RESEARCH PAPERS OF THE WROCŁAW UNIVERSITY OF ECONOMICS
No. 25 2008
Knowledge Acquisition and Management

Janina Anna Jakubczyc

Wrocław University of Economics, Wrocław, Poland
janina.jakubczyc@ue.wroc.pl

CONTEXTUAL CLASSIFIER ENSEMBLE FOR PREDICTING CUSTOMERS CHURN

Abstract: Problem of customers churn is becoming more and more important for many companies. Looking for appropriate technique to support an effective identification of migrants as well as causes of the migration is the subject of this article. We present appealing approach, which is contextual classifier ensemble that can find acceptable solution for this difficult problem.

Key words: churn, classification, classifier ensemble, contextual classifier ensemble.

1. Introduction

The markets that are characterized by free and easy opportunities to change supplier of specific services or goods, and high cost of client obtainment, and high demand saturation, they are becoming the challenge for researches. Masses of wandering customers looking for the satisfaction in the successive companies are increasing bearing almost no cost from this title. This situation is forcing the changes of the strategy to manage relations with customers. Acquiring the customer is not already enough, retention him is becoming the more important objective. Many companies are discovering that actions in this scope are failing [Berry, Linoff 2000; Customer 2007]. Causes of such a state of affairs are being suspected to the lack of:

– formulated strategy for customer churn;
– possibility of identification of customers generating the maximum value;
– communication going beyond a moment of sale of goods or the services or limited knowledge of external and competitive factors which are influencing the decision of the customer's churn;
– integration of the co-operation of customers and service staff which would ensure the customer's satisfaction.

For solving the formulated problem an effective identification of migrants as well as causes of the migration is the first step and this is making the subject of the

this article. Many causes of the migration can be identified what is resulting from researches carried out [Customer 2007]. These reasons can be bound up with different aspects of both rendered services and their recipients (a competitive price will not already be enough). So identifying the phenomenon is not less important from very prediction of customers walking away. Seeking the right method on account of the effectiveness, is not imposing great problem. At present classifier ensembles are these methods. They have proved their effectiveness both in the area of the theory [Dietterich 2000; Hansen 1990; Parikh et al. 2006; Shi Xin 2003], as well as real application in different fields [Giacinto, Roli 1999; Hung et al. 2007; Parikh et al. 2005; West et al. 2005]. Trouble with classifier ensemble is in loss of abilities of interpretations of received results. It has two reasons. Random rather than intentional techniques of creating classifier ensemble are the first one, and the second are combinations of the decision of a large number of classifiers into a single decision [Nock 2002]. In the author's opinion contextual classifier ensemble is well managing limitation of classifier ensemble [Jakubczyc 2006; Jakubczyc 2007].

The structure of the article is as follows: short characteristics of the contextual classifier ensembles are included in section 2; in section 3 a description of learning data and their contexts is introduced; the search of the most adequate and comprehensible model of the churn prediction is presented in section 4, and all is closed by the summary.

2. Contextual classifier ensemble

The classifier ensemble is a set of base classifiers which together are solving the problem of the discrimination. There are three basic mechanisms of creating and functioning of the classifier ensemble: creation of base classifiers, selection of base classifiers and combination of classifiers decisions. Applied mechanism of creating ensembles of classifiers is based on data manipulation with the use of internal or external data context [Jakubczyc 2006; Jakubczyc 2007]. The base classifiers are models of decision tree.

Selection of classifiers is the second mechanism. Differently than in the mechanism of combining, about the final decision one classifier chosen in some established way settles from the set of base classifiers is deciding [Giacinto et al. 2000; Giacinto, Roli 1999]. A new method was proposed for selection, which is withdrawing from selection the single classifier. It consists in dynamic creating team of classifiers commissioned for voting depending on contexts in which classified cases are appearing.

Combining single individual decisions of base classifiers into one final decision is the third mechanism. Applied techniques are determined by continuity or discreteness of class values [Kuncheva 2001; Kuncheva, Whitaker 2003; Leung, Parker 2003]. Since decision trees are a chosen algorithm, we are dealing with nominal value of classes. Voting schemata are accessible techniques of the combination of the

decision of base classifiers in this case. A majority voting is most popular (so-called democratic). There are many differently weighed voting schemata too.

The study was carried out with the use of the algorithm of generation of decision trees (version of Quinlan's algorithm C4.5 included in SAS Enterprise Miner 5.2).

