MCS: Multiple Classifier System to Predict the Churners in the Telecom Industry

Mehreen Ahmed
Department of Computer Software Engineering
National University of Sciences & Technology,
Islamabad, Pakistan
mahreenmcs@gmail.com

Imran Siddiqi
Department of Computer Science
Bahria University,
Islamabad, Pakistan
imran.siddiqi@gmail.com

Hammad Afzal
Department of Computer Software Engineering
National University of Sciences & Technology,
Islamabad, Pakistan
hammad.afzal@mcs.edu.pk

Behram Khan
BT Research,
United Kingdom
behram.khan@bt.com

Abstract—**Multiple classifiers for prediction or classification has gained popularity in recent years. Ensemble Technique perform best predictions as compared to traditional classifiers. This has resulted in the experimentation with new ways of ensemble creation. This paper presents a multiple classifier system (MCS) that can outperform traditional classifiers. Experiments are performed on a benchmark Customer Churn Dataset (available on UCI repository) and a newly created dataset from a South Asian wireless telecom operator. MCS achieved accuracies of 97% and 86% on the UCI churn dataset and private dataset, respectively. MCS as compared to existing best approaches realized the best results on the private and public datasets.**

Keywords—Churn prediction; multiple classifier system; classification; machine learning; telecom

I. Introduction

Customer churn prevention is a major part of the Customer Relationship Management (CRM) in any business. Same is the norm in telecom industry where, due to immense competition among the telecom carriers, there is a dire need for churn management for the operators to retain their current subscribers. Churn describes the subscribers who terminate their relationship with the service provider and move their subscription to the competitor. The subscribers choose the carrier that offers new incentives through advertisements, often fulfilling their needs with cheaper services as compared to their current operator. With such competition among the telecom operators, churn prediction has gained pivotal role in identifying beforehand if the current customer is planning to leave the carrier. The operators can then come up with win-back strategies to offer the customer some better deals. Churn prevention is an important factor as the cost of customer acquisition is much greater than the cost of customer retention [1].

Machine learning techniques have vastly been applied in predictive analysis that can provide accurate classification of the churners and non-churners. Among the most popular techniques are the decision trees [2-5]. Support Vector Machines and Neural Networks based solutions have also been employed in a number of studies [6-8]. Recently, ensemble technique such as Voting, Stacking, Bagging and Boosting are employed with greater success as they consider using a combination of learners instead of one that in turn increases the classification rate [9].

In this paper, a multiple classifier system (MCS) is introduced that is applied on two datasets; one publicly available dataset, that comprises of different types of 20 features and 5000 records. The other is a private dataset, collected from a South Asian wireless telecom operator, with 13 features and 2000 records, collected between Aug 2015 to Sep 2015.The results show that the proposed framework outperformed the current best results, giving an accuracy of 97% on the UCI Dataset. MCS was applied on a private churn dataset, which is a balanced dataset with equal number of churners and non-churners. The proposed technique produced 86% accurate results.

This paper is organized as: In Section 2, a literature review is presented on churn prediction. Section 3 presents the data sets used in this study and Section 4 introduces the multiple classifier system. The results are discussed in Section 5. Finally conclusions are drawn in Section 6.

II. Literature Review

Research shows that various traditional data mining classification algorithms have been applied on private and public telecom churn data sets. The popular techniques that have been applied include Decision Tree Based approaches, Artificial Neural Networks and Statistical methods. With these methods the highest accuracy of 93.7% was achieved on the public UCI data set using Multi-Layer Perceptron Neural Network [11]. However conflicts do arise among researchers as per which of these techniques performs better on the churn data sets, as for some Neural Network performs better while for others Decision Trees outperformed as can be seen in Table 1. Lately ensemble methods have been applied that produced

