©2018 Published in 6th International Symposium on Innovative Technologies in Engineering and Science 09-11 November 2018 (ISITES2018 Alanya – Antalya - Turkey)

Determination of Factors Causing Customer Loss in the Banking Sector Using Data Mining Methods

[1]Seher Arslankaya
[1]Faculty of Engineering, Department of Industrial Engineering Sakarya University, Turkey

Abstract

The need for customer satisfaction and continuity is very important in the banking sector as in other sectors. Increasing number of services and products are leading to increased competition in the sector and this causes customers to be more anticipated towards service providers. With globalization, banks have had to make new decisions for customer retention and keeping up with change. Implementing their new decisions requires strengthening their infrastructures and technologies and passing customer-focused systems. In this study, the characteristics of the customers who can leave firms in the banking sector are tried to be determined by using the information of the previous customers. The algorithm that makes the most accurate prediction was tried to be determined by applying ZeroR, OneR, Naive Bayes, J48 and Multilayer Perceptron algorithms to the bank customer data using Weka, one of the data mining software. According to the results of the algorithm, the reasons why the customers leave the bank were tried to be identified..

Keywords: Weka, data mining, banking sector, customer loss

1. Introduction

Until recently, the financial sector was either inadequate to meet the needs and desires of its customers or ignored them. However, nowadays, the businesses have embarked on new quests since the ones that neglect customers have lost their chances of competition. In new quests, issues such as customer focused culture, organizational structure, customer profitability, customer value, continuity of customer relations, development of customer information systems, customer differentiation and customer loyalty are influential (Bryan and Stone, 2002). In order to make improvements on all these issues, historical data must be used.

Customer relationship management is a management strategy that enables individualized relationships with customers and aims at increasing customer satisfaction and maximizing their profits besides providing the highest level of customer service (Eichorn 2004). It is an effective strengthening strategy for product, service and distribution systems to identify and meet the changing needs of customers (Nancarrow vd., 2003). Increasing competition in today's banking sector has brought up the effective management of existing and possible customer relationships. For customers, not only the service they receive, but also post-service relationships are important. The retention of existing customers and the addition of new customers to them depend on the value given to the customer. The first concept that a large quantity of data evokes is “Data Mining”. Data Mining is a process that explores patterns and relationships in data with the use of many analysis tools and uses them to make valid estimates (Koyuncugil, Özgülbaş, 2009).

*Corresponding author: Address: Faculty of Engineering, Department of Industrial Engineering Sakarya University, 54187, Sakarya TURKEY. e-mail address: aseher@sakarya.edu.tr, Phone: +902642955686

There are many studies in the literature on data mining. Emel and Taşkın (2005), conducted a sales analysis for a retailer business that included detailed and relative measurement results, using the database of the business with personalized sales movements according to the customer. Savaş and Topaloğlu (2011) have developed a program that creates data bases by extracting the gravity data of different GSM networks from mobile phones. Terzi (2012), used data mining in the detection of fraud and irregularities in business during the audit and integrated it with the audit. Akçetin ve Çelik [2014] determined the most suitable decision tree method in terms of accuracy and classification time by comparing their performance in order to identify spam e-mails of decision tree algorithms. Narlı et al (2014) identified learning styles of elementary school mathematics teacher candidates and explored the relationship between these styles using data mining techniques.

In this study, the second part is focused on data mining, the method used in the study. In the third part, data mining was conducted in Weka using the old customer information in the finance sector. In the fourth chapter statistical analysis of the model was made and statistical results were explained. The fifth section contains the discussion and suggestions, and the sixth section is the results section.

2. Materials and Method

2.1. Data Mining

Data mining is a data analysis technique that examines the relationships in very large amounts of data and helps to find the link between them, and allows the retrieval of information that is hidden in database systems (Kalikov, 2006). The most important feature of data mining for businesses is to determine similar trends and behavior patterns among data groups. At the same time, this process can be put into practice in an automatized way. This function is used extensively in marketing activities especially for target markets (İnan, 2003). Han and et al. (2001) and Delen et al. (2005) have listed the steps of data mining in the following way. 1-Data Cleaning 2-Data Integration 3-Data Selection 4-Data Transformation 5-Data Mining 6-Pattern Evaluation 7-Information Presentation. WEKA, incorporating machine learning algorithms, is an open source data mining program with a functional graphical interface developed by Waikato University in New Zealand (Witten vd., 2011). WEKA includes various data preprocessing, classification, regression, clustering, association rules and visualization tools.

