Predicting Customers Churn in Telecom Industry using Centroid Oversampling method and KNN classifier

Pragya Joshi
Department of Computer Engineering
Shri G.S. Institute of Tech. and Science
23, Sir Visvesvaraya Marg, Indore (MP)
Email: joshi.pragya@gmail.com

Surendra Gupta
Department of Computer Engineering
Shri G.S. Institute of Tech. and Science
23, Sir Visvesvaraya Marg, Indore (MP)
Email: sgupta@sgsits.ac.in

Abstract— Subscriber retention is a key issue for several service-based organizations mainly in telecom industry, where detecting customer's behaviour is one of the main interest. Customer churn occurs when an existing subscriber stops doing business or ends the relationship with a company. Since the proportion of churned customers are less as compared to the whole subscriber base, it leads to imbalanced data distribution.

To improve the classification performance of imbalanced data learning, a novel over-sampling method, Centroid Oversampling Technique, based on centroid of three nearest neighbor points, is proposed. It generates a set of representative data points to broaden the decision regions of small class spaces. The representative centroids are regarded as synthetic examples in order to resolve the imbalanced class problem. This approach addresses the problem of synthetic minority oversampling techniques, which overlaps the reflection of majority classes on training examples. Our comprehensive experimental results show that Centroid Oversampling Technique can achieve better performance than renowned multi-class oversampling methods.

Index-terms - Class Imbalance, Oversampling Techniques, Centroid method, K Nearest Neighbor.

I. INTRODUCTION

Techniques for handling class-imbalance problem can be broadly categorized into two sets basis their evaluation procedure, Oversampling and Undersampling. In oversampling, some of the data points of minority class are replicated to generate synthetic samples to form balanced dataset. Undersampling reduces the number of observations from majority class to balance the dataset, but removing observations might cause the training data to miss the significant information pertaining to majority class.

Best known oversampling techniques are SMOTE, ADASYN, MTDF. Synthetic Minority Oversampling Technique SMOTE[3] is considered as the most popular technique for handling CIP, as It produces new minority data points based on k-nearest neighbor(kNN), giving positive observations.

A major concern of SMOTE[3] is that the class with frequent samples dominate the neighboring data points of a test instance in-spite of distance measurements which leads to sub-optimal classification performance on the minority class. In this proposed approach, a modified oversampling technique is introduced which uses the concept of generating synthetic data samples by calculating centroid of three nearest points (kNN). The main objective of this approach is to correct the inherent bias to majority class on any distance measurement. So, the proposed work is aimed at modifying the existing oversampling technique to enhance the accuracy in predicting customer churn so that special tariffs can be assigned to retain them.

The main aim of this approach is to minimise the miss-classified errors while retaining a suitably high accuracy in representing the original features. Many machine learning algorithms yield low performance while dealing with large imbalanced datasets. Imbalanced classes can cause different classification approaches to face difficulties in learning, which can results in poor performance.

Therefore, in this paper, a technique has been proposed i.e. centroid oversampling technique. The basic idea of this section is to balance the class distribution using centroid method and kNN approach. In addition, we reduce the number of features and remove the not related, unnecessary or noisy information. Besides, this improves the presentation of information prediction with speeding up the processing algorithm. To improve the prediction accuracy, use PCA algorithm for dimension reduction.

This paper is ordered as follows. In section II, an overview of work already done in this area with their advantages and disadvantages is discussed. In the section, III the proposed approach is presented. Section IV presents the details of experimental analysis of the result. Conclusion and future work is discussed in section V.

II. Related Work

Substantial work has been carried out by earlier re-searchers in the field of classification techniques for Imbalanced datasets, handling miss-classified errors and related areas. Some of the most relevant techniques are discussed in this section.

SMOTE[3], Synthetic Minority Over-sampling method, is an oversampling method which uses the concept of creating new minority class samples by incorporating several minority class cases that lie close to each other. It is based on KNN approach, it chooses K nearest neighbors, merge them and generates new synthetic samples. The algorithm uses the feature vectors and its nearest neighbors, calculates the distance among these vectors. The variance is then multiplied by random number among (0, 1) and added back to feature.

