Research Article

Negative Correlation Learning for Customer Churn Prediction: A Comparison Study

Ali Rodan,[1] Ayham Fayyoumi,[2] Hossam Faris,[1] Jamal Alsakran,[1] and Omar Al-Kadi[1]

[1]*King Abdulla II School for Information Technology, The University of Jordan, Amman 11942, Jordan*
[2]*College of Computer and Information Sciences, Al Imam Mohammad Ibn Saud Islamic University, Riyadh 11432, Saudi Arabia*

Correspondence should be addressed to Ali Rodan; a.rodan@ju.edu.jo

Recently, telecommunication companies have been paying more attention toward the problem of identification of customer churn behavior. In business, it is well known for service providers that attracting new customers is much more expensive than retaining existing ones. Therefore, adopting accurate models that are able to predict customer churn can effectively help in customer retention campaigns and maximizing the profit. In this paper we will utilize an ensemble of Multilayer perceptrons (MLP) whose training is obtained using negative correlation learning (NCL) for predicting customer churn in a telecommunication company. Experiments results confirm that NCL based MLP ensemble can achieve better generalization performance (high churn rate) compared with ensemble of MLP without NCL (flat ensemble) and other common data mining techniques used for churn analysis.

1. Introduction

Technological improvements have enabled data driven industries to analyze data and extract knowledge. Data mining techniques facilitate the prediction of certain future behavior of customers [1]. One of the most important issues that reduces profit of a company is customer churn, which is also known as customer attrition or customer turnover [2]. Customer churn can also be defined as the business intelligence process of finding customers that are about to switch from a business to its competitor [3].

In todays industries, abundance of choices helps customers get advantage of a highly competitive market. One can choose a service provider that offers better service than others. Therefore, the profitmaking organizations which compete in saturated markets such as banks, telecommunication and internet service companies, and insurance firms strongly focused more on keeping current customers than acquiring new customers [4]. Moreover, maintaining current customers is proven to be much less expensive than acquiring new customers [5].

In order to keep their customers, companies need to have a deep understanding of why churn happens. There are several reasons to be addressed, such as dissatisfaction from the company, competitive prices of other companies, relocation of the customers, and customers' need for a better service which can lead customers to leave their current service provider and switch to another one [6].

Among the previous studies for churn analysis, one of the most frequently used method is artificial neural networks (ANNs). In order to fine-tune the models developed, several topologies and techniques were investigated with ANNs, such as building medium-sized ANN models which were observed to perform the best and making experiments in many domains such as pay-TV, retail, banking, and finance [7]. These studies indicate that a variety of ANN approaches can be applied to increase prediction accuracy of customer churn. In fact, the use of neural networks in churn prediction has a big asset in respect of other methods used, because the likelihood of each classification made can also be determined. In neural networks each attribute is associated with a weight and combinations of weighted attributes participate in the prediction task. The weights are constantly updated during the learning process. Given a customer dataset and a set of predictor variables the neural network tries to calculate a combination of the inputs and to output the probability that the customer is a churner.

On the other hand, data collected for churn prediction is usually imbalanced, where the instances in the nonchurner customer outnumber the instances in the churner class. This is considered as one of the most challenging and important problems since common classification approaches tend to get good accuracy results for the large classes and ignore the small ones [8]. In [9], the authors discussed different approaches for handling the problem of imbalanced data for churn prediction. These approaches include using more appropriate evaluation metrics, using cost-sensitive learning, modifying the distribution of training examples by sampling methods, and using Boosting techniques. In [8], authors added ensemble classifiers as a fourth category of approaches for handling class imbalance. It has been shown that ensemble learning can offer a number of advantages over a single learning machine (e.g., neural network) training. Ensemble learning has a potential to improve generalization and decrease the dependency on training data [10].

One of the key elements for building ensemble models is the "diversity" among individual ensemble members. Negative correlation learning (NCL) [11] is an ensemble learning technique that encourages diversity explicitly among ensemble members through their negative correlation. However, few studies addressed the impact of diversity on imbalanced datasets. In [12], authors indicated that NCL brings diversity into ensemble and achieve higher recall values for minority class comparing NCL to independent ANNs.

