Hindawi Publishing Corporation
Journal of Engineering
Volume 2013, Article ID 543940, 9 pages
http://dx.doi.org/10.1155/2013/543940

Research Article

Hierarchical Neural Regression Models for Customer Churn Prediction

Golshan Mohammadi,[1] Reza Tavakkoli-Moghaddam,[2] and Mehrdad Mohammadi[2]

[1] *Department of Finance Management, Faculty of Humanities and social Sciences, Islamic Azad University, Sanandaj Branch, Sanandaj, Iran*
[2] *Department of Industrial Engineering, College of Engineering, University of Tehran, P.O. Box 11155/4563, Tehran, Iran*

Correspondence should be addressed to Mehrdad Mohammadi; mehrdadmohamadi@ut.ac.ir

Received 25 November 2012; Revised 27 January 2013; Accepted 1 February 2013

Academic Editor: Jie Zhou

As customers are the main assets of each industry, customer churn prediction is becoming a major task for companies to remain in competition with competitors. In the literature, the better applicability and efficiency of hierarchical data mining techniques has been reported. This paper considers three hierarchical models by combining four different data mining techniques for churn prediction, which are backpropagation artificial neural networks (ANN), self-organizing maps (SOM), alpha-cut fuzzy *c*-means (α-FCM), and Cox proportional hazards regression model. The hierarchical models are ANN + ANN + Cox, SOM + ANN + Cox, and α-FCM + ANN + Cox. In particular, the first component of the models aims to cluster data in two churner and nonchurner groups and also filter out unrepresentative data or outliers. Then, the clustered data as the outputs are used to assign customers to churner and nonchurner groups by the second technique. Finally, the correctly classified data are used to create Cox proportional hazards model. To evaluate the performance of the hierarchical models, an Iranian mobile dataset is considered. The experimental results show that the hierarchical models outperform the single Cox regression baseline model in terms of prediction accuracy, Types I and II errors, RMSE, and MAD metrics. In addition, the α-FCM + ANN + Cox model significantly performs better than the two other hierarchical models.

1. Introduction

In today's competitive world, customer churn management (CCM) is an important task for each service provider to build long-term and profitable relationships with specific customers [1, 2]. The service providers in telecommunication industry suffer from attracting valuable customers with competitors; this is known as customer churn. Recently, there have been many changes in the telecommunications industry, such as, loyalty program for more profitable customers [3]. Loyal customers are the most fertile source of data for decision making. This data reflects the customers' actual behavior and those factors affect their loyalty. The potential value of customers can be evaluated by these data [3], also assessing the risk that they will stop paying their bills, and predicting their future needs [4].

Besides, because customer attrition will absolutely result in loss of incomes, customer churn management has received increasing attention in the whole marketing and management literature. Moreover, it has been proven that considerable impact on incomes is occurred by small change in retention rate [5].

The effective customer churn management for companies needs building more comprehensive and accurate churn prediction model. Recently, several customer churn prediction models have been presented in a number of domains such as telecommunications [6–8], retail markets [9, 10], subscription management [11, 12], banking service providers [13], and wireless commerce [14]. Among previous studies in the literature, statistical and data mining techniques have been applied to build the prediction models.

These techniques include artificial neural networks (ANNs) [7], Bayesian networks [6, 9], decision trees [15, 16], AdaBoosting [13], logistic regression [10, 11, 16, 17], random forest [10, 11], the proportional hazard model [5], and SVMs. Lessmann and Voß [18] gave a detailed review on this topic.

Two main tasks of data mining techniques are describing remarkable pattern or relationship in the data and also predicting a conceptual model which data followed up [2].

In the literature, it has been proven that hybrid data mining approaches by combining clustering and classification techniques have better performance in comparison with single clustering and classification data mining techniques. Hybrid approaches are particularly combined of two learning stages, in which the first one is preprocessing the data and the second one is the final prediction output [7]. Other hybrid data mining techniques for predicting customer churn model include using well-known metaheuristic algorithms (e.g., genetic algorithm) based on neural network which outperform traditional local search gradient descent/gradient ascent neural networks that use Rumelhart et al. [19] procedure for updating connection weights [20–22].