SMOTE algorithm is a pioneer algorithm and countless other algorithms are derivative of SMOTE.

SMOTE[3] is an oversampling method which yields "synthetic" data points basis distance measure between two given data points. Oversampling is done using each minority class samples, generating new data examples which connects the nearest neighbors of that class samples. Neighbors are chosen randomly basis the ratio of oversampling being done.

The key most objective of oversampling techniques is to strengthen the minority class. The motive behind this algorithm is quite simple. Oversampling causes overfitting, and because of recurrent data instances, the decision boundary gets constricted. So, in SMOTE[3] paper, it has been shown that these newly constructed data samples are not exact replicas, but new points and thus it relaxes the decision boundary, thereby helping the algorithm to estimate the theory more precisely.

However, some advantages and disadvantages of SMOTE are listed below.

Advantages: 1.Overcomes over-fitting by generating synthetic examples rather than recurrent data points. 2. Information about all the data instances remain intact. 3. Implementation and Interpretation is quite simple and understandable.

Disadvantages: 1. overlapping of neighboring classes can generate additional noise. Since SMOTE does not realise neighbouring data samples can be from different classes also. 2. Not recommended for high dimensional data, so SMOTE is not practically used for high dimensional data sets.

ADASYN[4] (Adaptive Synthetic) is an oversampling technique in which class weights are distributed for numerous minority data samples, which are being used based on their struggle level in learning. Most of the synthetic samples are generated for classes which are difficult to train as compared to other class samples which can be easily learned. It is an improved version of Smote. After creating those samples, it adds a random weight to the points thus making it more realistic. In other words instead of all the sample being linearly correlated to the parent they have a little more variance in them i.e they are bit scattered.

ADASYN[4] advances learning w.r.t. data distribution in below ways:

(1) Decreasing the bias lead by the imbalanced class.(2) Attentively moving the decision boundary of classification algorithm near the most problematic examples.

Based on previous research work, it has been observed that SMOTE[3] is one of the promising approach which is mostly used for creating new rules in classification problems. This approach can be widely used in the applications of knowledge discovery, data mining, and Telecom domain. But researches can be done in the context of testing this method using bulky databases and to compare this approach with different existing approaches.

III. Proposed Method

The Proposed approach is based on Centroid oversampling technique to generate synthetic samples by using kNN algorithm to identify the nearest neighbors and compute centroid of these nearest data points in the space.

Existing SMOTE[3] technique has a limitation while working with imbalanced dataset,i.e. the classes having frequent data examples overrules neighbors of testing instances despite of distance metrics.

The proposed approach is aimed to resolve this issue by generating centroid between the nearest data points rather than on the outlying areas. In this way, synthetic samples are uniformly distributed among the original data points which reduces the probability of selecting outlier in the data space.

The most important benefit of this proposed approach is its capability of correcting internal bias to majority classes of present kNN algorithm. So, this approach is aimed to produce better data samples than the existing techniques, and hence better classification results can be achieved.

Architecture of the projected system is illustrated in figure I.

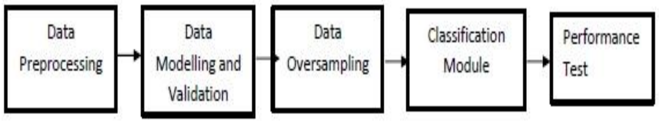

Fig. 1: Architecture Diagram of Proposed Approach

A. Data Pre-processing

The system takes customer churn related dataset as an input. It reads the dataset, and preprocess the categorical or non-numerical attributes and missing values so as to make the data appropriate for further processing.

B. Data Modelling and Validation

Apply 9-fold cross-validation to calculate the learning abilities of classification model on different data sets. A statistical method called Cross-validation is being used to estimate the learning skills of various models. It is mainly applied in machine learning techniques to evaluate a model for predictive learning and modelling issues.

