International Journal of Multimedia Data Engineering and Management
Volume 8 • Issue 1 • January-March 2017

Efficient Imbalanced Multimedia Concept Retrieval by Deep Learning on Spark Clusters

Yilin Yan, University of Miami, Department of Electrical and Computer Engineering, Coral Gables, FL, USA
Min Chen, University of Washington Bothell, Bothell, Computing and Software Systems, School of STEM, WA, USA
Saad Sadiq, University of Miami, Department of Electrical and Computer Engineering, Coral Gables, FL
Mei-Ling Shyu, University of Miami, Department of Electrical and Computer Engineering, Coral Gables, FL

ABSTRACT

The classification of imbalanced datasets has recently attracted significant attention due to its implications in several real-world use cases. The classifiers developed on datasets with skewed distributions tend to favor the majority classes and are biased against the minority class. Despite extensive research interests, imbalanced data classification remains a challenge in data mining research, especially for multimedia data. Our attempt to overcome this hurdle is to develop a convolutional neural network (CNN) based deep learning solution integrated with a bootstrapping technique. Considering that convolutional neural networks are very computationally expensive coupled with big training datasets, we propose to extract features from pre-trained convolutional neural network models and feed those features to another full connected neutral network. Spark implementation shows promising performance of our model in handling big datasets with respect to feasibility and scalability.

KEYWORDS

Apache Spark, Classification, Convolutional Neural Network (CNN), Deep Learning, Imbalanced Data, Semantic Indexing

INTRODUCTION

Skewness in data classes poses a significant challenge in major research problems pertaining to data mining and machine learning (Chen & Shyu, 2013; Chen & Shyu, 2011; Lin, Ravitz, Shyu, & Chen, 2007). Classes are rated as skewed or imbalanced when their data instances are non-uniformly associated to the class label. In real world cases, most applications have some degree of skewness inherently present in the data. Such datasets are often grouped into major and minor classes, where major classes have significantly greater numbers of instances associated with them as compared to minor classes. Some prominent imbalanced dataset use cases include fraud detection, network intrusion identification, uncommon disease diagnostics, critical equipment failure, and multimedia concept sensing. A number of famous classification methods are built to utilize the dataset statistics, which ends up being biased towards the majority classes. When identifying the minor classes, these classifiers often perform inaccurately even for very large datasets with considerable numbers of training instances.

Some notable frameworks aiming to solve this challenge are proposed in (Shyu, Haruechaiyasak, & Chen, 2003; Lin, Chen, Shyu, & Chen, 2011; Meng, Liu, Shyu, Yan, & Shu, 2014; Shyu, et al., 2003; Liu, Yan, Shyu, Zhao, & Chen, 2015; Yan, Chen, Shyu, & Chen, 2015). The authors of these

DOI: 10.4018/IJMDEM.2017010101

frameworks, along with others, target this issue from two different perspectives. The first type is algorithm-based approaches where the authors propose new frameworks or improve the existing methods using both supervised and unsupervised techniques. The second, very different type is towards the manipulation of the data itself to reduce the skewness in the class attribution. However, the problem of imbalanced classes is far from being conquered, especially in multimedia data. Multimedia data is particularly difficult because of the various data types that are layered with spatio-temporal features.

One path to handle this challenging situation would be to employ solutions from other domains of machine learning such as deep learning. Deep learning is the name of a whole family of algorithms that use graphs with multiple layers of linear and non-linear transformations to develop hierarchical learning models (Wan et al., 2014). Several frameworks have been proposed using the deep learning techniques that show promising results in application domains such as automatic speech recognition (Swietojanski, Ghoshal, & Renals, 2014), computer vision (Chen, Xiang, Liu, & Pan, 2014), and natural language processing (Mao, Dong, Huang, & Zhan, 2014). However, deep learning methods have not been used to address the problems of class-imbalance. As illustrated in Section IV of our empirical study and also presented in (Sun et al., 2013; Snoekyz et al., 2013) on the TRECVID 2015 datasets, even the famous deep learning methods such as convolutional neural network (CNN) which outperforms a multitude of conventional machine learning techniques face difficulties when dealing with the class-imbalance problems. Moreover, for big datasets in multimedia data mining, deep learning methods are very expensive on computations. The method proposed in (Karpathy et al., 2014) took more than 30 days to train with 1755 videos. The authors were only able to successfully train the deep learning framework using a near-duplicate algorithm.

Toward such demands, our method is proposed to improve the TRECVID dataset confidence scores by a CNN based deep learning framework. In addition, a big data deep learning approach coupled with a bootstrapping sampling technique is proposed to create a balanced set of batches using the training dataset. To the best of our knowledge, bootstrapping has not been used with the deep learning frameworks. To further facilitate the capability of handling the class imbalance problem in big datasets, a distributed computation framework using Apache Spark is also implemented to bind the novel qualities of CNN with the bootstrapping procedures. The proposed framework has shown to be highly impressive and comparatively economical in classifying highly skewed multimedia datasets. The Spark-based distributed computing capability enables a scalable architecture that can mine unstructured key-value confidence scores of multimedia data.

The remaining of the article is organized in the following manner. The following section discusses the related work in skewed data classification methods, followed by some recent progresses in deep learning. Our proposed framework is introduced in the Framework section with its performance evaluated using the experimental results in the next section. The last section concludes the findings and develops the direction for future research.

RELATED WORK ON CLASSIFICATION FOR CLASS-IMBALANCE DATASETS

As mentioned in the previous section, class-imbalanced data classification methods can be sorted into two categories, namely the algorithm techniques and the data manipulation techniques. The first type of approaches, i.e., the algorithm based techniques, either propose to build new or improve existing algorithms to attain superior classification of imbalanced datasets. Consider the cost-sensitive learning methods as an example, where the method maximizes the cost functions of the data to increase the accuracy of class prediction. The intention behind these frameworks is that practical applications do not treat misclassified instances equally. These methods typically evaluate a cost matrix and utilize it in training the model. A relevant approach to cost-sensitive training is to modify the bias to give an advantage to the minority class (Unsworth & Coghill, 2006). Some frameworks using this approach do show the likelihood of improving the classification accuracy, but they are restricted to few application domains.