In addition to predicting the customer churn model and determining that which customer belongs to which class (i.e., churned and nonchurned classes), companies are eager to know when, why, and with what probability their customers try to switch their subscription. Having knowledge about those factors which significantly affect customers churn behavior is more important than just knowing classes of customers. These effective factors are needed for companies to plan their long-term strategies for decreasing customer churn rate and above all, scheduling and adopting best marketing strategies based on when and why their customers like to break up their relationship because some companies suffer from marketing expenses in some especial times while they are not aware of what their customers want. On the other hand, having knowledge about effective factors and probability of attrition enables companies to focus on those customers who are more likely to churn. This useful information can be extracted using survival analysis of customers. In order to determine the hazard probability function of the customers and the above-mentioned information, the Cox proportional hazard method is applied as a last part of hierarchical methods because the ANNs are not able to calculate the churn probability of the customers. Another reason for using the Cox regression model is our used data. The customer churn data consists of censored data. Censored data occurs when you know that a measurement exceeds some threshold, but you do not know by how much. So in this study, each customer who has not churned till the end of the experiment is considered as a right censored data. Therefore, the Cox regression model is conducted on the customer data to cope with censored data.

However, few papers studied hierarchical data mining techniques for customer churn prediction. Therefore, in this paper, some data mining techniques are presented to create the hierarchical model of customer churn prediction. The hierarchical methods are based on combining clustering, that is, alpha-cut fuzzy c-means (α-FCM), self-organizing maps (SOM), and artificial neural network (ANN), classification techniques, that is, ANN, and survival analysis, that is, Cox proportional hazard regression model, which their combinations are α-FCM + ANN + Cox, SOM + ANN + Cox, and ANN + ANN + Cox. To evaluate the performance of the hierarchical models, an Iranian mobile dataset is considered for comparison between the hierarchical models and the single Cox regression baseline model in terms of prediction accuracy, Types I and II errors, RMSE, and MAD metrics. It also should be mentioned that some other well-known techniques, such as Fuzzy ARTMAP [23] and LLMF [24], were used in designing some other hierarchical methods (e.g., SOM + Fuzzy ARTMAP + Cox, ANN + Fuzzy ARTMAP + Cox, ANN + LLMF + Cox, SOM + LLMF + Cox, and α-FCM + LLMF + Cox), but just the above-mentioned hierarchical techniques are proposed and reported based on their better performance. Finally, some of contributions of this paper are as follows.

(i) Considering nonchurned customer as censored data and using Cox regression model as a first time in the literature in order to determine customers churn prediction.

(ii) Determining important factors affecting the customer churn in the Iranian telecommunication industry.

(iii) Determining the hazard and survival functions of each customer based on effective factors.

(iv) Proposing some new combination of data mining techniques containing ANN, SOM, α-FCM, and Cox regression as hierarchical methods.

(v) Conducting the proposed hierarchical methods on a dataset of Iranian telephony market.

(vi) Comparing different proposed hierarchical methods.

The rest of our paper is organized as follows. In Section 2, we describe the proposed data mining techniques in this paper. Section 3 describes the research methodology, and Section 4 presents the experimental results. Finally, the conclusion is provided in Section 5.

2. Proposed Data Mining Techniques

In order to create effective and accurate customer churn prediction models, many data mining techniques have been considered over the past time in the marketing and management literature (e.g., [12, 25]). The proposed data mining techniques are as follows.

2.1. Alpha-Cut Fuzzy C-Means Clustering. Clustering is an unsupervised learning technique that breaks down a set of patterns into groups (or clusters). Clustering technique refers to the partitioning of a set of data object into clusters. In particular, no predefined classes are assumed [26].

Classical clustering partitions each observation is assigned to a single group (cluster), without considering the degree of distinction or similarity of the observation

from all the other possible clusters. This type of clustering is often called hard or crisp clustering [5]. Nevertheless, fuzzy clustering methods based on the fuzzy set theory and on the concept of membership functions have been developed. In the fuzzy clustering, observations are allowed to belong to more than one cluster with different degrees of membership.

Fuzzy clustering of an observation X into c clusters is characterized by c membership functions μ_j as follows:

$$\begin{aligned} &\mu_j : X \longrightarrow [0,1], \quad j = 1, \ldots, c, \\ &\sum_{j=1}^{c} \mu_j (x_i) = 1, \qquad i = 1, 2, \ldots, n, \\ &0 < \sum_{i=1}^{n} \mu_j (x_i) < n, \qquad j = 1, 2, \ldots, c. \end{aligned} \tag{1}$$

Membership function is calculated based on the distance of observations from clusters' center. The well-known method of fuzzy clustering is the fuzzy c-means technique (FCM), initially proposed by Dunn [27]. FCM applies two consecutive steps including (a) calculation of the clusters' center and (b) assigning the observations to these clusters' center using specific form of distance, in order to minimize a standard loss function (SLF) as follows:

$$\text{SLF} = \sum_{k=1}^{c} \sum_{i=1}^{n} [\mu_k (x_i)]^m \left\| x_i - c_k \right\|^2, \tag{2}$$

where cluster center c_k and membership function of observation i in cluster k are calculated by (3) and (4), respectively

$$c_k = \frac{\sum_i [\mu_k (x_i)]^m x_i}{\sum_i [\mu_k (x_i)]^m}, \tag{3}$$

$$\mu_k (x_i) = \frac{(1/d_{ki})^{1/(m-1)}}{\sum_{k=1}^{c} (1/d_{ki})^{1/(m-1)}}, \tag{4}$$

where d_{ki} is the distance metric for observation i in cluster k.