Apply 9- fold cross-validation to build a model on 75% of original data and reserve 25% on testing purpose.

Original file gets divided into testing and training sub-files and so that further task can be applied on each set comprising of test and train dataset.

C. Data Oversampling

Centroid method is used to oversample the dataset. After first iteration of 9-fold cross-validation, oversampling is applied on 75% of original dataset, i.e. testing set. Detailed algorithm is explained here:

- Centroid Oversampling method is based on calculating the median of a triangle or three points.
- Select three nearest neighbor points from the given dataset.
- Calculate the median of these three points.
- The output should be a centroid.
- Continue with the next three neighbors and generate centroid in the same way.

By generating centroids of nearest points, original dataset is oversampled by 100%. The oversampled dataset is then passed to the classification module to train the classification model.

Algorithm 1 Centroid Oversampling Algorithm

```
Input: P={p1,p2....pn} Training positive examples.
n: Number of minority examples.
N: No. of synthetic data points to be generated for each positive data point.
k: 3
Output: Set of artificial instances.
Procedure:
    Preprocess dataset to change the categorical attributes into numerical one and traverse the dataset (each tuple) to preprocess the missing values.
    for i = 1 to min
    Find the two nearest neighbors of pi and store their indexes in an array.
    end for
    for m in indices
    Generate centroid of three points corresponding to their index values.
    newSample=(a[m]+a[m+1]+a[m+2])/3
    m=m+1
    Add new samples one by one to form Synthetic array.
 Synthetic ←newSample
    end for
```

D. Classification Module

1) Load the Testing dataset generated after cross fold validation on which oversampling is required to be done.
2) Use KNN (K-Nearest Neighbor) for classification.

KNN algorithms have been recognized as one of the top most powerful data mining algorithms for their capability of creating simple yet powerful classifiers. The k neighbors nearby a test instances are usually called prototypes.

Salient features of kNN are:

- Most Powerful Data mining algorithm.
- It is exceptionally easy to design and implement.
- Since the algorithm needs no training afore creation of predictions, new data can be further added seamlessly.
- Only two parameters are essential to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)
- Capability of constructing simple but dominant classifiers.
- KNN can be implemented for both classification and regression predictive problems.
- It is lazy learning algorithm and thus, needs no training preceding for constructing real time predictions.

This marks the KNN algorithm quicker than rest of the algorithms, which require training e.g SVM, linear regression, etc.

Classification module takes resultant dataset as an input. Dataset is segregated into two sets: training set and testing set for applying classification. Classifier is trained using training set and using testing set its accuracy is computed.

E. Performance Evaluation Module

There are many parameters available which can be used to check performance of classification. The parameters used in the project to check classification performance are :

Accuracy : Ratio of accurate prediction to the total no of prediction in a model.

Recall : It is considered as a measure of completeness. It tells about how many positive class samples are correctly classified.

Recall : Recall = sensitivity

Precision : It is a measure of exactness. It can be defined as, from the samples labelled as positive, how many actually belongs to positive class.

F-measure : Harmonic mean between recall and preci-sion.

