A Neuro-Fuzzy Classifier for Customer Churn Prediction

Hossein Abbasimehr
K. N. Toosi University of Tech
Tehran, Iran

Mostafa Setak
K. N. Toosi University of Tech
Tehran, Iran

M. J. Tarokh
K. N. Toosi University of Tech
Tehran, Iran

ABSTRACT
Churn prediction is a useful tool to predict customer at churn risk. By accurate prediction of churners and non-churners, a company can use the limited marketing resource efficiently to target the churner customers in a retention marketing campaign. Accuracy is not the only important aspect in evaluating a churn prediction models. Churn prediction models should be both accurate and comprehensible. Therefore, Adaptive Neuro Fuzzy Inference System (ANFIS) as neuro-fuzzy classifier is applied to churn prediction modeling and benchmarked to traditional rule-based classifier such as C4.5 and RIPPER. In this paper, we have built two ANFIS models including ANFIS-Subtractive (subtractive clustering based fuzzy inference system (FIS)) and ANFIS-FCM (fuzzy C-means (FCM) based FIS) models. The results showed that both ANFIS-Subtractive and ANFIS-FCM models have acceptable performance in terms of accuracy, specificity, and sensitivity. In addition, ANFIS-Subtractive and ANFIS-FCM clearly induce much less rules than C4.5 and RIPPER. Hence ANFIS-Subtractive and ANFIS-FCM are the most comprehensible techniques tested in the experiments. These results indicate that ANFIS shows acceptable performance in terms of accuracy and comprehensibility, and it is an appropriate choice for churn prediction applications.

General Terms
Data Mining & Churn

Keywords
Churn Prediction, Data mining, ANFIS, Fuzzy C-means, Subtractive clustering.

1. INTRODUCTION
In recent years, Due to the saturated markets and competitive business environment, Customer churn becomes a focal concern of most firms in all industries. Neslin et al. [1] defined customer churn as the tendency of customers to stop doing business with a company in a given time period. Churn prediction is a useful tool to predict customer at churn risk. Technically spoken, the purpose of churn prediction is to classify the customers into two types: customers who churn (leave the company) and customer who continue doing their business with company [2]. By accurate prediction of churners and non-churners, a company can use the limited marketing resource efficiently to target the churner customers in a retention marketing campaign.

Gaining a new customer costs 12 times more than retaining the existing one [3]; Therefore, a small improvement on the accuracy of churn prediction can result a big profit for a company [4].

Data mining techniques had been used widely in churn prediction context such as support vector machines (SVM) [5, 6, 7], decision tree [8], artificial neural network (ANN) [9, 10], Logistic regression [11, 12]. Accuracy is not the only important aspect in evaluating a churn prediction models. Churn prediction models should be both comprehensible and accurate. Comprehensibility of model causes it to reveal some knowledge about churn drivers of customers. Such knowledge can be extracted in the form of “if then” rules which allows developing a more effective retention strategy. In this study we apply Adaptive Neuro Fuzzy Inference System (ANFIS) as neuro-fuzzy classifier for customer churn prediction. Neuro-fuzzy systems have been deployed successfully in many applications, and yields a rule set that is derived from a fuzzy perspective inherent in data. Indeed, the main objective of this study is to compare the ANFIS as neuro fuzzy classifier with two states-of-the-art crisp classifiers including C4.5 and RIPPER. Furthermore, we introduce generating fuzzy inference system using fuzzy C-means clustering.

The reminder of this paper is organized as follows. Firstly, Executed methods are described in section 2. In section 3, the data preprocessing, the evaluation metrics and model building are described. The results of experiments are analyzed in section4. Conclusions are considered in section 5.

2. METHODS
2.1 Fuzzy c-means (FCM) clustering algorithm
Fuzzy c-means (FCM) is a data clustering method wherein each data point belongs to a cluster to some degree that is specified by a membership grade. This method was originally introduced by Jim Bezdek in 1981[13].

Suppose a collection of *n* data point $\{x_1, \ldots, x_n\}$ in an p-dimensional space.

The unknowns in FCM clustering are:

1- A fuzzy c-partition of the data, which is a c x n membership matrix U= [u(ik)] with c rows and n columns. The values in row *i* give the membership of all *n* input data in cluster for *k*=1 to *n*; the *k*th column of U gives the membership of vector *k* in all c cluster for *i*=1 to *c*. each of the entries in *U* lies in [0, 1]; each row sum is greater than zero; and each column of sum equals 1.