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

better accuracy than the traditional methods. On public data sets, C. Tsai and Y. Lu [12] used a hybrid technique which combined two Neural Networks. The first Neural Network performed the data reduction and the second predicted the potential churners. They achieved an accuracy of 94.32% using this hybrid model. A. Lemmens and C. Croux [13] found that ensemble techniques including Bagging and Boosting outperformed the basic CART after a few iterations while evaluating with measures like Top Decile and GINI coefficient. Recently, T. Vafeiadis et al. [14] concluded that boosted counterparts (using Adaptive Boosting) of the traditional classifiers like SVM, DT and Back-Propagation Network (BPN) gave much better performance. The highest accuracy achieved so far on the UCI data set is 96.86% with the boosted SVM (SVM-POLY kernel using AdaBoost) technique. Hence boosting technique when applied on traditional classifier methods, gave a higher accuracy.

Some studies use private telecom data sets which have different set of features and number of samples. However these data sets are unavailable and thus the results cannot be reproduced. Most of these studies only use a limited amount of techniques. In 2015, A. Rodan et al. [15] applied Negative Correlation Learning on an ensemble of ten Multi-Layer Perceptron networks to get an accuracy of 97.1% on a private Jordanian Telecom Company Dataset with 5000 samples.

Some studies use some unconventional ensemble approaches. Like in 2014, J. Xiao et al. [16] proposed a transfer learning approach named the Feature Selection Based Dynamic Transfer Ensemble (FSDTE) in which they combined transfer learning, Multiple Classifier Ensemble (MCE) and Group Method of Data Handling (GMDH) type neural network. Their approach achieves higher classification accuracy. An 80.8% sensitivity was achieved on the UCI data set. In 2015, A. Baumann et al. [17] proposed a framework called the Decision Centric Ensemble Selection (DCES), they employed the ensemble approach for predictive modeling in which they choose the candidates from a model library that gives the highest lift for a data set. Different combinations of ensemble models were observed for the UCI dataset, giving a Top Decile Lift of 6.821 with random forest and logistic regression ensemble.

TABLE I. RESULTS ON CHURN PREDICTION IN TELECOM INDUSTRY USING UCI PUBLIC DATA SET AND OTHER PRIVATE DATA SETS

Paper	Data Set	Technique	Results
Traditional Algorithms			
[1] 2008	UCI	SVM using RBF Kernel	90.9% Accuracy
[2] 2011	UCI	Neural Network	92.35% Accuracy
[3] 2014	UCI	Multi-Layer Perceptron	**93.7% Accuracy**
[4] 2006	Taiwan Wireless Telecom Company	Back Propagation Neural Network	Hit ratio: 98% LIFT: 9.96 (at 10%)
[5] 2006	British Telecom Company	Decision Tree	82% Accuracy
[6] 2010	Taiwan Telecom Company	Decision Tree	90.98% Accuracy
[7] 2010	HT Mostar	Decision Tree	91% Accuracy
[8] 2015	China Mobile Operator	C4.5	89% Accuracy
Ensemble Techniques			
[9] 2014	UCI	MCE + GMDH Type Neural Network	80.8% Sensitivity 86.9 AUC
[10] 2015	UCI	Boosted Support Vector Machine	**96.86% Accuracy**
[11] 2015	UCI	Regularized Logistic Regression + Random Forest	6.821 TDL
[12] 2012	East Asian Mobile Operator	Boosted Decision Tree ADT	97.4 AUC
[13] 2012	Singapore Telco Company	Random Forest	93.4% Accuracy
[14] 2015	Jordanian Telecom Company	10 MLP Ensemble with Negative Correlation Learning	**97.1% accuracy**

III. DATA SETS

In the present study, two different data sets of customer relationship management (CRM) are used to build predictive models and assess the performance. One publicly available data set from the UCI machine learning data repository that has

5000 samples and 20 attributes with 14.3% churn rate (data set 1). The second data set used is real life data collected from a major wireless telecom operator in South Asia (data set 2).

Most of the attributes in the data sets are associated with call detail records (CDR), billing and personal information. For data set 2, the carrier provided data containing 2000 subscribers. All of these subscribers were not contract based and had a monthly based subscription. The subscriber data was extracted from the time interval of two months, i.e. August and September 2015. To overcome the class imbalance problem in churn data sets, an equal amount of churners and active subscribers were collected from the carrier. So the data has 50% churner's information.

