chosen users of Biznes.Net social network. The carrier of the advertisement is an e-mail with a link redirecting to advertiser's web page. We state the response as a dependent attribute (variable). We define a positive response (YES category) as clicking the link and the opposite activity as a negative response (NO category). Hence, the response attribute is a categorical (binary) variable of two possible values {YES; NO}.

During the architecture stage, many C&RT models are obtained. All trees are constructed by applying learning set of 836 records, and they are tested on the set of 8150 records. Moreover, all models are built with the same stopping, validation and classification conditions. The tree is pruned when misclassification error occurs and the minimum number of cases in a node, to be considered for splitting is 83. In the validation process we apply 10-cross validation and while modelling, a misclassification costs are established as equal.

TABLE IV. RESULTS OF MARKETING CAMAPAIGN PERFORMED WITHOUT AND WITH CLASSIFIER

	Group chosen randomly	Group chosen by C&RT model	Summary
Sample size	1918	2052	3970
Response	71	131	202
Response proportion	**3.70%**	**6.38%**	**5.10%**

The C&RT algorithm implementation allows us to generate a classifier that nearly doubled the response of marketing campaign (see Table IV). By implementing this model on the test set we have obtained a 6.4% response ratio which is significantly better (significance level is 0.01) from the results of random response (3.7%). See the detailed description of the experiment in Surma and Furmanek paper [10]. Moreover, by analysing selected tree, we can observe that the best predictor is the days_from_last_login attribute. This situation has its rational explanation, since it indicates the most active users. All further splits bring better adjustments. The obtained results confirm that the data mining methods are justifiably approved as tools bringing tangible profits in business dimension.

B. Relational attributes analysis

After building the effective classifier, we would like to concentrate on examination of the significance of both individual and relational attributes in C&RT model construction. We would like to explore whether and how strong profits can bring us the recognition of different social interactions among the Biznes.Net members.

TABLE V. TYPES OF ATTRIBUTES EXPLOIT IN THE EXPERIMENT

Type of attribute	Number	Example
Individual	25	voivodeship, branch, title, age, gender, etc.
Relational	37	no_of_friends, sent_invitations, open_messages, groups_joint, etc.
Relational SNA	6	degree, fareness, nCloseness, clusterCoefficien, nPairs, Betweenness

The initial data set contains 62 attributes where 25 are individual and 37 are relational. In this stage we expand this list with 6 classical network indicators: degree, fareness, nCloseness, cluster Coefficien, nPairs, Betweenness. All of them are obtained with Ucinet program [11]. Eventually, we used of 68 attributes (see Table V) where 25 are individual and 43 are relational (including 6 SNA indicators). Afterwards, the new data set with 68 attributes and 795 network members is created. The 51 elements present positive response (response = YES) while the rest (744) are in the opposite category (response = NO).

TABLE VI. IMPORTANCE RANKING – THE BEST 10 PREDICTORS CONSIDERING THE DEPENDENT ATTRIBUTE (VARIABLE)

Attribute	Importance	Type
Voivodeship	1.000000	I
Created	0.796853	R
Branch	0.765648	I
position_current	0.656893	I
groups_joint	0.624998	R
Title	0.608506	I
credit_avg	0.602852	R
Clust coef	0.573255	R
credit_last_3months	0.571357	R
credit_last_6months	0.539559	R

Legend: I – individual attribute, R – relational attribute

The initial analysis indicates that if we consider both types of attributes (individual and relational) as predictors in model construction process, we can generally observe tangible impact on classifier architecture of the relational attributes (see Table VI). However, if we look at the structure of the tree built with both types of attributes, the relational predictor (group_joint) appears only at the fourth level and none of SNA indicators participates in the model construction process. The high positions of relational attributes in the importance ranking and their absence in the model architecture is an effect of the C&RT algorithm procedure. During each split the algorithm creates a local ranking and with the winner the split is proceeding. The relational attributes lose with the individual in most cases. Although, the sum of "the second positions" gives them relatively high position in global ranking of the attributes importance (see Table VI).