2- The second set of unknowns is a set of *c* cluster centers, arrayed as the *c* columns of a *p* x *c* matrix *V*. these cluster centers are data point in the input space of *p*-tuples. Pairs (*U*, *V*) of coupled estimates are found by alternating optimization through the first-order necessary conditions for *U* and *V*. the objective function of FCM is as follows.

FCM performs the clustering with the aim of minimization the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij}{}^m \left\| x_i - c_j \right\|^2, \quad 1 \le m \le \infty \quad (1)$$

Where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of p-dimensional data, c_j is the p-dimensional center of the cluster, and $\|*\|$ is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - x_j\|}{\|x_i - c_j\|} \right)^{\frac{2}{m-1}}}, \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}{}^m x_i}{\sum_{i=1}^{N} u_{ij}{}^m} \quad (2)$$

This iteration will stop when $max_{i,j}\left\{ \left| u_{ij}{}^{(k+1)} - u_{ij}{}^{(k)} \right| \right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, while k is the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

2.2 Subtractive clustering

Subtractive clustering is one of the automated data-driven based methods for constructing the primary fuzzy models proposed by chiu [14]. It is an extension of the Mountain Clustering in traduced by Yager and Filev [15]. This method avoids from rule-base explosion problem. It is a fast, one-pass algorithm for estimating the number of clusters and the cluster centers in a set of data. The main processes of subtractive clustering are as follows:

Consider a collection of m data point $\{x_1, \ldots, x_m\}$ in an N-dimensional space. The algorithm assumes each data point is a potential cluster center and calculates some measure of potential for each of them according to Eq.(3)

$$P_i = \sum_{j=1}^{m} exp^{-\alpha \|x_i - x_j\|} \quad (3)$$

Where $\alpha = 4/{r_a}^2$ and $r_a > 0$ defines the neighborhood radius for each cluster center. After calculating the potential for each vector, the one with the higher potential is selected as the first cluster center. Let $x_1 *$ be the center of the first group and $P_1 *$ its potential. Then the potential for each x_i is reduced according to Eq.(4)

$$P_i = P_i - P_1 * \; exp^{-\beta \|x_i - x_1 *\|} \quad (4)$$

Also $\beta = 4/{r_b}^2$ and $r_b > 0$ represent radius of neighborhood for which considerable potential reduction will happen. $r_b = 1.25\, r_a$ is regularly chosen to avoid obtaining closely spaced cluster centers.

After clustering, the clusters' information is used for determining the initial number of rules and antecedent membership function that is used for identifying the Fuzzy Inference System (FIS).

2.3 Adaptive Neuro Fuzzy Inference System (ANFIS)

Fuzzy logic (FL) and fuzzy inference systems (FIS), first proposed by Zadeh [16], provide a solution for making decisions based on vague, ambiguous, imprecise or missing data. FL represents models or knowledge using IF–THEN rules in the form of ''if X and Y then Z''. A fuzzy inference system mainly consists of fuzzy rules and membership functions and fuzzification and de-fuzzification operations. By applying the fuzzy inference, ordinary crisp input data produces ordinary crisp output, which is easy to be understood and interpreted. A more generalized description of fuzzy problems and uncertainty is provided in [17].

There are two types of fuzzy inference systems that can be implemented: Mamdani type and Sugeno type [18, 19]. Because the Sugeno system is more compact and computationally efficient than a Mamdani system, it lends itself to the use of adaptive techniques for constructing the fuzzy models.

A fuzzy rule in a Sugeno fuzzy model has the form of, *if* x *is* A *and* y *is* B *then* z=f(x, y) where A and B are input fuzzy set in antecedent and usually z=f(x, y) is a zero or first order polynomial function in the consequent.

Fuzzy reasoning procedure for the first order Sugeno Fuzzy Model is shown in Figure 1(a).

In order for a FIS to be mature and well established so that it can work appropriately in prediction mode, its initial structure and parameters (linear and nonlinear) need to be tuned or adapted through a learning process using a sufficient input-output pattern of data. One of the most commonly used learning systems for adapting the linear and nonlinear parameters of an FIS, particularly the first-order Sugeno fuzzy model, is the ANFIS. ANFIS is a class of adaptive networks that are functionally equivalent to fuzzy inference systems [20].