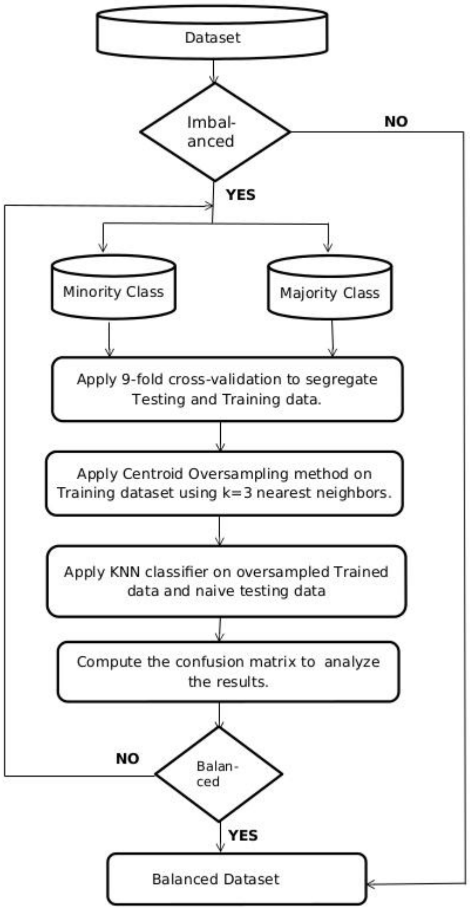
Dataset
Imbal-anced
NO
YES
Minority Class
Majority Class
Apply 9-fold cross-validation to segregate Testing and Training data.
Apply Centroid Oversampling method on Training dataset using k=3 nearest neighbors.
Apply KNN classifier on oversampled Trained data and naive testing data
Compute the confusion matrix to analyze the results.
NO
Balan-ced
YES
Balanced Dataset

Fig. 2: Flow Diagram of Proposed Approach

IV. Testing and Results

Proposed approach is implemented using Python. It is evaluated on three datasets, mainly on telecom churn datasets which are taken from UCI repository and Kaggle.

A comparison is made on proposed model by applying KNN classifier on the original data and data generated by the proposed centroid oversampling technique.

TABLE I: THE COMPARISON OF CLASSIFICATION ACCURACIES FOR CUSTOMER-CHURN DATASET

K=3

Parameters	KNN	SMOTE+KNN	Centroid+KNN
Accuracy	0.92	0.91	0.94
Recall	0.75	0.91	0.95
Precision	0.87	0.84	0.82
F-measure	0.65	0.91	0.89

K=5

Parameters	KNN	SMOTE+KNN	Centroid+KNN
Accuracy	0.90	0.87	0.91
Recall	0.75	0.88	0.90
Precision	0.84	0.79	0.77
F-measure	0.53	0.87	0.83

K=7

Parameters	KNN	SMOTE+KNN	Centroid+KNN
Accuracy	0.9	0.86	0.90
Recall	0.67	0.87	0.88
Precision	0.85	0.78	0.76
F-measure	0.49	0.86	0.80

TABLE III: THE COMPARISON OF CLASSIFICATION ACCURACIES FOR CUSTOMER-CHURN-MODELLING DATASET

K=3

Parameters	KNN	SMOTE+KNN	Centroid+KNN
Accuracy	0.84	0.87	0.87
Recall	0.67	0.85	0.87
Precision	0.70	0.83	0.79
F-measure	0.50	0.89	0.82

TABLE II: THE COMPARISON OF CLASSIFICATION ACCURACIES FOR TELCO-CUSTOMER-CHURN DATASET

K=3

Parameters	KNN	SMOTE+KNN	Centroid+KNN
Accuracy	0.92	0.88	0.94
Recall	0.75	0.84	0.95
Precision	0.87	0.86	0.82
F-measure	0.65	0.91	0.89

K=5

Parameters	KNN	SMOTE+KNN	Centroid+KNN
Accuracy	0.79	0.83	0.81
Recall	0.67	0.77	0.82
Precision	0.69	0.81	0.76
F-measure	0.52	0.88	0.79

K=7

Parameters	KNN	SMOTE+KNN	Centroid+KNN
Accuracy	0.77	0.80	0.78
Recall	0.63	0.74	0.79
Precision	0.63	0.78	0.71
F-measure	0.43	0.86	0.76

V. Statistical Analysis of the Result

From the table I,II and III, it has been observed that the classification accuracy for Centroid oversampling method is best as compared to SMOTE as well as KNN for the value of K=3.