2.2. Self-Organizing Maps. A new form of a neural network architecture called self-organizing map (SOM) was proposed by Kohonen [28], which has proved extremely efficient when the high degree of dimensionality and complexity accurses in input data. SOM is used to find out relationships in a dataset and cluster data according to the similarity of data (i.e., similar expression patterns) where the nature of the classification cannot be predicted by the model creators, or there may be more than one method to cluster the characteristics of a dataset [29]. Figure 1 shows an example of a 4×4 SOM.

2.3. Artificial Neural Network. Classification is one of the commonly used data mining techniques categorizing as supervised learning techniques. It determines the value of some variables and classifies according to results. The common algorithms of classification include decision trees, artificial neural networks (ANNs), and so on [30] in which artificial neural networks are the most recently applied methods in literature.

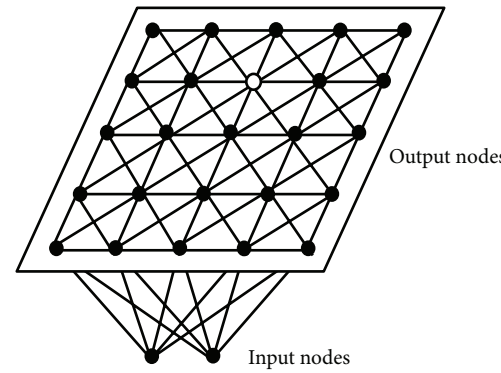

Figure 1: A 4×4 Kohonen's self-organizing map.

An ANN consists of some nodes and links between them. The ANN takes a number of input data and produces a single output data through an internal weighting system. ANNs can be categorized into single-layer perception or multilayer perception (MLP). The multilayer perception consists of multiple layers of simple, two taste, sigmoid processing nodes, or neurons that act together using internal weighted system. In addition, the neural network consists in one or more several intermediary layers between the input and output layers. Such intermediary layers are called hidden layers and nodes embedded in these layers are called hidden nodes. Figure 2 illustrates a multilayer neural network.

2.4. Cox Proportional Hazards Model. According to the Cox and Oakes [31], the Cox model is based on a modelling approach in order to analysing survival data. The purpose of the model is to simultaneously explore the effects of several variables on survival. The Cox model is a well-recognised statistical technique for analysing survival data. Survival analysis typically examines the relationship of the survival distribution to covariates. Most commonly, this examination entails the specification of a linear-like model for the log hazard. For example, a parametric model based on the exponential distribution may be written as follows:

$$\log h_i (t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{5}$$

or, equivalently,

$$h_i (t) = \exp (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}). \tag{6}$$

Equation (5) is a linear model for the log-hazard or a multiplicative model for the hazard. In (5), i is a subscript for observation, and the x's are the covariates. The constant α in this model represents a kind of log-baseline hazard, since $\log h_i(t) = \alpha$ [or $h_i(t) = e^{\alpha}$] when all of the x's are zero. Equation (6) is similar to parametric regression models based on the other survival distributions.

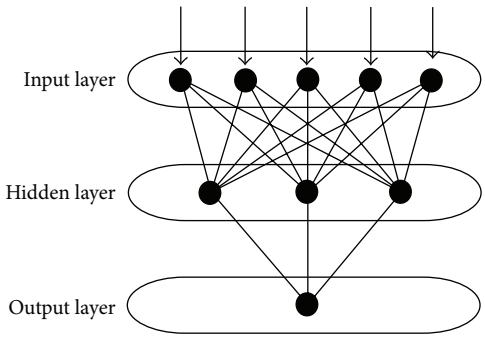

Figure 2: Multilayer neural network.

The Cox model, in contrast, leaves the baseline hazard function $\alpha(t) = \log h_0(t)$ unspecified

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \tag{7}$$

$$h_i(t) = h_0(t) \exp\left(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}\right), \tag{8}$$

where (8) is a semiparametric because while the baseline hazard can take any form, the covariates enter the model linearly.