A. Input Features

Churn occurs due to the dissatisfaction of a subscriber with their present service provider. The reason for this dissatisfaction may be due to a number of reasons which may include poor service or pricing. Over the years, researchers have introduced some unique set of features in their churn prediction models

For data set 2, the revenue information, customer account and usage details were used. Some new features introduced were the information regarding the data usage (data volume and revenue). One other important new feature used was the favorite other network information. Below these features are discussed in detail.

1) Revenue Details:
Revenue is the amount of the money that the company receives during a time period, in this case, for the months of August and September 2015. The revenue generated for SMS, calls, data, the off-network, on-network and the total overall monthly revenue was collected for both the months (Aug, Sep) and then aggregated.

a) Aggregate of Total Revenue: The overall monthly revenue earned in Rupees by the carrier in the months August & September 2015.

b) Aggregate of SMS Revenue: The revenue earned through the SMS service used by the subscriber.

c) Aggregate of Data Revenue: The revenue earned through the Data service used by the subscriber.

d) Aggregate of Off Net Revenue: The revenue earned by the calls, etc. made to the off-network (not the same network as the subscriber) customers by the carrier's present subscriber.

e) Aggregate of On Net Revenue: The revenue earned by the calls etc. made to the on-network (on the same network as the subscriber) customers by the carrier's present subscriber.

2) Subscriber's Account Details:

a) Network Age: The time passed since the subscriber started using the services of the carrier.

b) Package Name: The names of the packages the subscriber has registered. The carrier offers a number of packages. This information can help the carrier in knowing the demands of the subscriber and can have a huge impact on churn information.

c) User Type: This detail helps in knowing if the user is subscribed to a 2G or 3G service.

d) Aggregate of Complaint Count: The number of complaints made by the subscribers.

3) Subscriber's Usage Details:

a) Favorite Other Network: This information can certainly have a huge impact on churn ratio as it gives the information about which other network or operator the subscribers makes the most of the calls to and thus might influence the customer to move to that network to save money.

b) Aggregate of Data Volume: The volume of the data service used by the subscriber.

IV. Multiple Classifier System

Researchers have proposed techniques for generation of multiple classifier systems [15, 16]. MCS was built following the overproduce and choose approach [17]. This system is a hybrid ensemble created by combining the boosting and stacking techniques.

Adaptive Boosting (AdaBoost) [18] was introduced by Yoav Freund and Robert E. Schapire in 1995. The boosting algorithm works with labeled training data set $D = (s_i, t_i)$ where s_i are the instances of i samples, $i=1 \ldots N$ and t_i is the associated label with every instance s_i. On every iteration k = 1 … T, a weight w_k is assigned to each sample s_i of the training data set D. The weak learner is trained to get a weak hypothesis $h_k(s_i) = t_i$ .Then the learning error ε_k $\{where\ \varepsilon_k = p(h_k(s_i) \neq t_i)\}$ is calculated and the weights are updated. The weights are updated until the last iteration T is reached. The AdaBoost algorithm, adapts to the errors of the weak hypothesis that the weak learner produces, unlike its predecessors. A single prediction rule from the combination of the entire weak hypothesis is generated. AdaBoost solves over fitting problems by focusing on the misclassified examples. If the sample is misclassified, the assigned weights will increase and decrease for correctly classified samples. AdaBoost selects the most informative samples on every iteration *k*. AdaBoost converts a weak learning algorithm into a strong one [19].

Stacking or Stacked Generalization [20], proposed by David H. Wolpert, is the combination of diverse heterogeneous learning algorithms applied on a data set. This meta-model has base models referred as level-0 learners and a level-1 learner. Level-1 learner combines the set of outputs of the level-0 learners and corrects their mistakes there by improving the classification results. Stacking generates a meta-dataset, comprising the tuples of the original dataset using the predictions made by the classifiers as the input attributes. With the same target attribute for both the original and the stacked data set. This meta-model combines these output predictions of the level-0 learners into a single level-1 prediction.