Motivated by these possible advantages of NCL for class imbalance problems, in this paper, we apply the idea of NCL to an ensemble of multilayer perceptron (MLP) and investigate its application for customer churn prediction in the telecommunication market. Each MLP in the ensemble operates with a different network structure, possibly capturing different features of the input stream. In general, the individual outputs for each MLP of the ensemble are coupled together by a diversity-enforcing term of the NCL training, which stabilizes the overall collective ensemble output.

Moreover, the proposed ensemble NCL approach will be assessed using different evaluation criteria and compared to conventional prediction techniques and other special techniques proposed in the literature for handling class imbalance cases.

The paper has the following organization. Section 2 gives a background on data mining techniques used in the literature for churn analysis. In Section 3 we introduce negative correlation learning and how to use it to generate an ensemble of MLP with "diverse" members. Churn dataset description is given in Section 4. Experiments and results are presented in Section 5. Finally, our work is concluded in Section 6.

2. Related Work

Data mining techniques that are used in both researches and real-world applications generally treat churn prediction as a classification problem. Therefore, the aim is to use past customer data and classify current customers into two classes, namely, prediction churn and prediction nonchurn [7]. There

have also been made a few academic studies on clustering and association rule mining techniques.

After having done the necessary data collection and data preprocessing tasks and labeling the past data, features that are relevant to churn need to be defined and extracted. Feature selection, also known as dimension reduction, is one of the key processes in data mining and can alter the quality of prediction dramatically. Preprocessing and feature selection are common tasks that are applied before almost every data mining technique. There are different widely used algorithms for churn analysis, for instance, decision trees; by its algorithm nature, the obtained results can easily be interpreted and therefore give the researcher a better understanding of what features of customer are related to churn decision. This advantage has made decision trees one of the most used methods in this field. Some applications of decision trees in churn prediction include building decision trees for all customers and building a model for each of the customer segments [13]. Another relatively new classification algorithm is support vector machine (SVM). It is widely used in data mining applications, particularly in complex situations. SVM algorithm is proven to outperform several other algorithms by increasing the accuracy of classification and prediction of customers who are about to churn [14, 15].

Some other algorithms might be appropriate for customer churn prediction, as the artificial neural networks (ANN), which is another supervised classification algorithm that is used in predicting customer turnover. However, this algorithm is expected to give more accurate results when used in a hybrid model together with other algorithms or with another ANN classifier [16]. Another example could be genetic programming (GP). Genetic programming is an evolutionary method for automatically constructing computer program. These computer programs could be classifiers or regressors represented as trees. The authors in [17] used a GP based approach for modeling a churn prediction problem in telecommunication industry.

Moreover, Bayesian classification is one of the techniques which was also mentioned to be used in churn prediction. This method depends on calculating probabilities for each feature and its effect on determining the class value, which is the customer being a churner or not. However, Bayesian classification may not give satisfactory results when the data is highly dimensional [18].

Lastly, it is worth to mention k-nearest neighbor (k-NN) and random forests as another two classification methods applied in literature for churn prediction. k-NN is a classification method where an instance is classified by a majority vote of its neighbors, where, on the other hand, Random forests are an ensemble of decision trees which are generated from the bootstrap samples. The authors in [19] applied both k-NN and random forests to evaluate the performance on sampled and reduced features churn dataset. For some of related work for churn prediction methods, see Table 1.

3. Negative Correlation Learning (NCL)

NCL has been successfully applied to training multilayer perceptron (MLP) ensembles in a number of applications,

TABLE 1: Related work for churn prediction methods.