3. Research Methodology

3.1. Data Set. For the purpose of this paper, we consider a CRM data set provided by an Iranian mobile operator. Specifically, the dataset contains 3,150 subscribers, including 495 churners and 2,655 nonchurners, from September 2008 to August 2009. In addition, the subscribers have to be mature customers who were with the mobile operator for at least 2 months. Churn was then calculated based on whether the subscriber left the company during the 10 remained months. Churned customer is defined as a customer who has not made any contact with the operator (e.g., making a call, charging a credit, changing subscription, etc.).

3.2. Model Development

3.2.1. The Baseline. As the last part of all proposed hierarchical methods is Cox regression method and also the final aim of these hierarchical methods is determining better hazard and survival functions for customer churn prediction, therefore, we use the original dataset to create a Cox proportional hazards regression model as the baseline Cox model for comparison.

3.2.2. ANN + ANN + Cox. The first hierarchical model is based on combining two ANN models and Cox regression, in which the first ANN performs the data reduction task and the second ANN for churn classification and the last Cox regression for hazard function prediction. As there is no 100% accuracy, there are a number of correctly and incorrectly predicted data from the training set by the first ANN model. Consequently, the incorrectly predicted data can be regarded as outliers since the ANN model cannot predict them accurately. Then, the correctly predicted data by the first ANN model are used to train the second ANN model as the classification model. Finally, the corrected classified data from second ANN are used by Cox regression to predict hazard function.

3.2.3. SOM + ANN + Cox. For the second hierarchical model, a self-organizing map (SOM), which is a clustering technique, is used for the data reduction task. Then, the corrected clustered data are used to train the second model based on ANN. Finally, the classification result is used to hazard function prediction. To develop the SOM, the map size is set by $2 * 2$, $3 * 3$, $4 * 4$, $5 * 5$, and $6 * 6$, respectively, in order to obtain the highest rate of prediction accuracy. Then, two clusters of SOM which contain the highest proportion of the churner and nonchurner groups, respectively, are selected as the clustering result.

3.2.4. α-FCM + ANN + Cox. In the third hierarchical model, α-FCM, which is a clustering approach, is used for data reduction task. In the fuzzy c-means (FCM) clustering algorithm, almost none of the data points have a membership value of 1. Besides, noise and outliers may cause difficulties in obtaining appropriate clustering results from the FCM algorithm. Therefore, many studies have been done about the FCM algorithm in the literature [32]. Furthermore, studies about FCM can be divided into two categories. One is to extend the dissimilarity (or distance) measure $d(x_j, a_i)$ between the data point x_j and the cluster center a_i in the FCM objective function by replacing the Euclidean distance with other types of metric measures [33]. The other category is to extend the FCM objective function by adding a penalty term [34].

One of the best methods for assigning a data point to exactly one cluster is that if the membership value μ_{ij} of the data point x_j in the ith cluster is larger than a given value α, then the point x_j will exactly belong to the ith cluster with membership value of 1 and $\mu_{i'j} = 0$ for all $i \neq i'$. In order to guarantee that no two of these c cluster cores will overlap, the value of α is set to interval [0.5, 1] [35]. The cluster cores generated by FCMα can be calculated by (9)

$$\mu_{ij} = \frac{d\left(x_j, a_i\right)^{-1/(m-1)}}{\sum_{k=1}^{c} d\left(x_j, a_k\right)^{-1/(m-1)}} > \alpha \tag{9}$$

which is equivalent to

$$d\left(x_j, a_i\right)^{-1/(m-1)} < \frac{1}{\alpha \sum_{k=1}^{c} d\left(x_j, a_k\right)^{-1/(m-1)}}, \tag{10}$$

where m is the fuzziness index so its value is considered as 2. Interesting readers are referred to [35] for more detail. Then, the corrected clustered data are used to train second ANN model in order to customer classification. Finally, the hazard function using Cox regression is predicted based on the corrected classified result from ANN model.

3.2.5. Evaluation Method. To evaluate the proposed churn prediction models, prediction accuracy, and the Type I and II errors are considered. They can be measured by a confusion matrix shown in Table 1. The rate of prediction accuracy is defined as $(a + d)/(a + b + c + d)$.

The Type I error is the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In this paper, it means that the customer is not churned when the model has predicted that the hazard function of that customer is more than α (i.e., α is the alpha cut in fuzzy c-means clustering method). On the other hand, the Type II error is defined as the error of rejecting a null hypothesis when it is the true state of nature. It means that the customer is churned when the model has predicted that the survival function of that customer is more than α.

