Churn Prediction Model in Retail Banking Using Fuzzy C-Means Algorithm

Džulijana Popović
Zagrebačka banka d.d., Consumer Finance
Trg bana Josipa Jelačića 10, 10000 Zagreb, Croatia
E-mail: dzulijana.popovic@unicreditgroup.zaba.hr, www.zaba.hr

Bojana Dalbelo Bašić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
E-mail: bojana.dalbelo@fer.hr

Keywords: churn prediction, fuzzy c-means algorithm, fuzzy transitional condition of the first degree, fuzzy transitional condition of the second degree, distance of k instances fuzzy sum

Received: February 22, 2008

The paper presents model based on fuzzy methods for churn prediction in retail banking. The study was done on the real, anonymised data of 5000 clients of a retail bank. Real data are great strength of the study, as a lot of studies often use old, irrelevant or artificial data. Canonical discriminant analysis was applied to reveal variables that provide maximal separation between clusters of churners and non-churners. Combination of standard deviation, canonical discriminant analysis and k-means clustering results were used for outliers detection. Due to the fuzzy nature of practical customer relationship management problems it was expected, and shown, that fuzzy methods performed better than the classical ones. According to the results of the preliminary data exploration and fuzzy clustering with different values of the input parameters for fuzzy c-means algorithm, the best parameter combination was chosen and applied to training data set. Four different prediction models, called prediction engines, have been developed. The definitions of clients in the fuzzy transitional conditions and the distance of k instances fuzzy sums were introduced. The prediction engine using these sums performed best in churn prediction, applied to both balanced and non-balanced test sets.

Povzetek: Razvita je metoda mehke logike za uporabo v bančništvu.

1 Introduction

Due to intensive competition and saturated markets, companies in all industries realize that their existing clients database is their most valuable asset. Retaining existing clients is the best marketing strategy to survive in industry and a lot of studies showed it is more profitable to keep and satisfy existing clients than to constantly attract new ones [1,4,8,11]. Churn management, as the general concept of identifying those clients most prone to switching to another company, led to development of variety of techniques and models for churn prediction. Next generation of such models has to concentrate on the improved accuracy, robustness and lower implementation costs, as every delay in reaction means increased costs for the company [2].

The aim of this study was to show that the data mining methods based on the fuzzy logic could be successfully applied in the retail banking analysis and, moreover, that the fuzzy c-means clustering performed better than the classical clustering algorithms in the problem of churn prediction.

Although the clustering analysis is in fact an unsupervised learning technique, it can be used as the basis for classification model, if the data set contains the classification variable, what was case in this study.

To our best knowledge this is the first paper considering application of fuzzy clustering in churn prediction for retail banking. Studies of churn prediction in banking are very scarce, and the most of papers used models based on logistic regression, decision trees and neural networks [9,11]. Useful literature review of attrition models can be found in [11]. Some of them [9] reported the percentage of correct predictions varying from 14% to 73%, depending on the proportion of churners in the validation set. The others [3] obtained AUC performance in subscription services varying from 69,4% for overall churn to 90,4% but only for churn caused by financial reasons, which is much easier to predict. Results are not perfectly comparable due to differences in churn moment definitions, data sets sizes or industries, but still can provide valuable subject insight.

2 Fuzzy c-means clustering algorithm

Classical clustering assigns each observation to a single cluster, without information how far or near the observation is from all the other possible decisions. This type of clustering is often called hard or crisp clustering [1,10,12]. Two major classes of crisp clustering methods are hierarchical and optimization (partitive) clustering, with number of different algorithms, used in the study.
Based on the fuzzy set theory, firstly introduced by Zadeh in 1965. [5,6,10] and on the concept of membership functions, the fuzzy clustering methods have been developed. In fuzzy clustering entities are allowed to belong to many clusters with different degrees of membership.
Fuzzy clustering of X into p clusters is characterized by p membership functions μ_j, where

$$\mu_j: X \to [0,1], j = 1, \dots, p, \quad (1)$$

$$\sum_{j=1}^{p} \mu_j(x_i) = 1,\ i = 1, 2, \dots, n, \quad (2)$$

and

$$0 < \sum_{i=1}^{n} \mu_j(x_i) < n,\ j = 1, 2, \dots, p. \quad (3)$$

Membership functions are based on a distance function, such that membership degrees express proximities of entities to cluster centers (also called *cluster prototypes*).
The most known method of fuzzy clustering is the fuzzy c-means method (FCM), initially proposed by Dunn, generalized by Bezdek [5] and used in this study.
FCM involves two iterative processes: the calculation of cluster centers and the assignment of the observations to these centers using some form of distance. FCM is attempting to minimize a standard loss function

$$l = \sum_{k=1}^{p} \sum_{i=1}^{n} [\mu_k(x_i)]^m \|x_i - c_k\|^2 \quad (4)$$

from which two fundamental equations necessary to implement FCM are derived [5].