ANFIS architecture:

Assume a fuzzy inference system with two inputs x, y and one output z with the first order of Sugeno Fuzzy Model. Fuzzy rule set with two fuzzy if-then rules are as follows:

If x *is* A_1 *and* y *is* B_1, *then* $f_1 = p_1 x + q_1 + r_1$.

If x *is* A_2 *and* y *is* B_2, *then* $f_2 = p_2 x + q_2 + r_2$.

Where (p_1, q_1, r_1) and (p_2, q_2, r_2) are parameters of output functions.

As it is shown in Figure 1(b), we can implement the reasoning mechanism into a feed forward neural network with supervised learning capability, which is known as ANFIS architecture. The ANFIS has the following layers as illustrated in figure1(b).

Layer 0: it consists of plain input variable set.

Layer 1: The node function of every node i in this layer take the form [20]:

$$O_i^1 = \mu A_i\,(x) \quad (5)$$

Where x is the input to node *i*, μA_i is the membership function (which can be triangular, trapezoidal, Gaussian functions or other shapes) of the linguistic label A_i associated with this node. In other words, O_i^1 is the membership function of A_i and it specifies the degree to which the given x satisfies the quantifier A_i .

In this study, the Gaussian-shaped MFs defined below are utilized:

$$\mu A_i\,(x) = \exp\left(\frac{(x-c_i)^2}{2\sigma_i{}^2}\right) \qquad (6)$$

Where $\{c_i\,, \sigma_i\}$ are the parameters of the MFs governing the Gaussian functions. The parameters in this layer are usually referred to as premise parameters.

Layer 2. Every node in this layer multiplies incoming signals from layer 1and send product out as follows:

$$w_i = \mu A_i\,(x) \times \mu B_i\,(x) \qquad (7)$$

Where the output of this layer w_i represents the firing strength of a rule.

Layer 3. Every node *i* in this layer, determines the ratio of the ith rule's firing strength to the sum of all rules' firing strengths as:

$$\overline{w_i} = \frac{w_i}{w_1 + w_2} \text{ i=1, 2.} \qquad (8)$$

Where the output of this layer represents the normalized firing strengths.

Layer 4: Every node *i* in this layer is an adaptive node with a node function of the form

$$O_i^4 = \overline{w_i}\ f_i = \overline{w_i}\ (p_i\,x + q_i\,x + r_i\,) \qquad (9)$$

Where $\overline{w_i}$ is the output of layer 3, and $\{p_i\,, q_i\,, r_i\,\}$ is the parameter set. Parameters in this layer are referred to as the consequent parameters.

Layer 5: this layer consists of one single node that computes the overall output as the summation of all incoming signals from layer 4 as

$$\text{Overall output} = \sum_i \overline{w_i}\ f_i = \frac{\sum_i \overline{w_i}\ f_i}{\sum_i w_i} \qquad (10)$$

Bothe premise and consequent parameters of the ANFIS should be tuned, using a learning algorithm to optimally the relationship between input space and output space. Basically, ANFIS takes the initial fuzzy model and tunes it by means of a hybrid technique combining gradient descent back-propagation and mean least squares optimization algorithms. At each epoch, an error measure, usually defined as the sum of the squared difference between actual and desired output, is reduced. Training stops when either the predefined epoch number or error rate is obtained. There are two passes in the hybrid learning procedure for ANFIS. In the forward pass of the hybrid learning algorithm, functional signals go forward till layer 4 and the consequent parameters are identified by the least-squares estimate. In the backward pass, the error rates propagate backward and the premise parameters are updated by the gradient descent method.