2.2. Classification Algorithms

Classification algorithms can be listed as follows:

1. ZeroR Algorithm: It focuses on the ratio between the data available and the results, and the highest result is used as the predictor of the next data.
2. OneR Algorithm: It decides to classification based on a single feature.
3. Naive Bayes Algorithm: This algorithm, a statistical classification method, is a method that must be calculated again and again in dynamic systems (Seker, 2015).

4. Decision trees: J48 Algorithm: This algorithm called J48 in Weka is an algorithm that makes a classification with a distribution from top to bottom by classifying the data.
5. Artificial neural networks: These are computer systems which can learn events by doing "machine learning" and using realistic examples and they can be applied successfully in some subjects such as learning, associating, classifying, generalizing, characterizing, optimizing and estimating similar to the functional characteristics of the human brain (Öztemel,2006, Patterson vd., 2008; Hall vd., 2009).

3. Current situation analysis in finance sector

Nowadays, with the awareness of the fact that the businesses ignoring customers begin to lose their chances of competition, the development of customer information systems, the continuity of customer relations and customer loyalty have gained importance. Customer loyalty, which is also very important in the banking sector, is rapidly declining because it is easily possible for customers to move from one bank to another. Customer losses should be well analyzed to prevent existing customers from being directed to competitor firms and to maintain existing customer loyalty. In banking sector, competitiveness advantage will be achieved when the abandonment factors are determined, studied in detail and the weaknesses are improved.

3.1. Procurement and Extraction of Data to be used in the Problem

The data used in the study was taken from "Kaggle," a network of San Francisco-based data scientists, as well as a rich open dataset platform. The data set had a total of 10,000 customers and each customer had 14 qualities. The 14th quality indicated whether the customer left the firm or not. The number of qualifications was reduced to 10 by selecting the data that would be useful for the study. These qualifications indicated the credit score, country, gender, age, duration of membership, balance, number of accounts, whether they had credit cards, whether they were active members, and whether they had left the firm.

3.2. Application of ZeroR Algorithm

When the ZeroR algorithm is selected and run in the Weka program, the algorithm results are as shown in Fig. The ZeroR algorithm understood that most customers did not leave the bank and predicted that none of the following customers will abandon the bank.

3.3. Application of OneR Algorithm

When the classification is made with the OneR algorithm, it is found that the “number of products” is the one that best describes the output variable in the data set, that is, the most closely related to whether or not customers leave the bank. The algorithm predicted which customers are most likely to abandon by looking at this quality. Table 1 shows the OneR algorithm results of the number of products quality. It can be said that customers with accounts of 2.5 and below are satisfied with the bank, customers with accounts of 2.5 and above are not satisfied with the bank and left the bank. In order to find the most effective quality after the "Number of products", the number of product was removed from the

"preprocess" section and the algorithm was run again. As a result, it is seen that the age factor is also important in bank abandonment.

Table 1. OneR algorithm results of the number of accounts quality

Account number	1-2,5	2,5-4
Class	0	1

Table 2 shows the effect of age on customers' abandonment of the bank. It can be seen that the customers under the age of 48,5 are loyal to the bank and the customers between 48,5-57,5 years have left the bank.

Table 2. The OneR algorithm result of age quality

Age	0 - 48,5	48,5 - 57,5	57,5 - 59,5
Class	0	1	0

3.4. Naive Bayes Algorithm Application

As seen in the results, 82.66% of the dataset is correctly classified. While 8266 customers were in the expected behavior, the predictions of abandonment for the remaining 1734 customers were incorrect. Although membership of 1504 customers is expected to continue, their membership has been terminated and though 230 customers were estimated to terminate their membership, their membership continued. As seen in Table 3, when customers are compared in terms of the countries they live in, it is understood that most of the customers who have left the bank live in Germany. Customers living in Spain are loyal to the bank compared to those living in other countries.