We also compare the performance of the proposed model with pure Cox proportional hazards model in predicting the churn or survival probability of the customers. The observed outcome for each customer in the sample is either churn or survival (i.e., still active) by the end of the study period. We compute the deviation between observed and predicted outcomes (i.e., the probability of churn or survival as predicted by the model) for both proposed and pure Cox model. The Root Mean Squared Error (RMSE) and Mean Absolute Deviation (MAD) are calculated for comparing both models as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N_{\text{ch}}} \left(S_{\text{ch}}^{i} - 0\right)^2}{N_{\text{ch}}} + \frac{\sum_{j=1}^{N_{\text{Nch}}} \left(1 - S_{\text{Nch}}^{j}\right)^2}{N_{\text{Nch}}}},$$
$$\text{MAD} = \frac{1}{N_{\text{ch}}} \sum_{i=1}^{N_{\text{ch}}} \left|e_{\text{ch}}^{i} - \bar{e}_{\text{ch}}\right| + \frac{1}{N_{\text{Nch}}} \sum_{j=1}^{N_{\text{Nch}}} \left|e_{\text{Nch}}^{j} - \bar{e}_{\text{Nch}}\right|, \tag{11}$$

where S_{ch}^{i} and S_{Nch}^{j} are the survival probability of churned customer i and non-churned customer j, respectively; N_{ch} and N_{Nch} are the number of churned and nonchurned customer, respectively; e_{ch}^{i} is the deviation of churned customer i from zero (i.e., $e_{\text{ch}}^{i} = (S_{\text{ch}}^{i} - 0)$) and e_{Nch}^{j} is the deviation of non-churned customer j from one (i.e., $e_{\text{Nch}}^{j} = (1 - S_{\text{Nch}}^{j})$), and $\bar{e}_{\text{ch}}$ and $\bar{e}_{\text{Nch}}$ are the mean of the deviation of churned and non-churned customers, respectively.

4. Experimental Results

4.1. The Baseline. In order to create the Cox model, 2350 and remained 800 numbers of data are used for training and testing the Cox model, respectively. Table 2 shows the prediction performance of the baseline Cox proportional hazards model based on type I and II errors, accuracy, RMSE, and MAD metrics. On average, the baseline Cox proportional hazards model provides about 84% accuracy meaning that in 128 cases of data, the Cox model was unable to correctly predict the survival and hazard probability based on value of alpha-cut 0.7. The type I and II errors were equal to 87 and 41 cases of incorrectly predicted data. The baseline Cox model also provides 0.083 and 0.098 as the RMSE and MAD error metrics, respectively.

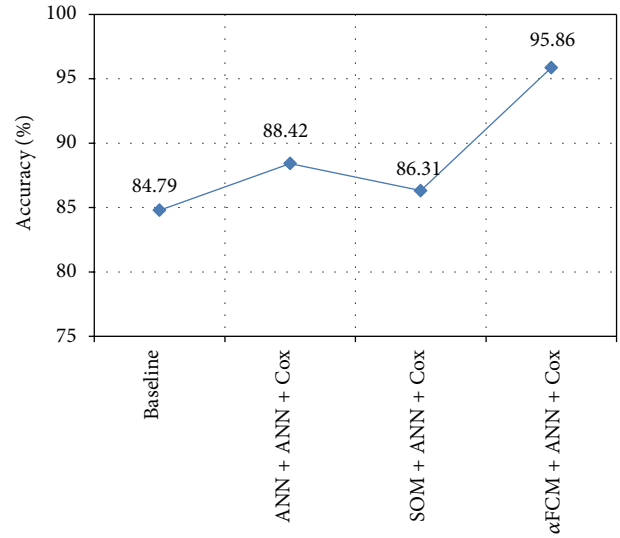

Figure 3: The accuracy of hierarchical models.

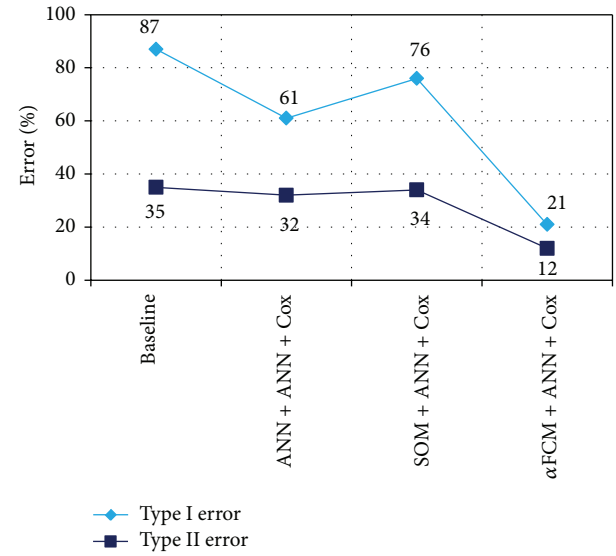

Figure 4: The Type I and II errors of hierarchical models.