| Author	Method	Description
Idris et al. [17]	GP	GP is applied with AdaBoost for churn prediction
Tsai and Lu [16]	ANN with BP	Applied as a hybrid approach in two stages (i.e., reduction and prediction)
Wang and Niu [14]	SVM	Least squares support vector machine (LS-SVM) is applied to establish a prediction model of credit card customer churn
Eastwood and Gabrys [31]	IBK	Authors apply simple k-nearest neighbor algorithm on a nonsequential representation of sequential data for churn prediction
Kraljević and Gotovac [32]	Decision trees	Decision trees (DT) were applied and compared with ANN and logistic regression. DT results outperform other models
Verbraken et al. [33]	Naive Bayes	Number of Bayesian Network algorithms, ranging from the Naive Bayes classifier to General Bayesian Network classifiers are applied for churn prediction
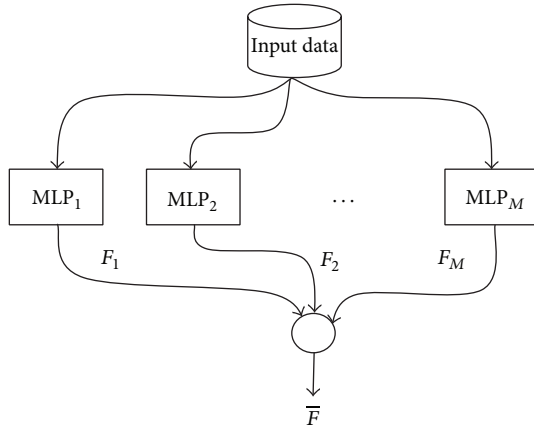
FIGURE 1: Ensemble of MLP networks.

including regression problems [20], classification problems [21], or time series prediction using simple autoregressive models [11].

In NCL, all the individual networks are trained simultaneously and interactively through the correlation penalty terms in their error functions. The procedure has the following form. Given a set of M networks and a training input set s, the ensemble output $F(t)$ is calculated as a flat average over all ensemble members (see Figure 1) $F_i(t)$:

$$F(t) = \frac{1}{M} \sum_{i=1}^{M} (F_i(t)). \tag{1}$$

In NCL the penalised error function to be minimised reads as follows:

$$E_i = \frac{1}{2} (F_i(t) - y(t))^2 + \lambda p_i(t), \tag{2}$$

where

$$p_i(t) = (F_i(t) - F(t)) \sum_{i \neq j} (F_j(t) - F(t)), \tag{3}$$

and $\lambda > 0$ is an adjustable strength parameter for the negative correlation enforcing penalty term p_i. It can be shown that

$$E_i = \frac{1}{2} (F_i(t) - y(t))^2 - \lambda (F_i(t) - F(t))^2. \tag{4}$$

Note that when $\lambda = 0$, we obtain a standard decoupled training of individual ensemble members. Standard gradient-based approaches can be used to minimise E by updating the parameters of each individual ensemble member.

3.1. Ensembles of MLPs Using NCL. Negative correlation learning (NCL) has been successfully applied to training MLP ensembles [10, 11, 20, 21]. We apply the idea of NCL to the ensemble of multilayer perceptron (MLPs) for predicting customer churn in a telecommunication company. Each MLP neural network operates with a different hidden layer, possibly capturing different features of the customer churn data. Crucially, the individual trained weights of the ensemble are coupled together by a diversity-enforcing term of the NCL training, which stabilises the overall collective ensemble output.

4. Dataset Description

The data used in this work is provided by a major Jordanian cellular telecommunication company. The data set contains 11 attributes of randomly selected 5000 customers for a time interval of three months. The last attribute indicates whether the customer churned (left the company) or not. The total number of churners is 381 (0.076 of total customers). The attributes along with their description are listed in Table 2.

The data is normalized by dividing each variable by its standard deviation. Normalization is recommended when data variables follow different dynamic ranges. Therefore, to eliminate the influence of larger values, normalization is applied to make all variables lie in the same scale.

4.1. Evaluation Criteria. In order to assess the developed model and compare it with different data mining techniques used for churn analysis, we use the confusion matrix shown in Table 3 which is the primary source for evaluating classification models. Based on this confusion matrix, the following three different criteria are used for the evaluation:

(1) accuracy: measuring the rate of the correctly classified instances of both classes,

$$\text{Accuracy} = \frac{(tp + tn)}{(tp + fn + fp + tn)}, \tag{5}$$

Table 2: List of attributes.

Attribute name	Description
3G	Subscriber is provided with 3G service (yes, no)
Total consumption	Total monthly fees (calling + sms) (JD)
Calling fees	Total monthly calling fees (JD)
Local sms fees	Monthly local sms fees (JD)
International sms fees	Monthly fees for international sms (JD)
International calling fees	Monthly fees for international calling (JD)
Local sms count	Number of monthly local sms
International sms count	Number of monthly international sms
International MOU	Total of international outgoing calls in minutes
Total MOU	Total minutes of use for all outgoing calls
On net MOU	Minutes of use for on-net-outgoing calls
Churn	Churning customer status (yes, no)

Table 3: Confusion matrix.