The data set was preprocessed in the form of numeric data transformation of textual attributes, feature selection and using sampling techniques. And other preprocessing tasks including

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

class labeling, formation of training/test sets and creation of derived variables, like aggregate monthly revenue that is the sum of the monthly revenues.

The proposed framework is made up of Boosted and Stacked Learners. MCS gathers the pre-processed churned data and assigns weights to form a weak model that goes through different iterations *T* forming diverse set of models. In each iteration the weights are modified and a meta-dataset is made, thus forming a new model. In a normal boosted learner, there is a single (base) learner. In MCS, the learner that is boosted is a 3-stacked learner. Stacked model takes three stacked learners (level 0 or base learners) to form a meta (level 1) learner. The three learners used as stacked (level 0) learners are neural network, decision tree and K-nearest neighbor heterogeneous algorithms. A majority vote is taken on the predictions of these diverse models. The framework can be seen in Fig. 1.

V. Experimental Setup and Results

The dataset after preprocessing is partitioned into training and test sets. Both the datasets (Dataset 1 and 2) were randomly subdivided into training (75%) and test sets (25%). The experiments are performed with individual and hybrid classifiers in comparison to the MCS. The test set is kept hidden and MCS is tested on the test set of the UCI dataset (Dataset 1) and the private dataset (Dataset 2).

Tables 2 and 3 show the result of the Multiple Classifier System as compared to the traditional classifiers on the Datasets 1 and 2. It can be seen that the accuracy increases in the case of MCS with both datasets (Dataset 1 and 2).

TABLE II. Results of MCS as Comparded to Individual Classifiers on Dataset 1 (UCI Churn Dataset)

Classifiers	**Accuracy (%)**
K Nearest Neighbor	91.73
Artificial Neural Network	**96.27**
Decision Tree	**95.87**
Naïve Bayesian	88.27
Logistic Regression	85.73
Multiple Classifier System	**97.2**

MCS was also compared with the best proposed techniques to date. Tables 4 and 5 show the comparison of MCS with these techniques. The proposed framework gave the best performance as compared to the literature.

TABLE III. Results of MCS as Comparded to Individual Classifiers on Dataset 2 (South Asian Dataset)

Classifiers	**Accuracy (%)**
K Nearest Neighbor	69.67
Artificial Neural Network	**81.3**
Decision Tree	72
Naïve Bayesian	56
Logistic Regression	75.33
Multiple Classifier System	**86.33**

TABLE IV. Comparison of MCS with State-of-the-Art Best Results on Dataset 1 (UCI Churn Dataset)

Reference	**Year**	**Technique**	**Accuracy (%)**
[2]	2011	Neural Network	92.35
[3]	2014	Multi-Layer Perceptron	93.7
[9]	2014	MCE + GMDH Type Neural Network	86.9 AUC
[10]	2015	Boosted Support Vector Machine	96.86
Proposed Technique	**2016**	**Multiple Classifier System (MCS)**	**97.2**

In Table 5, it can be seen that the accuracy is less as compared to the public datasets; still the proposed framework gave better results on both kinds of datasets either private or public.

TABLE V. Comparison of Proposed Technique with the Previous Techniques for SATO Dataset

Reference/Year	**Technique**	**Accuracy (%)**
[7] 2010	Decision Tree	75
[2] 2011	Neural Network	72
[9] 2012	Random Forest	68
[21] 2014	Hybrid model of Voted Perceptron (VP) and Logistic Regression (LR)	59
[3] 2014	Multi-Layer Perceptron	73
[10] 2015	Boosted Support Vector Machine	70
2016	**Multiple Classifier System (MCS)**	**86.3**