4.2. ANN + ANN + Cox. For the first hierarchical model based on combining two ANN models and Cox regression, the first ANN model performs the data reduction task. Therefore, we run the ANN model by a set of different hidden layer and learning epochs. The result of different combination of hidden layer and learning epochs is as Table 3 in which an ANN models with 16 and 12 hidden layer and 100 and 300 learning epochs are considered for two ANN model, respectively. Finally, the accuracy and other performance metrics for hierarchical ANN + ANN + Cox model are shown in Table 4.

4.3. SOM + ANN + Cox. To construct the second hierarchical model by combination of SOM, ANN, and Cox regression, $2 * 2$, $3 * 3$, $4 * 4$, $5 * 5$, and $6 * 6$ SOMs are used to

Table 1: Confusion matrix.

		Actual	
		Nonchurners	Churners
Predicted	Nonchurners	a	b (II: $S^* > \alpha$)
	Churners	c (I: $H^{**} > \alpha$)	d

*Probability of survival.
**Probability of hazard.

Table 2: The prediction performance of the baseline Cox proportional hazards model.

	Performance metrics				
	Accuracy	Error type I	Error type II	RMSE	MAD
Value	84.79%	87	35	0.091	0.098

cluster the data at first. We found that $4 * 4$ SOM performs the best which can provide the highest rate of accuracy for two clusters, that is, churner and nonchurner clusters. Then, the accurate clustered data are used for training classifier ANN and the result of different hidden layer and learning epochs is as Table 5 in which an ANN model with 16 hidden layer and 200 epochs is considered as classifier model. Finally, the accuracy and other performance metrics for hierarchical SOM + ANN + Cox model are shown in Table 6.

4.4. α-FCM + ANN + Cox. To construct the third hierarchical model based on alpha-cut fuzzy c-means, classifier ANN model and Cox proportional hazards model, the alpha-cut fuzzy c-means with alpha-cut equal to 0.7, and ANN with 16 hidden layer and 100 learning epochs were found regarding to the best reported accuracy. The *α-FCM*, used for data reduction task, results in two clusters including 2210 churners and 940 nonchurners. Then, 2350 and remained 800 numbers of data are, respectively, used for training and testing the classification ANN model where the prediction performance of ANN model is shown in Table 7. Finally, Table 8 shows the performance metrics of α-FCM + ANN + Cox hierarchical models.

On average, the α-FCM + ANN + Cox hierarchical model provides about 95.49% accuracy based on alpha-cut equal to 0.7 and ANN with 16 hidden layer and 100 learning epochs. The type I and II errors are equal to 21 and 12 cases of incorrectly predicted data. The α-FCM + ANN + Cox hierarchical model also provides 0.031 and 0.042 as the RMSE and MAD error metrics, respectively.

In order to show the high performance of α-FCM + ANN + Cox hierarchical model, the accuracy, errors type I and II, RMSE, and MAD metrics are illustrated in Figures 3, 4, and 5, respectively.

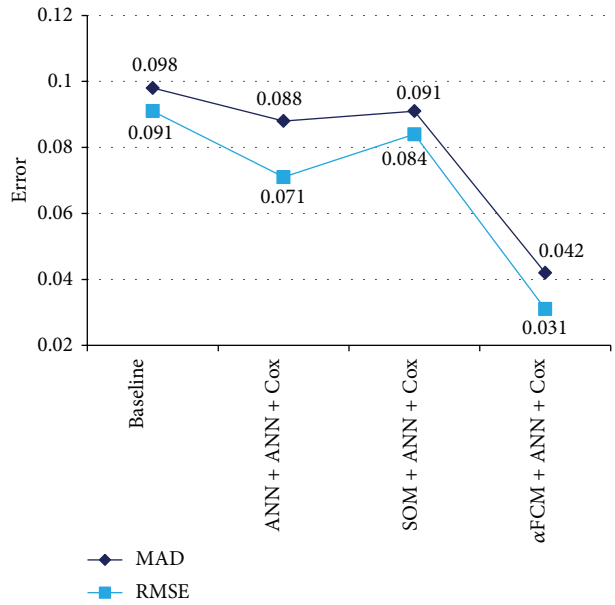

Figure 5: The RMSE and MAD errors of hierarchical models.