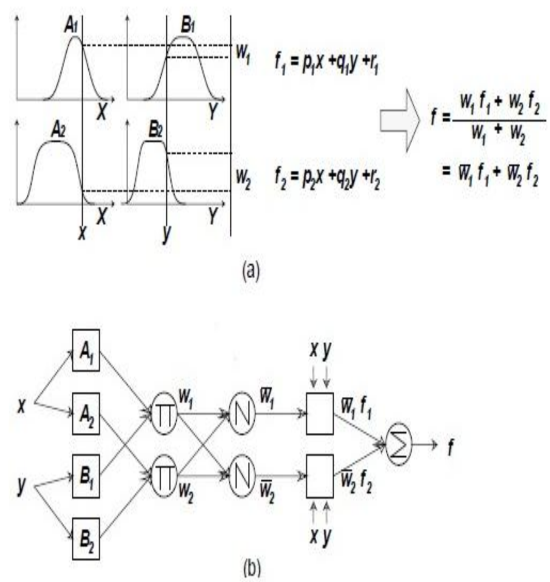

Figure 1:(a) the sugeno fuzzy model reasoning (b) Equivalent ANFIS structure[20]

3. EMPIRICAL ANALYSIS

3.1 Dataset

All algorithms used in this paper are applied on a publicly available dataset downloaded from the UCI Repository of Machine Learning Databases at the University of California, Irvine[1]. The data set contains 20 variables worth of information about 5000 customers, along with an indication of whether or not that customer churned (left the company). The proportion of churner in the dataset is 14.3%. For a full description of the dataset, one may refer to [21]. We first split the data set into 67%/33% training / test set split. The proportion of churners was oversampled in order to give the predictive model a better ability of discerning discriminating patterns. Therefore the proportion of churner and non-churner in training data set is 50%|50%. The test set was not oversampled to provide a more realistic test set; the churn rate remained 14.3%. All models constructed during this work are evaluated on this test set.

3.2 Data preprocessing

Data preprocessing is an essential phase in data mining. Low quality data will lead to low quality mining results. Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining. There are a number of data preprocessing techniques such as data cleaning, data transformation, data integration, data reduction [22]. In this paper, we have done feature subset selection to remove irrelevant attributes from dataset. Furthermore, we have used sampling techniques in order to make balance between positive and negative classes.

[1] http://www.ics.uci.edu/∼mlearn/MLRepository.html

3.3 Feature selection

We have used the PART (partial decision tree) algorithm, a novel data mining techniques for feature subset selection. This algorithm combines the divide-and-conquer strategy for decision tree learning with the separate-and-conquer one for rule learning. A detailed description about the PART algorithm is given in [23]. Berger, et al.,(2006) [24] have introduced feature selection by using PART algorithm. They showed that classifiers show comparable performance in their classification task when applied to the feature subset selected by using the PART algorithm. In this paper, we have obtained a reduced subset of features by applying the PART algorithm on dataset. First, a set of decision rules is built by applying the PART on the training set. Each rule contains a number of features. We then extract all features contained in the rule. Finally, the set of reduced features is derived. These features are shown in Table1.

3.4 Handling class imbalance

Customer churn is often a rare event in service industries, but of great significance and great value [25]. This means that real customer churn datasets have extremely skewed class distribution. For example, the dataset used in this study has extremely skewed class distribution; such that the Class distribution of churners versus non-churners is 14.3:85.7. This causes the classification modeling techniques experience difficulties in learning which customers are about to churn.

There are several data mining problems related to rarity along with some methods to address them [25]. The basic sampling methods include under-sampling and over-sampling. Under-sampling eliminates majority-class examples while over-sampling, in its simplest form, duplicates minority-class examples. Both of these sampling techniques decrease the overall level of class imbalance, thereby making the rare class less rare [26].

We applied the over-sampling techniques to make balance between churners and non-churners instances. By doing so the distribution of churners versus non-churners is the same.

3.5 Evaluation Criteria

If TP, FP, TN, and FN are the True Positives, False Positives, True Negatives and False Negatives in the confusion matrix, then accuracy is defined as (TP + TN)/(TP + FP + TN + FN).
The sensitivity is (TP/ (TP+FN)): the proportion of positive cases which are predicted to be positive.
The specificity is (TN/ ((TN+ FP)): the proportion of negative cases which are predicted to be negative [22].

In this study, we have used Accuracy, Sensitivity, and Specificity to quantify the accuracy of the predictive models. Furthermore, we have used the number of generated rules (#rules) to measure the comprehensibility of the constructed models.

