Review of Behaviour of Several Methods for Balancing Machine Learning Data

Nirupma Agarwal [1], Ajay Rastogi [2],Ankit kumar Jain [3]
[1.]Student, CCSIT, Teerthanker Mahaveer University,
Moradabad India
[2.] Assistant Prof. CCSIT
Teerthanker Mahaveer University, Moradabad, India
[3]. Student, CCSIT
Teerthanker Mahaveer University,
Moradabad, India

[1],nirupma896@gmail.com
[2.]ajayrastogimbd@gmail.com
[3].Ankittmubca001@gmail.com

Abstract— **Nowadays a huge volume of data is available all over the world. It is very necessary to consider or analyse this huge amount of big data or establish some kind of algorithm to analysis the data. We can analyse our data through machine learning or data mining. Machine learning is an essential part of artificial intelligence it is used to develop an algorithm placed on training data that learn from training data and predict the results on given real world situations or problems. Machine learning used in different types of fields they are detection of virus, in marketing or in playing game and so on. There are many methods or condition that may achieve the performance of leaning systems. In some of the method one method is associated to class inequality such as training data be a part of one class. In this situation real world define an important phase that the learning system may have problem to learn the concept or notion. In this review paper we will experimentally perform some methods form which some of them was planned by the producer to deal with the class inequality problem in thirteen UCI data sets. Experiments present that class inequality does not handle the performance of learning systems. Some cooperative experiments or methods are of two types over sampling method and under sampling method. Over sampling method produce scientific or accurate result or under sampling reflect the area covered by the ROC curve. This Review paper presents the behavioural methods or some algorithms to analyse the training data. The advantage applying machine learning is that once an algorithm prepares what to do with this training data the machine will do the work automatically.**

Keywords— **Machine learning, algorithms**

I. INTRODUCTION

Machine learning mainly use for training data for balancing. This is not real world problem where only one class defined large no of example but all other defined few. This is the class inequality or imbalanced issue in which machine learning algorithms works an obstacle. Mainly the machine learning goal is to analyse the structure of training data and proper that data into models so that people can easily understood and can utilized. Machine learning is different from other traditional popular approaches it is mainly a meadow of computer science. Machine learning algorithm mainly used for analysing the training data and give the output on a definite range. In previous years there are many methods that deal with class inequality problem in data mining and knowledge discovery database in which machine learning is best. Some suggested methods are SMOKE+ Tomes and SMOKE+ENN these methods provide good results. There are mainly two methods that deal with the class inequality problem over sampling method name Smote

[1] and Wilson neighbour rule

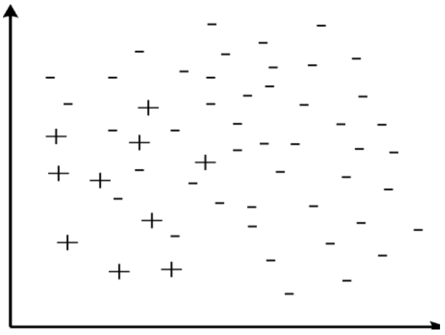
[2] The main motive to use these methods is to balance the training data and also to take down some noisy examples.

[3] and two others cleaning methods are – Tomsk

In this review paper we will describe some experimental and evaluation methods in which three methods are described by the writer in the class inequality problem in thirteen UCI data sets. From this we find that over sampling methods give proper result in the induction of classifier rather than under sampling methods. These two methods achieve well in data sets for some type of example. Machine learning is a progressing field. There are many technologies that benefits from machine learning are Optical character recognition technology (OCR), facial recognition technology, Recommendations engines, self-driving cars these all technology that are mostly using now are benefited from machine learning. We will also discuss type of machine learning supervised learning and unsupervised learning in which clustering and classification play an important role. Machine learning is basically an idea from which we can learn something from examples, experience or experiments rather doing programming. We can write some algorithms that will build some logic on training or generate some methods in place of writing codes.