K=5

Parameters	KNN	SMOTE+KNN	Centroid+KNN
Accuracy	0.81	0.81	0.82
Recall	0.59	0.79	0.80
Precision	0.63	0.77	0.73
F-measure	0.32	0.84	0.74

K=7

Parameters	KNN	SMOTE+KNN	Centroid+KNN
Accuracy	0.80	0.79	0.80
Recall	0.55	0.76	0.78
Precision	0.59	0.75	0.69
F-measure	0.21	0.83	0.71

VI. Visualization of the Result

Visualization graph has been plotted on the density distribution of available datasets where the data points of various features are plotted before and after applying the centroid oversampling method on churn.csv dataset.

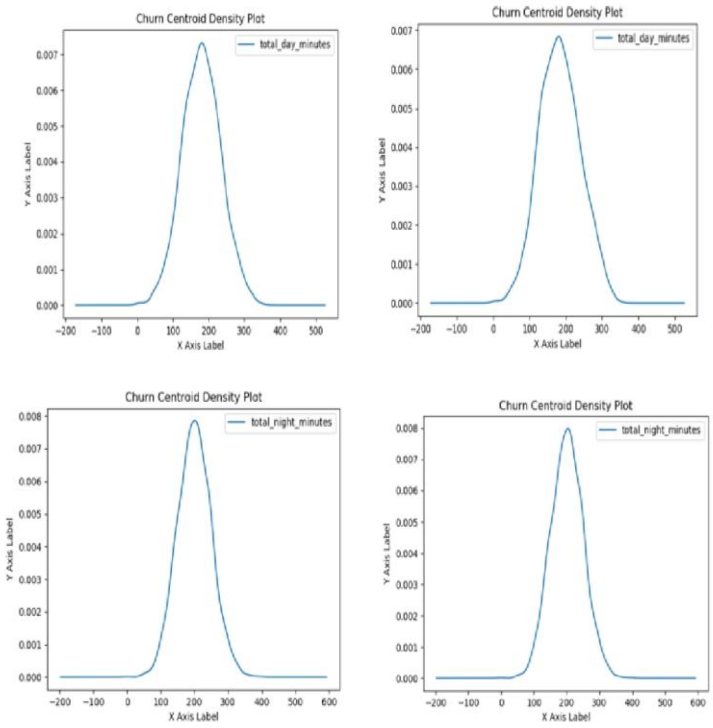

Fig. 3: Density distribution graph of Churn dataset before and after applying centroid oversampling method