Table 3. Naive Bayes algorithm result of Geography Quality

Current Status / Geography	Not abandoned	Abandoned
France	4205	811
Spain	2065	414
Germany	1696	815
TOTAL	7966	2040

As seen in Table 4, when customers leaving the bank are ranked according to their gender qualities, it is understood that female customers are more likely to leave the bank than male customers.

Table 4. Naive Bayes algorithm result of Gender Quality

Current Status / Gender	Not abandoned	Abandoned
Woman	3405	1140
Man	4560	899
TOTAL	7965	2039

3.5. J48 Algorithm Application

Figure 1 show that J48 algorithm correctly predicts 8552 out of 10000 estimates. To see how Weka's decision tree is created, right click on the corresponding process in the "result list" section and select "visualize tree". The decision tree of the dataset is as shown in Figure 2. When Figure 2 is examined, the decision tree model J48 algorithm is used in Weka. When the results are examined, the customer loss effects of the number of accounts, age, country, and gender characteristics that are understood in the results obtained by the OneR algorithm and the Naive Bayes algorithm are more clearly understood in the decision tree. The J48 results, which provide significant visibility into decision making, increase the reliability of results with other algorithms. Unlike other algorithms, the J48 algorithm evaluated all qualities at the same time.

```
Time taken to build model: 0.47 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8552               85.52   %
Kappa statistic                          0.4869
Mean absolute error                      0.2113
Root mean squared error                  0.3439
Relative absolute error                 65.128  %
Root relative squared error             85.395  %
Total Number of Instances            10000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,954    0,530    0,876      0,954    0,913      0,503    0,792     0,904     0
                 0,470    0,046    0,722      0,470    0,569      0,503    0,792     0,574     1
Weighted Avg.    0,855    0,432    0,844      0,855    0,843      0,503    0,792     0,836

=== Confusion Matrix ===

    a    b   <-- classified as
 7595  368 |    a = 0
 1080  957 |    b = 1
```

Figure 1. Result of J48 algorithm

Some of the conclusions that can be drawn when the decision tree is examined can be listed as follows:

- It is seen that 136 of the non-active customers, between the ages of 44 to 50, who used credit cards abandoned the bank.
- It is understood that only 14 of the customers under 43 who have been customers for longer than a year and use credit cards have left the bank.
- 9 of the male customers, whose age is older than 67 and who have a credit score less than 789, left the bank.