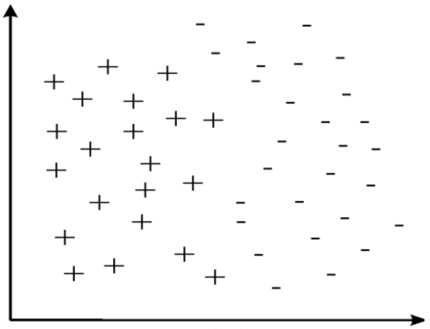
Learning from imbalanced training data sets is difficult –

learning form class imbalanced training data sets is a difficult problem. Class imbalanced problem can be understood by the diagram given below. In this diagram we have well defined training data sets that are not overlapping from one another. We can solve class imbalanced problem from decision tree also. Decision tree also proof a same problem. In class overlying problem Decision tree produce many test to distinguish the majority class from minority class**.**



Imbalance data sets (a)

In this diagram there is a class data imbalanced problem between few class (-) and more class (+) in training data sets and some amount of classes are overlying one and another**.** balance. We can balance our data with some other techniques. We can balance our class training data with well-defined cluster.



Well-defined data sets (b)

In the above diagram (b) majority class and minority class are not overlapping Each other data of training class are balanced in a well –defined clusters. This is positive case different from negative case. In some case minority class may use this method k-nearest neighbour (k-NN).Basically -1NN mainly classify the minority class problem because its neighbour may belong to majority class**.**

Evaluating Imbalanced domains –

We can evaluate the performance of imbalanced training data by confusion matrix. Confusion matrix is also known as error matrix. A confusion matrix is like a table that is used to represent the performance of classifying training data sets and it is also used to easily identify the confusion between classes that are positive and negative classes. It is mainly a brief predications result on coordination problem. Table is given below of positive and negative class.

	CLASS1 Predicted	**CLASS2 Predicted**

CLASS1	True positive (TP)	False Negative(FN)
CLASS 2	False positive (TP)	True Negative(TN)

Error rate and accuracy are the doubtful performance when we study the effect on class on learning system these mainly based on majority class. A 99% (majority class proportion) accuracy can be good or bad it will depend on type of problems.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

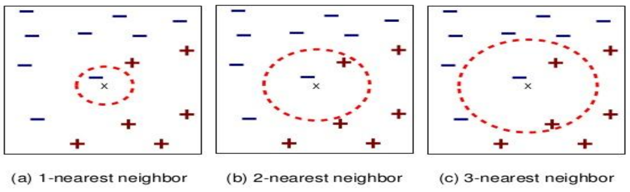
In confusion matrix there are some right or wrong predictions that may use to summarized by each class broken down and count values. Confusion matrix not only give the errors are made but also tell the type of errors that being made. In this matrix mainly the motive of classifier is to maximize the true negative and positive rates and also to minimize the false positive and negative rates. It can be calculate by making a prediction on each row in our test training data sets.

Methods –

In this area we will describe some methods or algorithms like decision tree, k-NN algorithms, and this algorithm is important in performance of methods. From this algorithm we can balance the class problem of our training data.

K-NN Algorithm –

K- Nearest Neighbour this algorithm is mainly used for analysing and regression problems. It is mainly used in industry for classifications problem. K-NN algorithm is a plain and simple algorithm or instance-based learning and it is one of the best data mining algorithms. K-NN is an idle algorithm it means that it will not generalize the training data or we can say that it is minimal. Training data must be kept when there is a loss of generalization. K-nearest neighbours reserve all available cases and analyse new cases placed on similarity measure.