3. Data and their contexts

A database of customers of the wired telecommunications company for the years 2003-2006 is the basis of the research. Data after preliminary tidying up is making 5006 records and 51 categories.

For creating the contextual classifier ensemble an identification of contexts of the phenomenon in data analyzed is the base. In this case, after analysis of collected data, it can be assumed that there were following interesting contexts of:

- size of the company determined by average amount of invoices (W: W_m – small, $W_ś$ – average, W_d – big);
- delaying in payments with taking the location into consideration (O: O_t – delay, O_n – lack of delay);
- level of the rotation of customers in provinces (R: R_w – high, $R_ś$ – average, R_n – low);
- indicator of comebacks relative to provinces (P: P_w – high, P_n – low);
- time at the operator (C: C_{m1} – 1 month, C_{m3} – 3 months, C_r – year, C_{l2} – 2 years, C_{l3} – 3 years, C_l – over 3 years).

Not all possibilities of the identification of the context were used up, but it is possible to begin with the assumption that they are paying attention to enough aspects of data gathered together.

4. In quest for the effective model of prediction

Identified contexts served for defining classification tasks. For each of them a decision tree was generated. This way 16 contextual classifiers were obtained. Decision tree learned on the entire data file, called farther W_s, is an additional classifier. So the final number of classifiers is 17. Grouping them up in frames of particular contexts, on account of their complementary character, seems intuitively correct. In this situation 5 classifiers, from which everyone describes entire learning material, are formed. So a set of six competitive classifiers was received (plus W_s), what means that each of them can perform the task of classification independently. Choice of the optimal subset of classifiers in designing classifiers ensembles in the analyzed approach can be treated as a secondary issue. Assuming that the created set of contextual classifiers is an optimal classifier ensemble is in a sense sound.

It is possible to accomplish testing the quality of the created contextual classifier ensemble analyzing the level of the diversity of base classifiers and their classification accuracy. There are many diversity measures [Kuncheva, Whitaker 2003; Ruta,

Gabrys 2002]. Some of the measures work on the whole group of classifiers (e.g. entropy measure Ent, measure of difficulty θ, Kohavi-Wolpert variance KW), whilst other measures consider the classifiers on a pairwise basis and then average the results (e.g. Q Yule statistics, correlation coefficient ρ, disagreement measure D), Two measures were chosen for analysis of pairwise classifiers: Q statistics and the correlation coefficient and one measure KW for the entire group of classifiers. Results of two first measures are presented in Table 1.

Table 1. The values of the Q and ρ measure for competitive contextual classifiers

Q	W_s	R	C	P	O	W	ρ	W_s	R	C	P	O	W
W_s	1	0.83	0.91	0.86	0.90	0.86		1	0.49	0.60	0.53	0.58	0.52
R	.	1	0.84	0.79	0.85	0.82		.	1	0.51	0.48	0.52	0.49
C	.	.	1	0.85	0.96	0.85		.	.	1	0.52	0.73	0.51
P	.	.	.	1	0.89	0.86		.	.	.	1	0.58	0.53
O	.	.	.	.	1	0.86		.	.	.	.	1	0.54
W	.	.	.	.	.	1		.	.	.	.	.	1

Explanations: W_s – complete classifier; W (a size of a company): $W_ś$ – medium, W_m – small, W_d – big; O (delaying in payments with location): O_t – delay, O_n – lack of delay; R (rotation of the customers in provinces): R_w – high, $P_ś$ – medium, R_n – low; P (comebacks relative to provinces): P_w – high, P_n – low; C (time at the operator): C_{m1} – 1 month, C_{m3} – 3 months, C_r – year, C_{l2} – 2 years, C_{l3} – 3 years, C_l – beyond 3 years.

Source: own study.

It is not possible to say that base classifiers are independent because values of Q statistics are high from 0.83 to 0.96. The situation is better in the case of correlation coefficient where the values are fluctuating in limits [0.39; 0.73]. It may be interesting to analyze less correlated pairs of classifiers (below value 0.50), e.g. W_s and R, R and P, R and C. The diversity measure KW for all seven classifiers equals 0.07. It shows little difference between base classifiers of contextual classifier ensemble.