The further analysis, where separate models, using either individual or relational attributes, are created, also depreciate significance of relational attributes. Moreover, the network indicators do not participate in classification process. Thus, we conclude that the relational attributes are not significantly useful for classifier creation and do not particularly influence the marketing campaign effectiveness.

TABLE VII. PROFIT SUMMARY FOR THREE CONSIDERED MODELS

	Tree based on individual attributes	Tree based on relational attributes	Tree based on both individual and relational attributes
Profit	42	24	43

Legend: Profit = Revenue - Cost.
Revenue: positive response (1z), Cost: misclassification (-1z) , cost of sent message (-0.01z)

Our observations are illustrated by Table VII, where a profit estimation for models with different predictors set is presented. This calculation is based on cost – revenue assumptions presented in Table III. Classifier built with only relational predictors is nearly twice less effective than model constructed with both types of attributes. Moreover, there is only a slight amelioration in model effectiveness when we

include relational attributes in the predictors set (42 vs. 43). Obtained results, slightly disappointing, may be affected by several factors:

- Weak representativeness of the learning set –Firstly, we decide to consider only the "active" users (last modification no later than 20 months). Secondly, we have got limited access to the information required to SNA indicators evaluation. Thus, we limited the set to 795 users which is less than 5% of the initial number of network actors and cannot be a guarantee of representative sample reflecting all the complexities of network relation.

- Weakness of the network – The SNA indicators (see Table 1) show low level of social interactions among the network members. Thus, there is a potential weak dependency between them and a dependent variable.

- Attributes simplicity – The relational attributes do not reflect the semantics of the network's relations because they reflect only a limited quantitative dimension of the phenomena. Moreover, with these attributes, we are not able to catch the complexity and semantic aspect of network interactions.

- Deficiency of correlation – It is hard to comprehend a possible relation between any relational attribute and the marketing response. On the contrary, the individual attributes can be rationally interpreted as the factors indicating the proper classification.

- Lack of measurement – The social network users have an opportunity to communicate and exchange information about proposed products that might affect significantly the response level. This crucial dimension of user referrals was not taken into account.

V. CONCLUSIONS AND FUTURE RESEARCH

The research presented in this paper confirms the usage of data mining techniques in marketing campaign optimization. This was justified by selecting individual attributes by C&RT model and significant improvement in response rate. We showed that simple extension of the user description by relational attributes did not improve the classical approach based on the individual attributes. The proper use of social relation knowledge is much more complex, and in the part B of chapter IV we have described this problem. Fulfilling all the factors mentioned at the end of previous chapter implicates non-trivial research problems. Research that were launched in this article will be continued not only in the context of the unusual intellectual challenge, but also because of relevance to business needs.

REFERENCES

[1] Wasserman, S., Faust, K. Social Network Analysis. Methods and Applications, Cambridge University Press, 2009.

[2] Larose, D.T. "Data Mining Methods and Models," 2006, Wiley, New York, 2006.

[3] Breiman, L.,Friedman, J., Stone, C.J. and Olshen R. "Classification and regression trees", Chapman & Hal, 1984.

[4] Chiu, S., Tavella, D. "Data mining and market intelligence for optimal marketing returns", Elsevier, 2008,

[5] Easley, D., Kleinberg, J. "Networks, Crowds, and Markets", Cambridge University Press, Cambridge, 2010.

[6] Kilduff, M., Wenpin, T. "Social Networks and Organizations", Sage, Los Angeles, 2003.

[7] Han, J., Kamber, M. Data Mining. Concepts and Techniques. Morgan Kaufmann, San Francisco, 2006.

[8] Watts, D.J., Strogatz, S.H. "Collective dynamics of 'small-world' networks,",Nature, 393, 1998, p.p. 440-442.

[9] Newman,, M.E.J. "The Structure and Function of Complex Networks," SIAM Review, 45(2), 2003, p.p. 167-256.

[10] Surma, J., Furmanek, A. "Improvig marketing response by data mining in social network", The 2nd International Conference on Mining Social Networks for Decision Support, Odense, 2010.

[11] Borgatti, S.P., Everett, M.G. and Freeman, L.C. "Ucinet for Windows: Software for Social Network Analysis", Harvard, MA: Analytic Technologies, 2002.