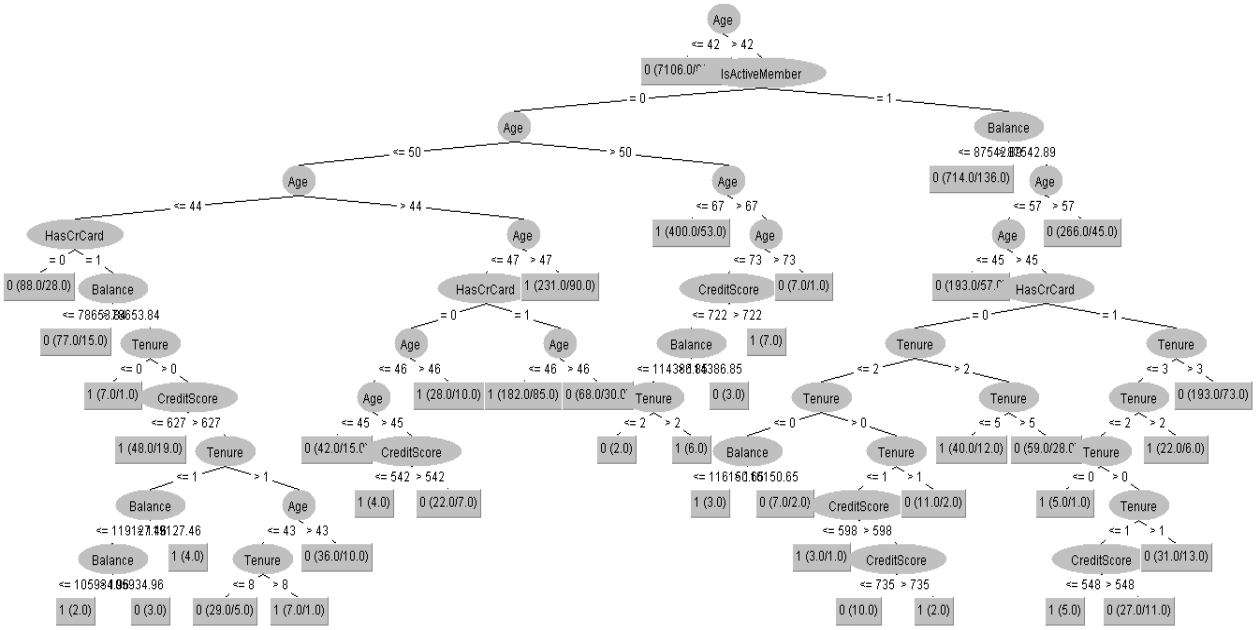

Figure 2. Decision Tree

3.6. Multilayer Perceptron Application

At this stage, experiments are made for the architecture of the artificial neural network. In this study, a 4-layer network topology gives the best result. As shown in Figure 3 the network consists of an input layer, 2 hidden layers and an output layer.

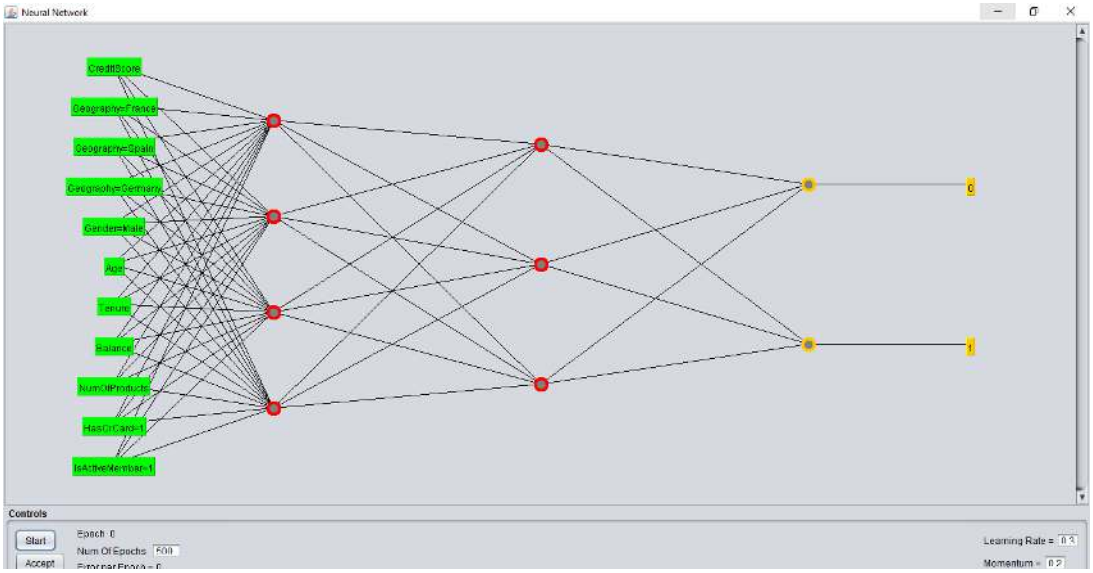

Figure 3 Layered artificial neural network

4. Evaluating Model Success

The basic concepts used to evaluate model performance are error rate, precision, sensitivity and F-criterion. The success of the model is related to the quantities of the number of samples assigned to the correct class and the number of samples assigned to the incorrect class [Coşkun, 2010].

4.1. Statistical Results of Classification Algorithms

The statistical results held in Weka are shown below (Table 5).