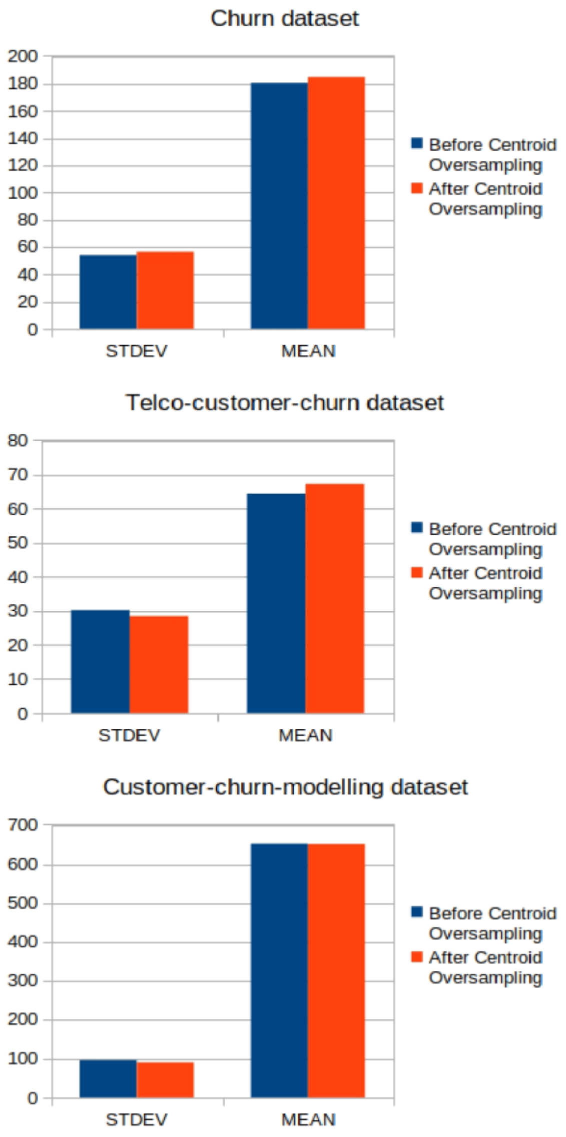

Fig. 4: Data Statics before and after applying Centroid Oversampling Technique

VII. Conclusion

The outcome of the project is focused on finding the customer churn who are going to discontinue the telecom services. Determining customer's nature empower companies to boost support of the customers and boost the gross performance of the industry. Expected result shows the increase in classification accuracy of proposed model as compared to the base model. Once the best oversampling method is identified for a particular dataset, the same can be used to enhance the classifier accuracy. The proposed system shows the better performance for different datasets than the existing oversampling techniques.

VIII. Future Work

Accuracy of many classification algorithms depend on the behavior of the dataset being analysed. For imbalanced datasets, main focus areas are the performance measures and density distribution of data. The proposed approach is based on binary classification of data, so further research can be done with muti-class classification also. Future scope can be in the direction to implement the technique on high dimensional data and to combine the centroid oversampling method with different classifiers to enhance the classification parameters. This method can be efficiently used in the areas of image processing and other data mining approaches.

References

[1] Adnan Amin , Sajid Anwar , Awais Adnan , Muhammad Nawaz , New-ton Howard , Junaid Qadir , (senior Member, Ieee), Ahmad Hawalah4 , And Amir Hussain , (Senior Member, IEEE), ”Comparing Oversam-pling Techniques to Handle the Class Imbalance Problem”,October 26, 2016.

[2] J. Burez and D. Van den Poel, ”Handling class imbalance in customer churn prediction”, Expert Syst. Appl.vol. 36, no. 3, pp. 4626–4636, 2009.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ”SMOTE:Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, no. 1, pp. 321–357, 2002.

[4] H. He, Y. Bai, E. A. Garcia, and S. Li, ”ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. IEEE Int.Joint Conf. Neural Network,Jun. 2008,pp. 1322–1328

[5] Lian Yan,Richard H. Wolniewicz and Robert Dodier, ”Predicting Customer Behavior in Telecommunications”. IEEE Intelligent Systems Volume: 19 , Issue: 2 , Mar-Apr 2004

[6] Jianping Gou, Zhang Yi ,Lan Du, Taisong Xiong ”A Local Mean-Based k-Nearest Centroid Neighbor Classifier” The Computer Journal archive Volume 55 Issue 9, September 2012 Pages 1058-1071

[7] Han, E., Karypis, G.: ”Centroid-based document classification.”In: Zighed, D.A., Komorowski, J., Zytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 116–123. Springer, Heidelberg (2000)

[8] Kiran Dahiya,Surbhi Bhatia ”Customer churn analysis in telecom industry” IEEE-International Conference on Reliability, Infocom Tech-nologies and Optimization (ICRITO) (Trends and Future Directions) 17 December 2015

[9] V.B.Surya Prasatha,, Haneen Arafat Abu Alfeilatb, Omar Lasass-mehb,Ahmad B.A.Hassanat, ”Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier- A Review”, 2017

[10] Liu, Z.G.; Pan, Q.; Dezert, J. ” A new belief-based K-nearest neighbor classification method. Pattern Recognition” 2013, 46, 834–844

[11] Wei Liu and Sanjay Chawla,”Class Confidence Weighted kNN Algo-rithms for Imbalanced Data Sets”, PAKDD 2011 LNCS (LNAI), vol 6635, pp 345-356. Springer, Heidelberg (2011)

[12] Asuncion, A., Newman, D.: ”UCI Machine Learning Repository”

[13] Demsar,ˇ J.: ”Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7”, 1–30 (2006)