5. Conclusion

As customers are the main competitive advantage of each industry, customer churn prediction is becoming a major task for companies to remain in competition with other industries. Therefore, building an effective customer churn prediction model, which provides an acceptable level of accuracy, has become a research problem for companies in recent years. In the literature, the better applicability and efficiency of hierarchical data mining techniques in order to predict customer attrition by combining two or more techniques has been reported over a number of different domain problems. In this paper, we consider three different hierarchical data mining techniques based on combination of some neural networks and regression model to examine their performances for telecommunication industry. In particular, backpropagation artificial neural networks (ANN), self-organizing maps (SOM), alpha-cut fuzzy c-means, and Cox proportional hazard model are considered. Consequently, ANN + ANN + Cox, SOM + ANN + Cox, and α-FCM + ANN + Cox hierarchical models are developed, in which the first component of the hierarchical models filter out unrepresentative data or outliers. Then, the corrected output clustered data are used to classify customer into churner and nonchurner groups.

To evaluate the performance of the hierarchical models, an Iranian mobile dataset is considered. The experimental

Table 3: Prediction performance of ANN + ANN hierarchical models.

Hidden layer	Learning epochs							
	Clustering ANN				Classification ANN			
	50	100	200	300	50	100	200	300
8	87.90	89.34	90.66	91.17	85.65	87.50	89.59	91.84
12	83.50	87.42	90.11	90.87	85.23	88.25	91.90	**92.49**
16	90.25	**92.69**	88.45	90.77	91.18	88.43	88.57	84.76
20	88.54	90.21	85.80	91.06	90.54	87.92	84.28	86.08
24	91.55	91.85	89.29	86.46	86.85	88.41	86.80	83.46
28	91.61	90.21	86.02	91.29	91.69	90.55	89.37	89.95
32	90.78	89.35	88.13	86.11	91.28	86.55	84.80	90.39

Table 4: Performance metrics of ANN + ANN + Cox hierarchical models.

	Performance metrics				
	Accuracy	Error type I	Error type II	RMSE	MAD
Value	88.42%	61	32	0.071	0.088

Table 5: Prediction performance of ANN in SOM + ANN + Cox hierarchical models.

Hidden layer	Learning epochs			
	50	100	200	300
8	83.40	85.30	89.30	90.67
12	84.31	86.46	89.31	89.77
16	87.45	90.28	**93.88**	91.18
20	90.30	90.29	91.87	84.06
24	85.76	89.44	83.37	93.87
28	90.99	91.57	87.59	90.20
32	91.46	83.28	85.71	89.60

Table 6: Performance metrics of SOM + ANN + Cox hierarchical models.

	Performance metrics				
	Accuracy	Error type I	Error type II	RMSE	MAD
Value	86.31%	76	34	0.084	0.091

Table 7: Prediction performance of ANN in α-FCM + ANN + Cox hierarchical models.

Hidden layer	Learning epochs			
	50	100	200	300
8	84.32	85.31	87.96	89.21
12	84.71	87.76	86.63	86.34
16	87.35	**93.20**	92.67	90.78
20	90.35	83.81	91.01	89.59
24	88.11	86.21	88.78	88.33
28	83.72	86.54	89.98	87.53
32	85.69	87.91	85.41	87.15

Table 8: Performance metrics of α-FCM + ANN + Cox hierarchical models.

	Performance metrics				
	Accuracy	Error type I	Error type II	RMSE	MAD
Value	95.86%	21	12	0.031	0.042

results show that the hierarchical models outperform the single Cox regression baseline model in terms of prediction accuracy, types I and II errors, RMSE, and MAD metrics. In addition, the α-FCM + ANN + Cox model significantly performs better than the SOM + ANN + Cox and ANN + ANN + Cox models.

For future work, other prediction techniques can be applied, such as support vector machines, genetic algorithms, logistic regression, and so forth. Finally, other domain datasets about churn prediction can be used for further comparison.

References

[1] K. Coussement and D. Van den Poel, "Integrating the voice of customers through call center emails into a decision support system for churn prediction," *Information and Management*, vol. 45, no. 3, pp. 164–174, 2008.

[2] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, "Application of data mining techniques in customer relationship management: a literature review and classification," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2592–2602, 2009.

[3] C. Hung and C. F. Tsai, "Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand," *Expert Systems with Applications*, vol. 34, no. 1, pp. 780–787, 2008.

[4] M. J. A. Berry and G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, 2003.

[5] D. Van den Poel and B. Larivière, "Customer attrition analysis for financial services using proportional hazard models," *European Journal of Operational Research*, vol. 157, no. 1, pp. 196–217, 2004.

[6] P. Kisioglu and Y. I. Topcu, "Applying Bayesian Belief Network approach to customer churn analysis: a case study on the telecom industry of Turkey," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7151–7157, 2011.