Table 1: Top nine features selected by PART

Feature	**Type**	**What**
InterPlan	Dichotomous Categorical	International Plan Subscriber(0=no, 1=yes)
VmailP	Dichotomous Categorical	VoiceMail Plan Subscriber(0=no, 1=yes)
TotalDayMins	Continuous	Daytime usage
TotalEveMins	Continuous	Evening usage
TotalEveCharge	Continuous	Charge for evening usage
TotalNightCharge	Continuous	Charge for night time usage
TotalInterMins	Continuous	International usage
TotalInterCalls	Continuous	Number of international calls
NumberofCalltoCS	Continuous	Number of calls to customer service

3.6 Model building

In this paper, we used the subtractive clustering technique with (genfis2) function. Given separated sets of input and output data, the genfis2 uses a subtractive clustering method to generate a fuzzy inference system (FIS). When there is only one output, genfis2 may be used to generate an initial FIS for ANFIS training by first implementing subtractive clustering on the data. The genfis2 function uses the subclust function to estimate the antecedent membership functions and a set of rules. This function returns an FIS structure that contains a set of fuzzy rules to cover the feature space. The parameters of subtractive clustering were set as follows: the range of influence is 0.5, squash factor is 1.25, accept ratio is 0.5; rejection ration is 0.15.

The number of epoch is equal to 100. We name the FIS generated by subtractive clustering and trained by ANFIS as ANFIS-Subtractive model.

We also used the FCM clustering technique with (genfis3) function. genfis3 generates a FIS using fuzzy c-means (FCM) clustering by extracting a set of rules that models the data behavior. Similar to genfis2 this function requires separate sets of input and output data as input arguments. When there is only one output, you can use genfis3 to generate an initial FIS for ANFIS training. The rule extraction method first uses the fcm function to determine the number of rules and membership functions for the antecedents and consequents [27].

We set the number of cluster for FCM equal to 6. The number of epoch is equal to 100. We name the FIS generated by FCM clustering and trained by ANFIS as ANFIS-FCM model.

C4.5 decision tree, RIPPER, and logistic regression with default parameters were executed in WEKA (Waikato Environment for Knowledge Analysis) data mining software [23].

4. RESULTS AND ANALYSES

4.1 Predictive power

As can be seen from the table, the highest accuracy is obtained using RIPPER rule learner (Accuracy = 95%). However, C4.5 decision tree, ANFIS- Subtractive, and ANFIS-FCM follow closely, and except for logistic regression, all results lies in interval between 91% and 95%. Because accuracy implicitly assumes a relatively balanced class distribution among the observations and equal misclassification costs, it alone is not an adequate performance measure to evaluate the experimental results [28].

A Real churn dataset, has a skewed distribution, therefore the supposition of equal misclassification costs cannot be sustained. Typically, for a customer relationship manager, the most important issue is the correct detection of future churner. Since the costs related with the misclassification of churners are clearly higher than the costs related to misclassification of a non-churner, we should assume unequal misclassification costs. As a result, a high sensitivity is preferred to a high specificity from a company's view point. Of course this does not mean that specificity can be completely ignored. Indeed, a reasonable tradeoff has to be made between specificity and sensitivity. A churn prediction model that predicts all customers as churners might performs well in including all churning customers in retention campaign, but this lead to the an extremely high retention marketing costs.

The highest sensitivity in our experiments is obtained with C4.5 (Sensitivity=87%). RIPPER, ANFIS- subtractive, ANFIS- FCM and logistic regression don't perform significantly worse.

The highest specificity in our experiments is reached with RIPPER (Specificity= 97.5%). C4.5, ANFIS- subtractive, and ANFIS-FCM models don't differ significantly in terms of specificity, and except for logistic regression, all results lies in interval between 92 % and 95.6%.

In sum, ANFIS-Subtractive and ANFIS-FCM models have reasonable performance in terms of accuracy, specificity, and sensitivity.

4.2 Comprehensibility

Accuracy, sensitivity, specificity are not the only important aspect in evaluating a churn prediction models [28]. A churn prediction model should be both comprehensible and accurate. Comprehensibility of model causes it to reveal some knowledge about churn drivers of customers. Such knowledge can be extracted in the form of "if then" rules which allows developing a more effective retention strategy. Therefore, comprehensibility of the classification model is an important requirement in churn prediction modeling.