Tablo 5. Confusion matrix algorithm

Konfüzyon Matrisi			
	a	b	Sınıfları
ZeroR	7963	0	a=0
	2037	0	b=1
OneR	7917	46	a=0
	1757	280	b=1
NaiveBayes	7733	230	a=0
	1504	533	b=1
J48	7595	368	a=0
	1080	957	b=1
Multilayer Perceptron	2581	133	a=0
	373	333	b=1

a: TP (True Positive) sample number b: FN (False Negative) sample number
c: FP (False Pozitive) sample number d: TN (True Negative) sample number

It is seen that most of the customers do not leave the firm because of the large number of "0" class in the matrix in Table 5. The results show that the loss of the customer in the problem, the class of “1”, cannot be underestimated. The matrix, on the other hand, shows sample numbers that are classified correctly and incorrectly. When the confusion matrix is examined;

• In ZeroR algorithm the ones desired to be a are analyzed as 7963 whereas the ones that became b are analyzed as 2037. Incorrectly classified ones are b.
• However, in OneR algorithm the ones desired to be a become 7917 a, the ones desired to be b become 280 b. Here, 1757 desired b became a and 46 desired a became b as a result of misclassification.
• As to Navie Bayes algorithm, the ones desired to be a become 7733 a, the ones desired to be b become 1242 b. When it comes to misclassified ones, 1504 desired b became a and 230 desired a became b.
• As for J48 algorithm, the ones desired to be a become 7595 a, the ones desired to be b become 957 b. As to misclassified ones, 1080 desired b became a and 368 desired a became b.
• In Multilayer Perceptron algorithm, the ones desired to be a become 2581 a, the ones desired to be b become 333 b. 373 desired b became a and 133 desired a became b as a consequence of misclassification.

In Table 6, layered cross-validation results are given as a result of applying ZeroR, OneR, Naive Bayes, J48 and Multilayer Perceptron classification algorithms.

• Multilayer Perceptron gives the best result with 80,7% according to correctly classified samples. Then J48, Naive Bayes and OneR and ZeroR algorithms gave good results respectively.
• Examples of incorrectly categorized classifications are given in the table to complement the correct classification examples.
• When the Kappa statistic results are examined, the highest result is the Multilayer Perceptron algorithm with 0.4973.

According to these results, it is decided that the most reliable algorithm is the Multilayer Perceptron algorithm with an accuracy of 85.7%

Table 6.Layered cross-validation

Layered Cross Validation										
Summary	**ZeroR**		**OneR**		**Naive Bayes**		**J48**		**Multilayer Perceptron**	
Correct classified samples	7963	79,6%	8197	81,9%	8266	82%	8552	85%	2914	85,7%
Incorrectly categorized examples	2037	20,3%	1803	18%	1734	17%	1448	14%	486	14,3%
Kappa statistic	0		0,1915		0,3034		0,4869		0,4973	
Average absolute error	0,3245		0,1803		0,2493		0,2113		0,1953	
Square root mean error	0,4027		0,4246		0,3557		0,3439		0,3308	
Relative absolute error	100 %		55,5707 %		76,8419 %		65,128%		59,9702%	
Root relative error	100 %		105,4299 %		88,3133 %		85,395%		81,5451%	
Total number of cases	10000		10000		10000		10000		3400	

5. Discussion and Suggestions

Bank employees should work to increase customer satisfaction in order to prevent customer loss. For this reason, it is necessary to investigate the reasons why customers who have a large number of accounts are not satisfied with the bank. The reasons of general dissatisfaction in Germany should be analyzed well and solved in the light of survey results and complaints to customer services. The best aspects of Spanish banks that are different from others should be examined and tried to be applied in the branches of other countries. Special campaigns should be organized to eliminate the tendency of female customers to leave. The reasons for 44-50 aged customers leaving the bank should be investigated and campaigns should be organized to

make them active members. Studies should be carried out on this policy to correct the problems by observing whether there are problems that trigger customer abandonment except number of accounts, geography, gender and age factors mentioned above.