978-1-5090-6435-9/17/$31.00 ©2017 IEEE

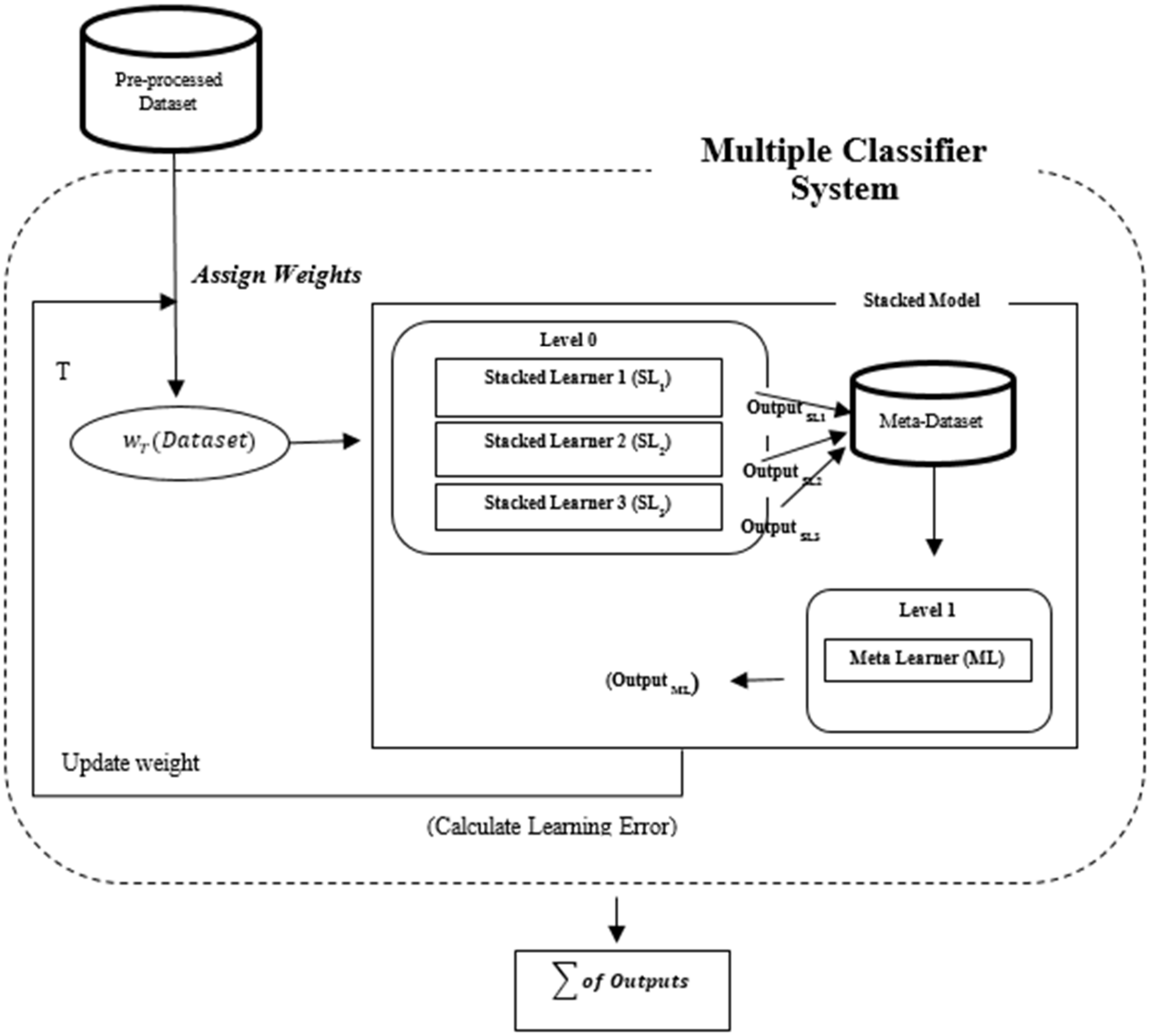

Fig. 1. System architecture of proposed Multiple Classifier System (MCS) with T iterations.