To sum up, one can state that base classifiers are similar to each another so their decisions are wrong on the same cases. Such classifieres are not complementary in their descriptions and they are keeping certain level of the lack of the knowledge about the studied phenomenon or the applied model is not good enough. The lack of clear-cut results confirming the relation between the diversity and the quality of base classifiers do not has to determine lower final quality of classifier ensembles. That is confirmed in many research conducted among others by Shipp and Kuncheva [2002], Rue and Gabrys [2001]. Therefore in the consecutive steps the quality of single contextual classifiers and possibilities of its improvement were checked by applying the contextual classifiers ensembles in different configurations.

The accuracy of single classifiers presented in Table 2 shows almost equal though not very high general level. Accuracy of the prediction of customers which will stay

in the communications company (9999) is very high and is fluctuating from 0.85 for the *R* classifier (of rotation) to 0.99 for the *C* classifier (time at the operator).

However prediction of migrating customers (1111) is positively on the too low level, it is not crossing the 37% value even. Random point out one of two alternatives is more effective because it is taking out ca. 50%.

Table 2. The classification accuracy of single contextual classifiers

	W_s	*R*	*C*	*P*	*O*	*W*
Og.	0.75	0.70	0.72	0.71	0.72	0.74
9999	0.92	0.85	0.99	0.87	0.93	0.88
1111	0.33	0.32	0.02	0.33	0.19	0.37

Explanations: W_s – complete classifier; *W* (a size of a company): $W_ś$ – medium, W_m – small, W_d – big; *O* (delaying in payments with location): O_t – delay, O_n – lack of delay; *R* (rotation of the customers in provinces): R_w – high, $P_ś$ – medium, R_n – low; *P* (comebacks relative to provinces): P_w – high, P_n – low; *C* (time at the operator): C_{m1} – 1 month, C_{m3} – 3 months, C_r – year, C_{l2} – 2 years, C_{l3} – 3 years, C_l – beyond 3 years.

Source: own study.

The accuracy of single classifiers can be improved by applying different classifiers groups and voting schemata. In the case of this research, on account of the transparency of deliberations we decided on the majority voting schema, in which every classifier has one vote at its disposal, and this class which will get the most of them is winning. In the situation of equal amount of votes, a toss is deciding (more about outlines of the voting can be found in [Leung, Parker 2003; Kuncheva et al. 2001]).

The task consists in finding the group of classifiers for which the accuracy of the prediction of migrating customers will be higher than random. On account of exceptionally low prediction value for migrating customers and a small number of classifiers, we decided to analyze all possible configurations. Results of voting groups for the best representatives from every possible cardinality were put in Table 3 (two, three, four, five).

Table 3. Comparing results of the majority voting for the best groups of contextual classifiers

	W_sO	RW	W_sOW	RPO	POW	W_sRCP	*CPOW*	W_sRPOW
Og.	0.74	0.72	0.73	0.68	0.71	0.71	0.71	0.74
9999	0.83	0.81	0.80	0.75	0.78	0.87	0.89	0.84
1111	0.45	0.42	0.55	0.53	0.54	0.38	0.37	0.49

Explanations: W_sO – complete classifier; *W* (a size of a company): $W_ś$ – medium, W_m – small, W_d – big; *O* (delaying in payments with location): O_t – delay, O_n – lack of delay; *R* (rotation of the customers in provinces): R_w – high, $R_ś$ – medium, R_n – low; *P* (comebacks relative to provinces): P_w – high, P_n – low; *C* (time at the operator): C_{m1} – 1 month, C_{m3} – 3 months, C_r – year, C_{l2} – 2 years, C_{l3} – 3 years, C_l – beyond 3 years.

Source: own study.

The level of the improvement in the effectiveness of the prediction of migrating customers is positive. It is fluctuating from 0.18 to 0.35. However, in spite of such a considerable improvement in comparison with single contextual classifiers, only three groups are achieving the effectiveness higher than 0.50. The substantial contexts of the prediction customer churn are: of delaying in payments O, size of companies W, the level of the rotation R and the indicator of comeback P. There are also appearing as far as in four ensembles general classifier W_s. It is worthwhile to notice that the preferred cardinality of the classifier ensemble is low and it is taking out three. Achieved results are not acceptable.

The idea of coping with too high prediction error is referring to principles of creating single contextual classifiers. The contextual classifiers are compositions of complementary classifiers of given context (e.g.: contextual classifier referring to delay O consist of two classifiers which represent delay values of O_t and O_n). Since predictive effectiveness of complementary classifiers can considerably differ, we decided to consider them as separate units. At first the classification effectiveness was calculated (Table 4).