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

In the above diagram of k-Nearest neighbour the majority class (+) and minority class (-) are defined. In 1 –nearest neighbour diagram of k-NN x that have the training data is nearest to minority class but they are away from majority class. In 2- nearest neighbour diagram of k-NN the majority class is not so far from the minority class(-) they are nearest to data sets. In 3-NN diagram the majority class and minority class are nearest to data sets. KNN algorithm use (HVDM) Heterogeneous value different metric for implementation distance function. K-NN is also used to generate the predictions. K-NN algorithm has different names they are –

Instance-based learning_– K-NN mainly refers to a case based learning or instance based learning. It is used to generate predications at raw training instance.

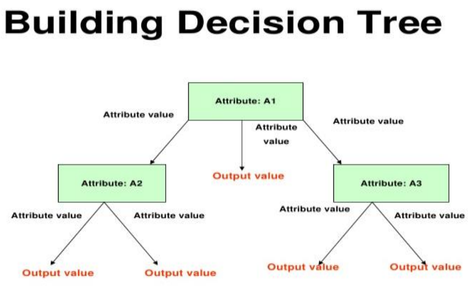
Lazy based learning_– K-NN also known as lazy based learning. In lazy learning all work is done at the time when predictions are needed and no learning model is suitable.

Non-Parametric – K-NN prepare no assumptions that the
KNN know as non-parametric algorithm.

Decision tree –

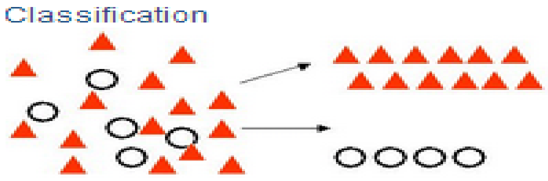
Decision tree is a type of supervised Machine learning in this we can tell what input we are giving and what similar output is of training data sets. Decision tree mainly have two entities that are leaf and the nodes used for decision (Decision nodes).

Decision nodes are used for splitting the data and the leaf are used form final outcomes. There are two types decision tree – classification trees and regression trees.



Decision tree is a tree that has some properties they are first is the Inner node that expresses the properties (attribute), second is the Edge it represents some test on the parent node and the third leaf node that represent the classes. Decision is of two types-

1. Classification trees_– Classification tree is the purpose of variable where the variable is categorized and it is use to describe the class within which a purpose of variable would possible fall into.



2. *Regression trees*_- When the purpose of variable is regular and tree is passed to predict its value.

Smote –

Smote stands for Synthetic Minority Over-sampling Technique [1]. It is part of random over sampling method. In smote technique they perform few operations on real data and generate some extra training data. Its main world is to plan new minority class by including some minority class example that is all lie together. SMOTE main work is to over sampling of the minority class. In smote over fitting complications is hided that matter the decision boundaries for minority class that increase further into majority class.

Neighbourhood Cleaning Rule –

combined with the original class N_i then N_i is rejected. If N_i apply to minority class and its three nearest neighbour discard N_i then all its closet neighbour that belong to majority class are dispatched or removed.

IV .Conclusion

In this review paper we figure out the behavioural of several over under-sampling method to handle the problem of imbalanced training data sets. The result show that these two over sampling behavioural methods SMOTE and SMOTE+ENN balance our data sets and gives a better result as compared to others. These methods are implied on training data sets on positive majority and negative minority class. Random over sampling method also expressed an unprosperous method and provide ambitious result with more confused method and it is also is less costly as compared to other meaningful results.

REFERENCES

[1] Tomek, I. Two Modifications of CNN. IEEE Transactions on Systems Man and Communications SMC-6(1976), 769–772.

[2] Wilson, D. L. Asymptotic Properties of Nearest Neighbour Rules Using Edited Data. IEEE Transactions on Systems, Man, and Communications 2, 3 (1972), 408–421.

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. JAIR 16 (2002), 321–357.

[4] Laurikkala, J. Improving Identification of Difficult Small Classes by Balancing Class Distribution. Tech. Rep. A-2001-2, University of Tampere, 2001.

[5] Weiss, G. M., and Provost, F. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. JAIR 19 (2003), 315–354.