Expression (5) is used to calculate a new cluster center value:

$$c_j = \frac{\sum_i [\mu_j(x_i)]^m x_i}{\sum_i [\mu_j(x_i)]^m} \quad (5)$$

and expression (6) to calculate the membership in the $j-th$ cluster:

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p} \left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}} \quad (6)$$

The symbols in the equations (4), (5) and (6) denote:

l	is the minimized loss value;
p	is the number of fuzzy clusters;
n	is the number of observations in the data set;
$\mu_k( )$	is a function that returns the membership of x_i in the $k-th$ cluster;
m	is the fuzzification parameter;
c_k	is the centre of the $k-th$ cluster;
d_{ji}	is the distance metric for x_i in cluster c_j;
d_{ki}	is the distance metric for x_i in cluster c_k.

3 Churn prediction problem in retail banking and input data set

There is no unique definition of churn problem, but generally, term churn refers to all types of customer attrition whether voluntary or involuntary [1,3]. How to recognize it in practice depends on industry and case. In this study, client is treated as churner if he had at least one product (saving account, credit card, cash loan etc.) at time t_n and had no product at time t_{n+1}, meaning that he cancelled all his products in the period $t_{n+1} - t_n$. If client still holds at least one product at time t_{n+1}, he is considered to be non-churner.
The programs for all research phases, as well as prediction engines, were written in SAS 9.1. [12].

3.1 Input data set

The input data set has 5000 clients, chosen by random sampling from the client population in 2005, aged between 18 and 80 years, preserving the distribution of population according to introduced auxiliary variable which was product level of detection (PLOD). The class imbalance problem [2,9] was solved in the way that precisely 2500 churners and 2500 non-churners entered the final data set, what is in line with results in [2]. Regarding the moment of churn for 2500 churners five sample data sets, with different configuration of churners, have been explored. The analysis showed that the "clear" set, with churners all lost in the same quarter of the year, is best for further clustering. All clients who quitted the relationship with bank in some period, but returned after 6 months or later, were removed from the sample, as the analysis showed they behave similarly to real churners and introduce the noise.

3.2 Variable selection

Not all variables of interest were allowed for study, and the availability of more transactional variables would surely lead to better model performance [11]. This was partially proved through inclusion of derived variables (differences in time, ratios, etc.), what led to the improved accuracy of fuzzy clustering in comparison to clustering results with only original variables. All variables were measured in five equidistant points of time: t_0 to t_4. Preliminary clustering analysis showed that, as far as the original variables are statical in their character, including the values of more than two periods

leads to more noise than to greater precision. All the combinations of two periods were tested and finally the values in t_0 and t_2 were chosen for the further analysis.

Table 1. gives the description of the 73 variables finally chosen for the further work.

time	socio-demographic	banking products in charge	financial	bad-behavior	derived for Δt=t2-t0
t_0	3	16	14	2	6
t_2	—	16	14	2	—
total	3	32	28	4	6

Table 1: Variables for the final FCM.

3.3 Canonical discriminant analysis

Canonical discriminant analysis (CDA) finds the linear combinations of the variables that provide maximal separation between clusters [12]. CDA helped in identifying the variables that best describe each of two classes/clusters: churners and non-churners. In some way CDA confirmed that the combination of $t_0 - t_2$ variable values is more adequate for FCM then the other combinations. All the coefficients and corresponding variables have been carefully examined.

3.4 Detection and removal of outliers

Although the majority of variables is not normally distributed, the standard deviation [10] in combination with CDA and k-means revealed most serious outliers better than other methods.

The most serious outliers were detected for all 73 variables separately and the intersection of those 73 sets was found. For all the outliers the data values have been checked in the data warehouse. The check confirmed that all the data are correct and that the outliers are not the consequence of the errors in database. Top 50 outliers from that intersection were removed from the data set. The outlier removal significantly improved the performance of classical clustering and slightly improved the performance of FCM.

3.5 Results of the hierarchical and crisp k-means algorithm

To prove that fuzzy clustering performs better than the classical methods on the real retail banking data, hierarchical clustering and k-means clustering were done. Applied to all 5000 clients, almost all hierarchical methods, as well as repeated k-means, failed on the same outliers. Most of them separated only one client in the first cluster and all other 4999 were appointed to the second cluster. All the methods were repeated on the data set without top 50 outliers and some of them performed better.