Among the five algorithm used in this paper, logistic regression doesn't support a rule based representation. On the other hand, RIPPER, C4.5, ANFIS-Subtractive, and ANFIS-FCM induce comprehensible rules from a dataset. As the results shows ANFIS-Subtractive and ANFIS-FCM clearly induce much less rules than C4.5 and RIPPER. Hence ANFIS-Subtractive and ANFIS- FCM which result in a comparable number of rules are the most comprehensible techniques tested in the experiments. The if-then rules generated from ANFIS-Subtractive clustering were shown in figure2. These results indicate that ANFIS has acceptable performance in terms of accuracy and comprehensibility, and it is an appropriate choice for churn prediction applications.

Table 2: Performance of algorithms

Technique	Accuracy	Specificity	Sensitivity	#rules
C4.5	**94%**	**95.6%**	**87%**	25
RIPPER	**95%**	**97.5%**	**85.7%**	18
Logistic regression	77.3%	76.6%	**82%**	----
ANFIS-Subtractive	**92%**	**93%**	**84%**	**6**
ANFIS-FCM	**91%**	**92%**	**84%**	**6**

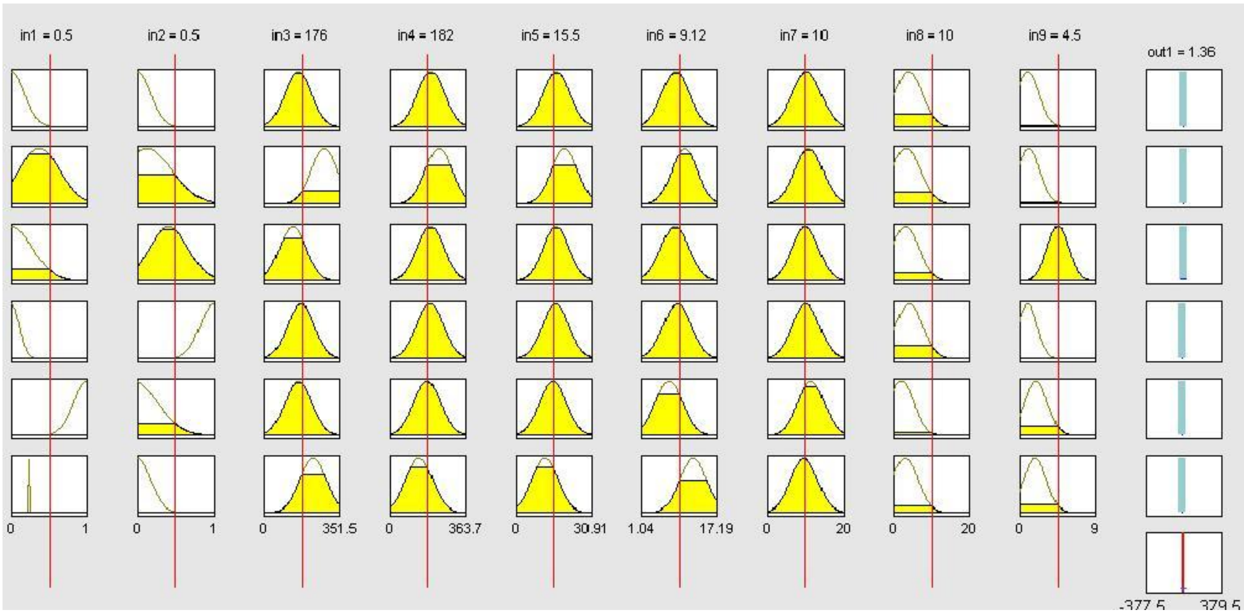

Figure 2 :The if-then rules generated from ANFIS-Subtractive results

5. CONCLUSIONS

Both accuracy and comprehensibility are two important requirements in churn prediction modeling. This paper presents application of ANFIS in churn prediction context. Particularly, we compared ANFIS as a neuro-fuzzy classifier with two state-of-the-arts crisp classifiers including C4.5 and RIPPER rule learner. The results showed that both ANFIS-Subtractive and ANFIS-FCM models have acceptable performance in terms of accuracy, specificity, and sensitivity. In addition, ANFIS-Subtractive and ANFIS-FCM clearly induce much less rules than C4.5 and RIPPER. Hence ANFIS-Subtractive and ANFIS-FCM which result in a comparable number of rules are the most comprehensible techniques tested in the experiments. These results indicate that ANFIS showed acceptable performance in terms of accuracy and comprehensibility, and it is an appropriate choice for churn prediction applications.