	Predicted	
	Nonchurn	Churn
Actual nonchurn	tp	fn
Actual churn	fp	tn

(2) hit rate: measuring the rate of predicted churn in actual churn and actual nonchurn,

$$\text{Hit rate} = \frac{tn}{(fn + tn)}, \qquad (6)$$

(3) actual churn rate: measuring the rate of predicted churn in actual churn,

$$\text{Churn rate} = \frac{tn}{(fp + tn)}. \qquad (7)$$

5. Experiments and Results

5.1. Experimental Setup. In literature, some authors studied the effect of ensemble size on the performance. For example Hansen and Salamon in [22] suggested that ensembles with a small size as ten members were adequate to sufficiently reduce test-set error [23]. In our current paper we used empirical approach to investigate the appropriate size of the ensemble. For ensemble of networks, we tried $(4, 6, 8, 10, 12, \ldots, 20)$ networks in the hidden layer and then checked the performance each time. The best performance reached without

Table 4: Selected ensemble of MLP using NCL parameters based on 5-fold cross validation.

Parameter	Value
Hidden layers	1
Ensemble size (M)	10
Decay	0.001
hidden Layer nodes (N)	10
Activation function	$1/(1 + e(-x))$
Learning rate (η)	0.3
Momentum	0.2
λ	0.5

overfitting was for an ensemble of size of 10 networks. Therefore, the ensemble used in all our experiments consists of $M = 10$ MLP networks. In all experiments we use MLPs with hidden layer of $N = 10$ units. We used NCL training via backpropagation learning algorithm (BP) on E with learning rate $\eta = 0.3$. The output activation function of the MLP is sigmoid logistic.

We optimize the penalty factor λ and the number of hidden nodes using 5-fold cross validation, and λ is varied in the range $[0, 1]$ (step size 0.1) [10]. The number of hidden nodes is varied from 1 to 20 (step 1). Based on 5-fold cross validation, the details of the selected ensembles of MLPs using NCL parameters are presented in Table 4. Note that the selected parameters for the flat ensembles of MLPs (without NCL) are the same as with NCL except that there are no NCL λ parameters. The ensembles used in our experiments are also compared with the following common data mining techniques used in the literature:

 (i) k-nearest neighbour (IBK),

 (ii) Naive Bayes (NB),

 (iii) random forest (RF),

 (iv) genetic programming,

 (v) single MLP neural network trained with BP,

 (vi) C4.5 decision trees algorithm,

 (vii) support vector machine (SVM).

As churn data is imbalanced, NCL is also compared with other special techniques proposed in the literature for handling class imbalance cases. These techniques include:

 (i) AdaBoost algorithm with C4.5 Decision Tree as base classifier (AdaBoost), [24].

 (ii) Bagging algorithm with C4.5 Decision Tree as base classifier (Bagging) [25],

 (iii) MLP for cost-sensitive classification (NNCS) [26],

 (iv) synthetic minority oversampling technique with MLP as base classifier (SMOTE + MLP) [27],

 (v) neighborhood cleaning rules with constricted particle swarm optimization as base classifier (NCR + CPSO) [28–30],

 (vi) neighborhood cleaning rules with MLP as base classifier (NCR + MLP) [28].

TABLE 5: Tuning parameters for data mining techniques used in the comparison study.

Method	Parameters
GP	Population size = 1000, Maximum number of generations = 100, functions = {*, /, −, +, IF, <, >}, tree max depth = 10, tree max length = 30 elites = 1, selection mechanism = tournament selection crossover point probability = 90%, mutation = probability 15%
ANN with BP	Activation function = Sigmoid, Epoches = 5000, Learning Rate = 0.3, Momentum = 0.2
SVM	Cost = 1, Gamma = 10000
IBK	Number of neighbours = 1, nearest neighbor search algorithm = linear search (brute force search)
AdaBoost	Number of classifiers = 10
Bagging	Number of classifiers = 10
NNCS	Hidden layers = 2, hidden nodes = 15
SMOTE	Number of neighbors = 5
NCR + CPSO	Number of neighbors = 5 for SMOTE, number of particles = 75 for CPSO

TABLE 6: Evaluation results (results of best five models are shown in bold).