VI. Conclusion

The experiments are conducted on test sets and the results are evaluated. The Multiple Classifier System is applied on the test sets of both the public and private data set. The model is compared to individual classifiers and other best proposed techniques in previous studies. Combination of the Boosting and Stacking meta-classifiers proves that heterogeneous ensembles are more diverse and accurate. The remarkable result achieved with this amalgamation proves the strength of ensemble techniques. An accuracy of 97% was achieved with the proposed multiple classifier system on a benchmark dataset. In the future, further combinations or hybrids of classifiers can be explored to improve the predictive performance of private datasets and solutions to resolve the class imbalance issues in the churn datasets can be investigated.

References

[1] G.-e. Xia and W.-d. Jin, "Model of Customer Churn Prediction on Support Vector Machine," *Systems Engineering - Theory & Practice,* vol. 28, pp. 71-77, 2008/01 2008.

[2] A. Sharma and D. P. K. Panigrahi, "A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services," *International Journal of Computer Applications (0975 – 8887),* vol. 27, 2011.

[3] Ionut B. Brandusoiu and G. Toderean, "A Neural Networks Approach for Churn Prediction Modeling in Mobile Telecommunications Industry," *Annals of The University of Craiova,* vol. 11, pp. 9-16, 2014.

[4] S. Y. Hung, D. C. Yen, and H. Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications,* vol. 31, pp. 515-524, 2006/10 2006.

[5] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Churn prediction: Does technology matter," *International Journal of Intelligent Technology,* vol. 1, pp. 104-110, 2006.

[6] C.-F. Tsai and M.-Y. Chen, "Variable selection by association rules for customer churn prediction of multimedia on demand," *Expert Systems with Applications,* vol. 37, pp. 2006-2015, 2010.

[7] G. Kraljević and S. Gotovac, "Modeling data mining applications for prediction of prepaid churn in telecommunication services," *AUTOMATIKA: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije,* vol. 51, pp. 275-283, 2010.

[8] Y. Liu and Y. Zhuang, "Research Model of Churn Prediction Based on Customer Segmentation and Misclassification Cost in the Context of Big Data," *JCC,* vol. 03, pp. 87-93, 2015.

[9] J. Xiao, Y. Xiao, A. Huang, D. Liu, and S. Wang, "Feature-selection-based dynamic transfer ensemble model for customer churn prediction,"

Knowledge and Information Systems, vol. 43, pp. 29-51, 2014/01/16 2014.

[10] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory,* vol. 55, pp. 1-9, 2015/06 2015.

[11] A. Baumann, S. Lessmann, K. Coussement, and K. W. D. Bock, "Maximize What Matters: Predicting Customer Churn With Decision-Centric Ensemble Selection," presented at the ECIS 2015 2015.

[12] 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, 2012.

[13] C. Phua, H. Cao, J. B. Gomes, and M. N. Nguyen, "Predicting Near-Future Churners and Win-Backs in the Telecommunications Industry," 2012.

[14] A. Rodan, A. Fayyoumi, H. Faris, J. Alsakran, and O. Al-Kadi, "Negative Correlation Learning for Customer Churn Prediction: A Comparison Study," *The Scientific World Journal* vol. 2015 2015.

[15] L. Xu, A. Krzyżak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *Systems, Man and Cybernetics, IEEE Transactions on,* vol. 22, pp. 418-435, 1992.

[16] J. Kittler and F. Roli, "Multiple classifier systems," *Lecture notes in computer science,* 2002.

[17] A. J. Sharkey, N. E. Sharkey, U. Gerecke, and G. O. Chandroth, "The 'test and select' approach to ensemble combination," in *Multiple Classifier Systems*, ed: Springer, 2000, pp. 30-44.

[18] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational learning theory*, 1995, pp. 23-37.

[19] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence,* vol. 14, p. 1612, 1999.

[20] D. H. Wolpert, "Stacked generalization," *Neural networks,* vol. 5, pp. 241-259, 1992.

[21] G. Olle, "A Hybrid Churn Prediction Model in Mobile Telecommunication Industry," *IJEEEE,* 2014.