Figure 1 shows the standard measures [7] for comparison of the results in churn prediction, as stated in [2] and Table 2 shows some of the results of classical clustering.

$$\textbf{tp rate}\ (recall, hit\ rate) = \frac{positives\ correctly\ classified}{total\ positives}$$

$$\textbf{fp rate}\ (false\ alarm\ rate) = \frac{negatives\ incorrectly\ classified}{total\ negatives}$$

$$\textbf{accuracy} = \frac{true\ positives + true\ negatives}{positives + negatives}$$

$$\textbf{specificity} = 1 - fp\ rate$$

Figure 1: Common performance metrics calculated from confusion matrix.

CLUSTERING ALGORITHM	STANDARDIZATION METHOD	tp rate (recall)	fp rate	accuracy	specificity
Average Linkage	standard deviation	99,96%	100,00%	50,44%	0,00%
Average Linkage	range	100,00%	99,47%	50,73%	0,53%
Centroid Linkage	standard deviation	99,96%	100,00%	50,44%	0,00%
Centroid Linkage	range	100,00%	99,96%	50,48%	0,04%
Ward's Minimum Variance	standard deviation	84,67%	66,92%	59,11%	33,08%
Ward's Minimum Variance	range	73,58%	60,81%	56,55%	39,19%
Complete Linkage	standard deviation	99,96%	99,92%	50,48%	0,08%
Complete Linkage	range	87,39%	70,07%	58,93%	29,93%
Flexible Beta	standard deviation	81,55%	64,89%	58,55%	35,11%
Flexible Beta	range	72,18%	59,01%	56,73%	40,99%
McQuitty's Similarity Analysis	standard deviation	99,96%	100,00%	50,44%	0,00%
McQuitty's Similarity Analysis	range	98,08%	89,27%	54,81%	10,73%
Median Linkage	standard deviation	99,96%	100,00%	50,44%	0,00%
Median Linkage	range	100,00%	99,96%	50,48%	0,04%
Single Linkage	standard deviation	99,96%	100,00%	50,44%	0,00%
Single Linkage	range	100,00%	99,96%	50,48%	0,04%
Crisp k-means*	standard deviation	100,00%	99,96%	50,02%	0,04%
Crisp k-means	standard deviation	99,88%	80,67%	59,61%	19,33%

**** performed on complete data set, without outlier removal***

Table 2: Results of preliminary classical clustering.

To get the full comprehension on algorithm performance several measures have to be considered simultaneously. Recall rate of 100% means unsuccessful churners recognition, if comes in combination with specificity under 1%. Losing one client causes greater losses for the bank, then investing in marketing campaign for several clients incorrectly classified as possible churners, which means that costs of false negatives are much higher then costs of false positives. In real clients population there are much less positive then negative instances, so liberal classifiers obtaining high recall rate and acceptable specificity are considered successful in business.

4 Model setup and prediction results

FCM has been repeatedly applied on the complete data set and on the data set without top 50 outliers, with 10 different values of the fuzzification parameter m, and different initial cluster seeds. It performed slightly better without outliers, what means that FCM is very robust against outliers' presence. From application point of view that is very good property of FCM, since it will not always be profitable for the bank to detect and remove outliers, not to mention the fact that these outliers are sometimes the most active and profitable clients and they need to be included in the model development. With crisp k-means it would not be possible, because it performed incredibly poorly with these clients.

Data set was splitted into two parts: training set and test set, in three different ratios. The ratio of 90% of clients in the training set and 10% of clients in test set was chosen. According to the values of the membership functions, the clients in fuzzy transitional conditions (FTC) were detected. For that purpose two new definitions were proposed.

Definition 1. *Let p be the number of clusters in the FCM algorithm. Let us denote $max_{j=1}^{p}\{\mu_j(x_i)\} = \mu_{MAX}^1$ and $max_{j=1}^{p}\{\mu_j(x_i) \setminus \mu_{MAX}^1\} = \mu_{MAX}^2$ for the entity x_i. The entity x_i is said to be in the fuzzy transitional condition of the 1^{st} degree if, for arbitrary small $\varepsilon > 0$, holds that $\mu_{MAX}^1 - \mu_{MAX}^2 < \varepsilon$.*

Definition 2. *Let p be the number of clusters in the FCM algorithm. Let us denote $max_{j=1}^{p}\{\mu_j(x_i)\} = \mu_{MAX}^1$. The entity x_i is said to be in the fuzzy transitional condition of the 2^{nd} degree if, for arbitrary small $\varepsilon > 0$, holds that $\mu_{MAX}^1 - \frac{1}{p} < \varepsilon$.*

Subsets of clients in the FTC of both degrees, and with floating ε values, were further analyzed and the information gained from the fact about their membership values helped in explaining their behavior. Four prediction models were developed, based on the main idea of the distance of the new client from the clients in the training data set. For the predictive purpose in the 4^{th} model, the definition of *distance of k instances* (DOKI) *sums* was introduced.