6. ACKNOWLEDGMENTS

We thank the Iran Telecommunication Research Center for financial support.

7. REFERENCES

[1] Neslin, S.A. Gupta, S. Kamakura, W. Lu, J. Mason, C., 2006. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 43(2), 204–211.

[2] Coussement, K. F. Benoit, D. Van den Poel, D., 2010. Improved marketing decision making in a customer churn prediction context using generalized additive models, Expert Systems with Applications 37, 2132–2143.

[3] Torkzadeh, G., Chang, J. C.-J., & Hansen, G. W., 2006. Identifying issues in customer relationship management at Merck-Medco. Decision Support Systems, 42(2).

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

[5] Coussement, K. Van den Poel, D., 2008a. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications 34, 313–327.

[6] Xie, Y. Li, X. Ngai, E. Ying, W., 2009. Customer churn prediction using improved balanced random forests, Expert Systems with Applications 36, 5445–5449.

[7] Yu, X., et al. 2010, An extended support vector machine forecasting framework for customer churn in e-commerce. Expert Systems with Application , doi:10.1016/j.eswa.2010.07.049.

[8] [8]Huang, B. Buckley, B. Kechadi, T., 2010.Multi-objective feature selection by using NSGA-II for custom-er churnprediction in telecommunications, Expert Systems with Applications 37, 3638–3646.

[9] Tsai, C. Lu, Y., 2009. Customer churn prediction by hybrid neural networks, Expert Sys-tems with Applications, 36, 12547–12553.

[10] Pendharkar, P., 2009,Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services, Expert Systems with Applications 36, 6714–6720.

[11] Lemmens,A., and Croux,C., 2006. "Bagging and boosting classification trees to predict churn," Journal of Marketing Research, vol. 43, no. 2, pp. 276-286, 2006.

[12] Coussement, K. Van den Poel, D., 2008b. Integrating the voice of customers through call center emails into a decision support system for churn prediction- Information & Management, 45 , 164–174.

[13] Bezdec, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York

[14] Chiu, S., 1994. "Fuzzy Model Identification Based on Cluster Estimation," Journal of Intelligent & Fuzzy Systems, Vol. 2, No. 3, Spet.

[15] Yager, R. Filev, D., 1994. Generation of fuzzy rules by mountain clustering, J. Intell. Fuzzy Syst. 2 (3) , 209–219.

[16] Zadeh, L.A., 1965. Fuzzy sets, Information and Control 8, 338–353.

[17] Zadeh, L.A., 2005. Toward a generalized theory of uncertainty (GTU) – an outline, Information Sciences 172, 1–40.

[18] Mamdani, EH., Assilian. S., 1975. An experiment in linguistic synthesis with a fuzzy logic controller. Int J Man_Machine Studies, 7(1), 1-13.

[19] Sugeno, M., 1985. Industrial applications of fuzzy control. Elsevier Science Pub., Co.

[20] Jang, J.-S. R., 1993. "ANFIS: Adaptive-Network-based Fuzzy Inference Systems," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 23, No. 3, pp. 665-685.

[21] Larose, D., 2005. Discovering knowledge in data: An introduction to data mining. NewJersey, USA: Wiley.

[22] Han, J., & Kamber, M., 2006. Data Mining Concepts and Techniques. Morgan Kaufmann.

[23] Witten, I. H. & Frank, E., 2005. Data mining: Practical machine learning tools and techniques. San Francisco: Morgan Kaufmann. 0-12-088407-0.

[24] Berger, H., Merkl, D., Dittenbach, M. 2006. Exploiting Partial Decision Trees For Feature Subset Selection in e-Mail Categorization, In Proceedings of the ACM Symposium on Applied Computing (SAC).

[25] Burez, J. Van den Poel , D., 2009.Handling class imbalance in customer churn prediction, Expert Systems with Applications 36, 4626–4636

[26] Weiss, G. M., 2004. Mining with rarity: A unifying framework. SIGKDD Explorations, 6(1), 7–19.

[27] Fuzzy logic toolbox user's guide for use with MATLAB 2010.

[28] Verbeke, W., et al., 2011. Building comprehensible customer churn prediction models with ad-vanced rule induction techniques. Expert Systems with Applications, 38, 2354–2364.