ABSTRACT

An imbalanced dataset is relevant primarily in the context of supervised machine learning involving two or more classes. Imbalance means that the number of data points available for a different class is different, Imbalanced data sets is a special case for classification problem where the class distribution is not uniform among the classes. Imbalanced classes are a common problem in machine learning classification where there is a disproportionate ratio of observations in each class. Class imbalance can be found in many different areas including medical diagnosis, spam filtering, and fraud detection. Typically, they are composed of two classes, the majority (negative) class, and the minority (positive) class. Information about experimental studies sets. These types of data sets are typically found on websites that collect and aggregate data sets. These aggregators tend to have data sets from multiple sources, without much creation. In this case, that‟s a good thing too much creation gives us overly neat data sets that are hard to label. Active learning is undoubtedly effective, but several recent studies have indicated that active learning is failed when it is applied to data. In our project, Human Annotator will collect the data‟s from the public post and he will separate labeled and unlabelled data sets. User needs to register their details and they can view their learning materials. The labelled and unlabelled data‟s have been analyzed by human annotator and matched with labelled appropriate data sets are achieved and it will be available for learners.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABBREVIATIONS	EXPANSION
ANSI	American National Standards Institute
AOW-ELM	Active Online Weighted-Extreme Learning Machine
ASP	Active Server Page
ASR	Automatic Speech Recognition
BCL	Base Class Library
BLOBs	Binary Large Objects
CBIR	Content Based Image Retrieval
CLR	Common Language Runtime
ELM	Extreme Learning Machine
FCL	Framework Class Library
MLP	Multiple Level Perceptron
SQL	Structured Query language
SRS	Software Requirement Specification
SVM	Support Vector Machine
WCF	Windows Communication Foundation

CHAPTER 1

INTRODUCTION

1.1. OUTLINE OF THE PROJECT

Imbalanced data typically refers to classification tasks where the classes are not represented equally. Most of the real-world classification problems display some level of class imbalance, which happens when there are not sufficient instances of the data that correspond to either of the class labels. Therefore, it is imperative to choose the evaluation metric of your model correctly. If it is not done, then you might end up adjusting/optimizing a useless parameter. In a real business-first scenario, this may lead to a complete waste.

There are problems where a class imbalance is not just common it is bound to happen. For example, the datasets that deal with fraudulent and non-fraudulent transactions, it is very likely that the number of fraudulent transactions as compares to the number of non-fraudulent transactions will be very much less. And this is where the problem arises. When the dataset has underrepresented data, the class distribution starts skew.

Due to the inherent complex characteristics of the dataset, learning from such data requires new understandings, new approaches, new principles, and new tools to transform data. And moreover, this cannot anyway guarantee an efficient solution to your business problem.

1.2. PURPOSE OF THE PROJECT

Dealing with imbalanced datasets includes various strategies such as improving classification algorithms or balancing classes in the training data (essentially a data pre-processing step) before providing the data as input to the machine learning. The latter technique is preferred as it has a broader application and adaptation. Moreover, the time taken to enhance is often higher than to generate the required samples. But for research purposes, in our project Human Annotator will

collect the imbalanced data, he will segregate labeled and unlabelled data to provide a complete learning material.

1.3. ACTIVE LEARNING PARADIGM

Active learning is a popular machine learning paradigm and it is frequently deployed in the scenarios when large scale instances are easily collected, but labeling them is expensive and/or time-consuming. By adopting active learning, a classification model can iteratively interact with human experts to only select those most significant instances for labeling and to further promote its performance as quickly as possible. Therefore, the merits of active learning lie in decreasing both the burden of human experts and the complexity of training instances but acquiring a classification model that delivers superior or comparable performance to the model with labeling all instances.

Past research has accumulated a large number of active learning models, and generally, we have several different taxonomies to organize these models. Based on different ways of entering the unlabeled data, active learning can be divided into pool-based and stream-based models. The former previously collects and prepares all unlabeled instances, while the latter can only visit a batch of newly arrived unlabeled data at each specific time point.

According to different numbers of the labeled instances in each round, we have single-mode and batch-mode learning models. As their names indicate, the single-mode model only labels one unlabeled instance on each round, while the batch-mode labels a batch of unlabeled examples once. In addition, we have several different significance measures to rank unlabeled instances, including uncertainty, representativeness, inconsistency, variance, and error. Each significance measure has a criterion for evaluating which instances are the most important for improving the performance of the classification model. For example, uncertainty considers the most important unlabeled instance to be the nearest one to the current classification boundary; representativeness considers the unlabeled instance that can represent a new group of instances, e.g., a cluster, to be more important, and inconsistency

considers the unlabeled instance that has the most predictive divergence among multiple diverse baseline classifiers to be more significant. In addition, active learning models can also be divided into different categories according to which kind of classifier has been adopted.

Some popular classifiers, including naive Bayes, *k*-nearest neighbors, decision tree, multiple level perceptron (MLP), logistic regression, support vector machine (SVM), and extreme learning machine (ELM), have all been developed to satisfy the requirements of active learning. In the past decade, active learning has also been deployed in a variety of real world applications, such as video annotation, image retrieval, text classification, remote sensing, image annotation, speech recognition, network intrusion detection, and bioinformatics.

The proposed algorithm is named active online weighted ELM (AOW-ELM), and it should be applied in the pool-based batch-mode active learning scenario with an uncertainty significance measure and ELM classifier. In AOW-ELM, we first take advantage of the idea of cost-sensitive learning to select the weighted ELM (WELM) as the base learner to address the class imbalance problem existing in the procedure of active learning. Then, we adopt the AL-ELM algorithm presented in our previous paper to construct an active learning framework. Next, we deduce an efficient online learning mode of WELM in theory and design an effective weight update rule. Finally, benefiting from the idea of the margin exhaustion criterion, we present a more flexible and effective early stopping criterion. Moreover, we try to simply discuss why active learning can be disturbed by skewed instance distribution, further investigating the influence of three main distribution factors, including the class imbalance ratio, class overlapping, and small disjunction. Specifically, we suggest adopting the clustering techniques to previously select the initially labeled seed set, and thereby avoid the missed cluster effect and cold start phenomenon as much as possible. Experiments are conducted on 32 binary-class imbalanced data sets, and the results demonstrate that the proposed algorithmic framework is generally more effective and efficient than several state-of-theart active learning algorithms that were specifically designed for the class imbalance scenario.

1.4.1. Machine Learning Methods

Some of the methods of Machine Learning algorithm are categorized as

- SUPERVISED LEARNING

A Supervised learning algorithm learns from labelled training data, helps you to predict outcomes for unforeseen data. It is highly accurate and trustworthy method.

- UNSUPERVISED LEARNING

Unsupervised learning algorithm is the type of self - organized with the help of previously unknown patterns in dataset without pre-existing labels.

- SEMI-SUPERVISED LEARNING

Semi-supervised learning is the combination of both supervised and unsupervised which means labeled and unlabelled data.

- REINFORCEMENT MACHINE LEARNING

Reinforcement machine learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

1.4.2. Applications of Machine Learning

- Video Surveillance
- Social Media Services
- Email Spam and Malware Filtering
- Financial Services
- Health Care
- Retail
- Transportation

1.4.3. Advantages

- Computational property is cheaper and more powerful.
- Affordable data storage.
- It can analysis complex data quickly and automatically.
- It produces more accurate results.