Table 4. The classification accuracy of contextual classifiers

	R_w	$R_ś$	R_n	C_{m1}	C_{m2}	C_r	C_{l2}	C_{l3}	C_l
Og.	0.60	0.74	0.69	0.74	1	0.68	0.73	0.8	0.94
9999	0.66	0.88	0.88	0.95	1	1	1	1	1
1111	0.51	0.29	0.25	0.18	1	0	0	0.1	0
	P_w	P_n	O_n	O_t	W_m	$W_ś$	W_d		W_s
Og.	0.70	0.72	0.72	0.75	0.74	0.91	0.67		0.75
9999	0.90	0.84	0.93	0.91	0.88	0.91	0.65		0.92
1111	0.26	0.39	0.18	0.32	0.4	1	0.71		0.33

Explanations: W_s – complete classifier; W (a size of a company): $W_ś$ – medium, W_m – small, W_d – big; O (delaying in payments with location): O_t – delay, O_n – lack of delay; R (rotation of the customers in provinces): R_w – high, $P_ś$ – medium, R_n – low; P (comebacks relative to provinces): P_w – high, P_n – low; C (time at the operator): C_{m1} – 1 month, C_{m3} – 3 months, C_r – year, C_{l2} – 2 years, C_{l3} – 3 years, C_l – beyond 3 years.

Source: own study.

Received results are pointing at the big diversity of classifiers. Looking only at migrating customers they are able to predict costumer abandoning with the maximum or zero accuracy in frames of one context. It is possible to choose classifiers that support prediction of migrants up from this group of 17 classifiers. Since every complementary contextual classifier is describing only a fragment of teaching material, what is determined by a given context, a basic W_s classifier will be a guarantor of the full description. It will appear in every contextual classifier ensemble.

All possible pair "basic classifier – contextual classifier" were examined. Table 5 contains results about the accepted level of the prediction. On 16 as many as

8 classifier ensembles achieved the high classification accuracy which is higher or equal 0.70.

Table 5. Supporting the general classifier with contextual classifiers

	W_sR_w	W_sC_r	W_sR_n	W_sC_{l3}	$W_sW_ś$	W_sP_w	W_sO_n	W_sO_t
Og.	1,00	0,84	0,85	0,86	0,85	0,83	0,84	0,84
9999	1,00	0,85	0,87	0,89	0,88	0,86	0,87	0,88
1111	1,00	0,82	0,81	0,78	0,75	0,74	0,74	0,70

Explanations: W_s – complete classifier; W (a size of a company): $W_ś$ – medium, W_m – small, W_d – big; O (delaying in payments with location): O_t – delay, O_n – lack of delay; R (rotation of the customers in provinces): R_w – high, $P_ś$ – medium, R_n – low; P (comebacks relative to provinces): P_w – high, P_n – low; C (time at the operator): C_{m1} – 1 month, C_{m3} – 3 months, C_r – year, C_{l2} – 2 years, C_{l3} – 3 years, C_l – beyond 3 years.

Source: own study.

Only one full context is supporting the prediction of migrants up: delaying in payments. Rotation of customers and the time at the operator are representing two values, appropriately out of three and of five values of contexts. The context of a medium size of companies and of a high indicator of comeback gives support too. In the situation of such a high support it is possible to choose one of classifiers which would take final decision. But it can mean the resignation from many additional pieces of information they hold. Treating every pair as base classifier of the contextual classifier ensemble is a good idea which will use all supporting contexts. The final decision is a result of applying majority voting schema. Scores of tests on such a classifier were very high and they were located in a range [0.77; 0.89] for migrants.

In the case when classified new customer is not applying to the context appearing in given folding, the classification task is passed down to the general classifier W_s. In order to make the influence of the general classifier smaller a selection of classifiers voting was suggested. The team entitled to vote consists of classifiers, in which the complementary contextual classifier is active.

In the exceptional situation, when no context describes new case, a general classifier is applicable. The accuracy of the prediction of migrants increased and it fluctuated in the scope 0.80-0.95. Since the effectiveness turned out very much satisfactory research were not carried out with the different cardinality of classifier ensembles respecting the principle of the simplicity of the description.