Definition 3. *Let p be the number of clusters in the FCM algorithm and X be the set of n entities with assigned membership values μ_j, $j = 1, \ldots, p$. Distance of k instances sum i.e. $DOKI_j^k(x)$ sum for the new entity x_{n+1} is defined as the sum of membership values $\{\mu_j\}$ in the $j-th$ cluster of the k nearest entities from X, according to distance metric used in FCM.*

Calculation of DOKI sums requires the input parameter k and several different values were applied. Table 3 presents the results of FCM on the training set and prediction engine with DOKI sums applied on balanced and non-balanced test sets. Concept of DOKI sums might seem similar to k nearest neighbors approach, but DOKI sums up values of membership functions and not the pure distances. Recall rate for test sets were even higher then recall rate obtained with FCM on the training set. Improvement in recall was paid in slight decrease in specificity. As mentioned previously, it is more important to hit churners, even if it is paid by hitting some percentage of loyal clients. The cost minimization can be achieved later through more intelligent and multi-level communication channels.

DATA SET	tp rate (recall)	fp rate	accuracy	specificity
training	79,64%	55,61%	62,00%	**44,39%**
test - nonbalanced	87,60%	60,82%	**63,64%**	39,18%
test - balanced	**88,52%**	62,45%	63,04%	37,55%

Table 3: Results of FCM and DOKI prediction model.

5 Conclusions and further work

It is always challenging to deal with real data and business situations, where classical methods can rarely be applied in their simplest theoretical form. The main idea of the study – to prove that fuzzy logic and fuzzy data mining methods can find their place in the reality of retail banking – was completely fulfilled. FCM performed much better than the classical clustering and provided more hidden information about the clients, especially those in fuzzy transitional conditions. Three new definitions were introduced and had the impact on the overall work. Implementation of DOKI sums increased hit rate (recall) by 8,88% in comparison to pure FCM. A lot of work still needs to be done. In the near future every client and every selling opportunity will become important. Methods which require a lot of preprocessing and, above all, removing many outlying clients, will lose the battle with more efficient and robust methods. More accuracy should be obtained through better information exploitation of clients in fuzzy transitional conditions, and not through clients removal. Monitoring clients in FTCs and reacting as they approach to churners could be a way for more intelligent churn management. This requires analysis on larger data sets, including more transactional variables into the model and tuning ε. Model should also include costs of positive and negative misclassifications. Different segments of clients or clients having similar product lines could be modeled

on their own, to find empirically best FCM parameters for each segment/product line.

Acknowledgement

The first author's opinions expressed in this paper do not necessarily reflect the official positions of Zagrebačka banka d.d.

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grant No. 036-1300646-1986 and 098-0982560-2563.

References

[1] Berry J.A.M., Linoff S.G. (2004) *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*, 2nd Ed. Indianapolis: Wiley Publishing, Inc.

[2] Burez J., Van den Poel D. (2008) "Handling class imbalance in customer churn prediction", *Expert Systems with Applications*, In Press, available online 16 May 2008.

[3] Burez J., Van den Poel D. (2008) "Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department", *Expert Systems with Applications* 35 (1-2), pp. 497-514.

[4] Coussement K., Van den Poel D. (2006) "Churn Prediction in Subscription Services: an Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques", *Working Paper 2006/412*, Ghent University.

[5] Cox E. (2005) *Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration*, San Francisco: Morgan Kaufmann Publishers.

[6] De Oliveira, J.V., Pedrycz W. (editors) (2007) *Advances in Fuzzy Clustering and its Applications*, John Wiley & Sons Ltd.

[7] Fawcett T. (2004) ROC Graphs: Notes and Practical Considerations for Researchers, Netherlands: *Kluwer Academic Publishers*.

[8] Hadden J., Tiwari A., Roy R., Ruta D. (2005) "Computer assisted customer churn management: State-of-the-art and future trends", *Computers & Operations Research* 34, pp. 2902-2917.

[9] Mutanen T., Ahola J., Nousiainen S. (2006) "Customer churn prediction - a case study in retail banking", *ECML/PKDD 2006 Workshop on Practical Data Mining: Applications, Experiences and Challenges*, Berlin.

[10] Theodoridis S., Koutroumbas K. (2003) *Pattern Recognition*, 2nd Ed., San Diego, USA: Academic Press, Elsevier Science.

[11] Van den Poel D., Larivière B. (2004) "Customer attrition analysis for financial services using proportional hazard models", *European Journal of Operational Research* 157 (1), pp. 196-217.

[12] Yeo D. (2005) Applied Clustering Techniques Course Notes, Cary NC, USA: SAS Institute Inc.