Model	Accuracy	Actual churn rate	Hit rate
k-Nearest neighbour (IBK)	0.927	0.022	0.067
Naive Bayes (NB)	0.597	0.901	0.115
Random Forest (RF)	0.940	0.006	0.109
Genetic programing (GP)	0.759	0.638	0.142
Single ANN with BP	0.941	0.625	0.607
Decision tress (C4.5)	**0.975**	0.703	**0.964**
Support vector machine (SVM)	**0.977**	0.703	**0.992**
AdaBoosting	**0.972**	0.719	**0.898**
Bagging	**0.975**	0.703	**0.954**
MLP for cost-sensitive classification (NNCS)	0.496	0.819	0.113
SMOTE + MLP	0.722	0.724	0.177
NCR + CPSO	0.894	0.827	0.694
NCR + MLP	0.642	0.751	0.144
Flat ensemble of ANN	0.958	0.732	0.725
Ensemble of ANN using (NCL)	**0.971**	**0.803**	**0.814**

Table 5 presents the empirical settings of the selected models parameters of these common data mining techniques based on 5-fold cross validation. For SVM, cost and gamma parameters are tuned using simple grid search. In C4.5 algorithm, the attributes are computed using the discriminant $P(\omega_0 \mid A = 0)$ where $P(\cdot)$ is the conditional probability and ω_i is the classification type. For naive Bayesian, the transcendent is $p(x \mid w_i)P(w_i) - p(x \mid w_j)P(w_j) = 0$ where w_i and w_j are the classification types, $p(x \mid w)$ is the condition probability density, and $P(w)$ is the transcendent probability.

5.2. Results and Comparison.

Table 6 summarizes the results of negatively correlated ensemble of MLPs, independent ensemble of MLPs ($\lambda = 0$), and other data mining models used in the literature for customer churn prediction. In order to assess the improvement achieved by using a genuine NCL training against independent training of ensemble members ($\lambda = 0$), the MLP networks are initialized with the same weight values in both cases (see Table 4). The MLP ensemble trained via NCL outperformed independent ensemble of MLPs model and all the other models used in this study.

Note that the two MLP ensemble versions we study share the same number of free parameters, with the sole exception of the single diversity-imposing parameter λ in NCL based learning.

According to Table 6, it can be noticed that NCL achieved high accuracy rate (97.1%) and ranked among the best five techniques (results are shown in bold font) with a slight difference (i.e, 0.6%) behind the best which is SVM. The flat ensemble ANN comes after with 95.8% accuracy. As mentioned earlier, accuracy is not the best evaluation metric when the examined problem is highly imbalanced. Therefore, we need to check other criteria; in our case, they are the churn rate and the hit rate. Looking at the obtained churn rates, NCL comes second with 80.3% churn rate and significant increase of 10% over SVM. It is important to indicate here that

although NB is the best in churn rate with 90.1%, it is the 2nd worst in terms of accuracy rate which is 59.7%. This gives a great advantage of NCL over NB. Finally, the hit rates of IBK, NB, RF, GP, NNCS, SMOTE + MLP, and NCR + MLP show very poor results so they can be knocked off the race. On the other hand, NCL and ensemble ANN show acceptable results for hit rate of 81.4% and 72.5%, respectively.

6. Conclusion

Customer churn prediction problem is important and challenging at the same time. Telecommunication companies are investing more in building accurate churn prediction model in order to help them in designing effective customer retention strategies. In this paper we investigate the application of an ensemble of multilayer perceptron (MLPs) trained through negative correlation learning (NCL) for predicting customer churn in a telecommunication company. Experimental results confirm that NCL ensembles achieve better generalization performance in terms of churn rate prediction with highly acceptable accuracy error rate compared with flat ensembles of MLPs and other common machine learning models from the literature.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

References

[1] 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.

[2] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Computer assisted customer churn management: state-of-the-art and future trends," *Computers and Operations Research*, vol. 34, no. 10, pp. 2902–2917, 2007.

[3] S. A. Qureshi, A. S. Rehman, A. M. Qamar, and A. Kamal, "Telecommunication subscribers' churn prediction model using machine learning," in *Proceedings of the 8th International Conference on Digital Information Management (ICDIM '13)*, pp. 131–136, September 2013.