5. Summary

The research carried out on the prediction of migrating customers of the communications company turned out very much fruitful. First of all a high classification accuracy was achieved without loss of interpretable results, which takes place in the case of applying other rules of creating classifier ensemble (see [Nock 2002]). In-

teresting and effective in effects a resignation from complementary treating classifiers turned out in frames of given of context. Complementary models of the context are supporting basic classifier considerably more effective then classifiers of full context. Work in this field of contextual classifier ensemble is continued.

References

Berry M.J.A., Linoff G. (2000). *Mastering Data Mining*. John Wiley & Sons, New York.

Customer Retention Assessment, KANBAY People & Technology Powering Transformation. www.kanbay.com 11.11.2007.

Dietterich T.G. (2000). Ensemble methods in machine learning. [In:] *Proceedings of 1st International Workshop on Multiple Classifier Systems.* Springer Berlin/Heidelberg pp. 1-15.

Giacinto G., Roli F. (1999). Methods for dynamic classifier selection. [In:] *Proceedings of the 10th International Conference on Image Analysis and Processing ICIAP'99, Venice, Italy.* IEEE Computer Society, Washington, pp. 659-664.

Giacinto G., Roli F., Fumera G. (2000). Selection of image classifiers. *Electronic Letters*, vol. 36, no. 5, pp. 420-422.

Hansen L., Salamon P. (1990). Neural network ensembles. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, pp. 993-1001.

Hung C., Chen J., Wermter S. (2007). Hybrid probability based ensembles for bankruptcy predict. [In:] *Proceedings of the International Conference on Business and Information*, vol. 4/1, Academy of Taiwan Information System Research, Taiwan.

Jakubczyc J.A. (2006). Kontekstowy klasyfikator złożony. [In:] *Nowoczesne technologie informatyczne w zarządzaniu.* Eds. E. Niedzielska, H. Dudycz, M. Dyczkowski. Prace Naukowe Akademii Ekonomicznej nr 1121, AE, Wrocław, pp. 313-322.

Jakubczyc J.A. (2007). Contextual classifier ensembles. [In:] *LNCS 4439 – Business Information Systems.* Ed. W. Abramowicz, Springer, Berlin/Heidelberg, pp. 562-569.

Kuncheva L.I. (2001). Using measures of similarity and inclusion for multiple classifier fusion by decision templates. *Fuzzy Sets and Systems*, vol. 122, no. 3, pp.401-407.

Kuncheva L.I., Bezdek J.C., Duin R.P.W. (2001). Decision templates for classifier fusion: An experimental comparison. *Pattern Recognition*, vol. 34, no. 2, pp. 290-314.

Kuncheva L.I., Whitaker C.J. (2003). Measures of diversity in classifier ensembles. *Machine Learning*, vol. 51, pp. 181-207.

Leung K.T., Parker D.S. (2003). *Empirical Comparisons of Various Voting Methods in Bagging*. SIGKDD'03, ACM, Washington.

Nock R. (2002). Inducing interpretable voting classifiers without trading accuracy for simplicity: Theoretical results, approximation algorithms, and experiments. *Journal of Artificial Intelligence Research*, vol. 17, AI Access Foundation and Morgan Kaufmann Publishers.

Parikh D., Kim M.T., Oagaro J., Mandayam S., Polikarp R. (2006). *Multiple Classifiers for Multisensor Data Fusion*. IEEE – Sensors Application Symposium, Houston.

Parikh D., Stepenosky N., Topalis A., Green D., Kounios J., Clark Ch., Polikar R. (2005). *Ensemble based data fusion for early diagnosis of Alzheimer's disease*. IEEE- Engineering in Medicine and Biology, Shanghai.

Ruta D., Gabrys B. (2001). Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems. [In:] *Proceedings of the 4th International Symposium on Soft Computing*, Academic Press, Paisley, UK.

Ruta D., Gabrys B. (2002). New measure of classifier dependency in multiple classifier systems. [In:] Proceedings of the 3rd International Workshop on Multiple Classifier System, LNCS 2364. Eds. F. Roli, J Kittler. Springer Verlag, pp. 127-136.

Shipp C.A., Kuncheva L.I. (2002). Relationship between combination methods and measures of diversity in combining classifiers. *Information Fusion*, vol. 3, no. 2, pp. 135-148.

Shi Xin Y. (2003). *Feature Selection and Classifier Ensembles: A Study on Hyperspectral Remote Sensing Data*. PhD, The University of Antwerp.

West D., Dellana S., Qian J. (2005). Neural network ensemble strategies for financial decision applications. *Computers and Operations Research*, vol. 32, no. 10.