[7] C. F. Tsai and Y. H. Lu, "Customer churn prediction by hybrid neural networks," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12547–12553, 2009.

[8] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: a profit driven data mining approach," *European Journal of Operational Research*, vol. 218, pp. 211–229, 2011.

[9] B. Baesens, G. Verstraeten, D. Van den Poel, M. Egmont-Petersen, P. Van Kenhove, and J. Vanthienen, "Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers," *European Journal of Operational Research*, vol. 156, no. 2, pp. 508–523, 2004.

[10] W. Buckinx and D. Van den Poel, "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," *European Journal of Operational Research*, vol. 164, no. 1, pp. 252–268, 2005.

[11] J. Burez and D. Van den Poel, "CRM at a pay-TV company: using analytical models to reduce customer attrition by targeted marketing for subscription services," *Expert Systems with Applications*, vol. 32, no. 2, pp. 277–288, 2007.

[12] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques," *Expert Systems with Applications*, vol. 34, no. 1, pp. 313–327, 2008.

[13] N. Glady, B. Baesens, and C. Croux, "Modeling churn using customer lifetime value," *European Journal of Operational Research*, vol. 197, no. 1, pp. 402–411, 2009.

[14] X. Yu, S. Guo, J. Guo, and X. Huang, "An extended support vector machine forecasting framework for customer churn in e-commerce," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1425–1430, 2011.

[15] J. Qi, L. Zhang, Y. Liu et al., "ADTreesLogit model for customer churn prediction," *Annals of Operations Research*, vol. 168, no. 1, pp. 247–265, 2009.

[16] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, 2012.

[17] A. Keramati and S. M. S. Ardabili, "Churn analysis for an Iranian mobile operator," *Telecommunications Policy*, vol. 35, no. 4, pp. 344–356, 2011.

[18] S. Lessmann and S. Voß, "A reference model for customer-centric data mining with support vector machines," *European Journal of Operational Research*, vol. 199, no. 2, pp. 520–530, 2009.

[19] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1, pp. 318–362, MIT Press, Cambrige, Mass, USA, 1986.

[20] P. C. Pendharkar and J. A. Rodger, "An empirical study of impact of crossover operators on the performance of non-binary genetic algorithm based neural approaches for classification," *Computers and Operations Research*, vol. 31, no. 4, pp. 481–498, 2004.

[21] P. Pendharkar and S. Nanda, "A misclassification cost-minimizing evolutionary-neural classification approach," *Naval Research Logistics*, vol. 53, no. 5, pp. 432–447, 2006.

[22] P. C. Pendharkar, "A comparison of gradient ascent, gradient descent and genetic-algorithm- based artificial neural networks for the binary classification problem," *Expert Systems*, vol. 24, no. 2, pp. 65–86, 2007.

[23] E. Granger, M. A. Rubin, S. Grossberg, and P. Lavoie, "A What-and-Where fusion neural network for recognition and tracking of multiple radar emitters," *Neural Networks*, vol. 14, no. 3, pp. 325–344, 2001.

[24] J. Sharifie, C. Lucas, and B. N. Araabi, "Locally linear neurofuzzy modeling and prediction of geomagnetic disturbances based on solar wind conditions," *Space Weather*, vol. 4, no. 6, pp. 1–12, 2006.

[25] S. Y. Hung, D. C. Yen, and H. Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications*, vol. 31, no. 3, pp. 515–524, 2006.

[26] A. Jain, M. Murty, and P. Flyn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[27] J. Dunn, "A fuzzy relative of the ISO-data process and its use in detecting compact, well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[28] T. Kohonen, "Adaptive, associative, and self-organizing functions in neural computing," *Applied Optics*, vol. 26, no. 23, pp. 4910–4918, 1987.

[29] X. Zhang, J. Edwards, and J. Harding, "Personalised online sales using web usage data mining," *Computers in Industry*, vol. 58, no. 8-9, pp. 772–782, 2007.

[30] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2001.

[31] D. R. Cox and D. Oakes, *Analysis of Survival Data*, Chapman and Hall, London, UK, 1984.

[32] J. Yu and M. S. Yang, "Optimality test for generalized FCM and its application to parameter selection," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 1, pp. 164–176, 2005.

[33] K. L. Wu and M. S. Yang, "Alternative c-means clustering algorithms," *Pattern Recognition*, vol. 35, no. 10, pp. 2267–2278, 2002.

[34] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517–530, 2005.

[35] M. S. Yang, K. L. Wu, J. N. Hsieh, and J. Yu, "Alpha-cut implemented fuzzy clustering algorithms and switching regressions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 38, no. 3, pp. 588–603, 2008.