6. Results

- In this study, customer loss analysis was conducted in the banking sector with Weka, one of the data mining programs. Information on data mining and Weka is provided. The working principles and applications of ZeroR, OneR, Naive Bayes, J48 and Multilayer Perceptron algorithms used in Weka are shown respectively. As a result, the causes of the customer loss in the problem are revealed.
- As a result of the analysis, the most accurate and high values are given by the Multilayer Perceptron algorithm, which is one of the classification algorithms.
- Multilayer Perceptron algorithm is followed by J48, Naive Bayes, OneR and ZeroR algorithms.
- ZeroR algorithm showed that the majority of the end customers were loyal to the bank.
- As a result of the analysis of the OneR algorithm, it is concluded that the number of accounts is the most important quality for customer loss.
- As a result of the OneR algorithm, it is understood that customers with 1 or 2 accounts are satisfied with the bank, and customers with 3 or 4 accounts are not loyal to the bank.
- According to the results of the Naive Bayes algorithm, it is understood that the majority of customers who abandon their banks are from Germany when compared to the country they live in. It is seen that the customers living in Spain are more satisfied with the bank than the ones living in other countries.
- Another consequence of the Naive Bayes algorithm is that female customers are more likely to leave the bank than male customers when the customers who left the bank are compared according to their genders.
- As a result of the J48 algorithm, it is seen that 136 of the non-active customers, between the ages of 44 to 50, who used credit cards abandoned the bank.
- Another result of the J48 is that only 14 of the customers under 43 who have been customers for longer than a year and use credit cards have left the bank.
- According to J48 algorithm, 9 of the male customers, whose age is older than 67 and who have a credit score less than 789, left the bank.

References

[1] Bryan F, Stone M, CRM in Financial Services: A Practical Guide to Making Customer Relationship Management Work, Kogan Page Limited, Milford, Ct, USA, 2002.

[2] Eichorn FL, Internal Customer Relationship Management (IntCRM) A Framework for Achieving Customer Relationship Management from the Inside Out, Problems and Perspectives in Management 1, (2004), ss. 154 177.

[3] Nancarrow C, Rees S, Stone M, New Directions in Customer Research and the Issue of Ownership: A Marketing Research Viewpoint, Database Marketing- Customer Strategy Management, (2003), Vol. 11, ss. 26 39.

[4] Koyuncugil, A, S, Özgülbaş N, Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları, bilişim teknolojileri dergisi, cilt: 2, sayı: 2, Mayıs 2009
[5] Emel G.G. Taşkın Ç., Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması, Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi Cilt: 6 Sayı: 2 Aralık 2005
[6] Savaş S., Topaloğlu N., Veri Madenciliği Yöntemi İle GSM Şebekelerinin Performans Analizi, Gazi Üniv. Müh. Mim. Fak. Der. Cilt 26, No 4, 741-751, 2011
[7] Terzi S., Hile ve Usulsüzlüklerin Tespitinde Veri Madenciliğinin Kullanımı, The Journal of Accounting and Finance April / 2012
[8] Akçetin E., Çelik U., The performance benchmark of decision tree algorithms for spam e-mail detection , İnternet uygulamaları ve Yönetim Dergisi, Cilt 5, Sayı 2, 2014
[9] Narli S., Aksoy E., Ercire Y E., Investigation of Prospective Elementary Mathematics Teachers' Learning Styles and Relationships between Them Using Data Mining, International Journal of Educational Studies in Mathematics, 2014, 1 (1), 37-57, ISSN: 2148-5984
[10] Kalıkov A, Veri Madenciliği ve Bir E-Ticaret Uygulaması, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, 2006
[11] İnan, O, Veri Madenciliği, Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, 2003.
[12] Han J, Kamber M, Pei j, Data Mining Concepts and Techniques, Morgan Kaufmann, Elsevier, London 2001
[13] Delen D, Walker G, Kadam A, Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine, 34(02), s (113-127), 2005.
[14] Witten I H, Frank E, Hall M A, Data mining: practical machine learning tools and techniques, ISBN: 9780128043578, Morgan Kaufmann, Elsevier, London, 2011.
[15] Seker S E, Weka İle Veri Madenciliği, ISBN: 9781311400574, Bilgisayar Kavramları Yayınları, Ataşehir, İstanbul, 2015.
[16] Öztemel E, Yapay Sinir Ağları, Papatya Yayınları, s.29, 2006.
[17] Patterson D, Liu F, Turner, D, Concepcion A, Lynch, R, Performance Comparison of the Data Reduction System, Proceedings of the SPIE Symposium on Defense and Security, Mart, Orlando, FL, 2008.
[18] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I H, The WEKA Data Mining Software: An Update, ACM SIGKDD Explorations Newsletter, 11(1), 10–18, 2009