[4] G.-E. Xia and W.-D. Jin, "Model of customer churn prediction on support vector machine," *System Engineering—Theory & Practice*, vol. 28, no. 1, pp. 71–77, 2008.

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

[6] Z. Kasiran, Z. Ibrahim, and M. S. M. Ribuan, "Mobile phone customers churn prediction using elman and Jordan Recurrent Neural Network," in *Proceedings of the 7th International Conference on Computing and Convergence Technology (ICCCT '12)*, pp. 673–678, December 2012.

[7] A. Sharma and P. K. Panigrahi, "A neural network based approach for predicting customer churn in cellular network services," *International Journal of Computer Applications*, vol. 27, no. 11, pp. 26–31, 2011.

[8] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man and Cybernetics C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.

[9] J. Burez and D. van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009.

[10] G. Brown and X. Yao, "On the effectiveness of negative correlation learning," in *Proceedings of the 1st UK Workshop on Computational Intelligence*, 2001.

[11] Y. Liu and X. Yao, "Ensemble learning via negative correlation," *Neural Networks*, vol. 12, no. 10, pp. 1399–1404, 1999.

[12] S. Wang, K. Tang, and X. Yao, "Diversity exploration and negative correlation learning on imbalanced data sets," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '09)*, pp. 3259–3266, June 2009.

[13] 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.

[14] N. Wang and D.-X. Niu, "Credit card customer churn prediction based on the RST and LS-SVM," in *Proceedings of the 6th International Conference on Service Systems and Service Management (ICSSSM '09)*, pp. 275–279, Xiamen, China, June 2009.

[15] A. Rodan, H. Faris, J. Alsakran, and O. Al-Kadi, "A support vector machine approach for churn prediction in telecom industry," *International Journal on Information*, vol. 17, no. 8, pp. 3961–3970, 2014.

[16] 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.

[17] A. Idris, A. Khan, and Y. S. Lee, "Genetic programming and adaboosting based churn prediction for telecom," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC '12)*, pp. 1328–1332, Seoul, Republic of Korea, October 2012.

[18] 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.

[19] A. Idris, M. Rizwan, and A. Khan, "Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies," *Computers & Electrical Engineering*, vol. 38, no. 6, pp. 1808–1819, 2012.

[20] G. Brown, J. L. Wyatt, and P. Tiňo, "Managing diversity in regression ensembles," *The Journal of Machine Learning Research*, vol. 6, pp. 1621–1650, 2005.

[21] R. McKay and H. Abbass, "Analyzing anticorrelation in ensemble learning," in *Proceedings of the Conference on Australian Artificial Neural Networks and Expert Systems*, pp. 22–27, 2001.

[22] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[23] D. Opitz and R. Maclin, "Popular ensemble methods: an empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[24] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[25] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[26] Z. H. Zhou and X. Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63–77, 2006.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[28] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Proceedings of the 8th Conference on AI in Medicine in Europe (AIME '01)*, vol. 2001 of *Lecture Notes on Computer Science*, pp. 63–66, Springer, 2001.

[29] T. Sousa, A. Silva, and A. Neves, "Particle swarm based data mining algorithms for classification tasks," *Parallel Computing*, vol. 30, no. 5-6, pp. 767–783, 2004.

[30] H. Faris, "Neighborhood cleaning rules and particle swarm optimization for predicting customer churn behavior in telecom industry," *International Journal of Advanced Science and Technology*, vol. 68, pp. 11–12, 2014.

[31] M. Eastwood and B. Gabrys, "A non-sequential representation of sequential data for churn prediction," in *Knowledge-Based and Intelligent Information and Engineering Systems*, J. D. Velásquez, S. A. Ríos, R. J. Howlett, and L. C. Jain, Eds., vol. 5711 of *Lecture Notes in Computer Science*, pp. 209–218, Springer, Berlin, Germany, 2009.

[32] G. Kraljevíc and S. Gotovac, "Modeling data mining applications for prediction of prepaid churn in telecommunication services," *Automatika*, vol. 51, no. 3, pp. 275–283, 2010.

[33] T. Verbraken, W. Verbeke, and B. Baesens, "Profit optimizing customer churn prediction with Bayesian network classifiers," *Intelligent Data Analysis*, vol. 18, no. 1, pp. 3–24, 2014.