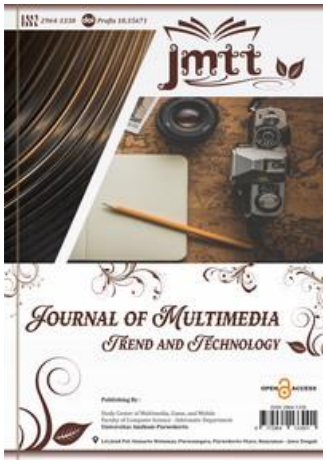
Journal of Multimedia Trend and Technology - JMTT
Vol. 3, No. 1, April 2024, ISSN 2964-1330
https://journal.educollabs.org/index.php/jmtt/

Digging Deeper With Machine Learning for Unbalanced Multimedia Data Categorization

Nihayatusyifa[1], Dita Febrianti[2]

[1,2]Departement of Informatic, Universitas Perwira Purbalingga, Indonesia
Email: nihaya08@gmail.com[1], dita008@gmail.com[2]

ARTICLE INFO



History :

Submit on 20 December 2023
Review on 2 February 2024
Accepted on 5 March 2024

Keyword :

Classification,
Imbalance Data,
Video,
Tracking

ABSTRACT

Since many real-world data sets have skewed class distributions—in which the majority of data instances (examples) belong to one class and considerably fewer instances belong to others—classifying unbalanced data is an important area of research. While the minority instances (fraud in banking operations, abnormal cell in medical data, etc.) in many applications actually represent the concept of interest, a classifier induced from an imbalanced data set is more likely to be biassed towards the majority class and show very poor classification accuracy on the minority class. Unbalanced data classification, particularly for multimedia data, continues to be one of the most difficult issues in data mining and machine learning despite substantial research efforts. In this research, we present an extended deep learning strategy to address this difficulty and get encouraging results in the classification of skewed multimedia data sets. In particular, we examine the combination of advanced empirical research on convolutional neural networks (CNNs), a cutting-edge deep learning technique, and bootstrapping techniques. Given that deep learning techniques, like CNNs, are typically computationally costly, we suggest feeding low-level features to CNNs and demonstrate that this may be done in a way that saves a significant amount of training time while still producing promising results. The experimental findings demonstrate how well our methodology performs in the TRECVID data set when it comes to categorising highly unbalanced data.

Corresponding Author:

Nihayatusyifa ✉
Departement of Informatic, Universitas Perwira Purbalingga, Indonesia
Email : *nihaya08@gmail.com*

INTRODUCTION

The issue of class imbalance has drawn a lot of attention lately in data mining and machine learning research[1]. When data instances (examples) in a set are not nearly evenly distributed among classes/categories, as is frequently the case in real-world data sets, the data set is said to be imbalanced[2]. The class having significantly less data instances in such a data set is referred to as a minority class, whereas the class with more data instances is referred to as a majority class in this context. The majority of classifiers exhibit very low classification accuracy on the minority classes because they are primarily designed to explore data statistics, which may lead to bias towards the majority classes[3][4]. In contrast to the majority class, minority class instances are typically more significant and fascinating in a variety of applications, such as rare disease diagnosis data, fraud detection in banking operations, network intrusion detection, risk management, technical equipment failure prediction, multimedia concept detection[5][6].

Numerous strategies have been put out in the literature to address this problem. Generally speaking, they fall into two categories: data manipulation techniques, which seek to alter the data distribution in order to lessen the imbalance in data sets, and algorithm/model oriented approaches, which seek to suggest new learning mechanisms or adapt current methods to work for imbalanced data sets[7]. Unbalanced data classification is still a difficult research issue, nevertheless. Multimedia data presents much more of a challenge because of its varied media types and spatiotemporal properties[8][9].

There have been significant advancements in machine learning techniques in recent years. "Deep learning" is a significant breakthrough approach that encompasses a set of machine learning algorithms that use deep architectures made up of several non-linear transformations to try and model high-level abstractions in data. Applying deep learning techniques to a range of applications, such as object detection, speech recognition, and natural language processing, among others, has shown promising results in a number of recent research[10].

To the best of our knowledge, however, imbalanced data categorization problems have not been addressed by deep learning techniques. As we have seen in our empirical study (in Section IV) and presented on the TRECVID data (a benchmark data set with severely imbalanced data distributions)[11], deep learning approaches, like convolutional neural networks (CNNs), actually perform worse in imbalanced data, even though they frequently outperform traditional machine learning methods in many applications. In addition, using deep learning techniques to big multimedia data sets may be computationally prohibitively expensive[10][12]. The authors, for instance, stated that it took them more than a month to train the deep learning models on 200 videos, and that they could not have completed their task without using a near-duplicate locating approach to reduce the size of the training set.

In order to enhance multimedia data classification, we present an expanded CNN-based deep learning architecture in this research[13]. CNNs are combined with a bootstrapping sample method in this framework to provide a series of balanced training batches, each with a small number of successful examples. As far as we are aware, deep learning techniques have not benefited from bootstrapping when applied to unbalanced data sets. Furthermore, our suggested bootstrapping sample technique matches the distinct features of CNNs, enabling the expanded deep learning framework to operate efficiently on the experimental data set[14]. It is demonstrated to be very successful and efficient at classifying multimedia material with a distribution of data that is very unbalanced[15][16].

This is how the remainder of the paper is structured. Several approaches to the classification of unbalanced data are addressed in Section 2. In Section 3, the suggested

framework is presented. The experimental findings and analysis are provided in section 4. This paper is finally summarised in the last part.

METHOD

A. Related work for imbalanced data classification.

The fundamental goal of algorithm/model oriented techniques was to improve the performance of unbalanced data classification by analysing and adjusting the training procedures. To enhance classification performance, cost-sensitive learning techniques, for example, aim to maximise the loss functions related to a given data set. The fact that most real-world applications lack consistent consequences for incorrect classifications serves as the inspiration for these learning techniques[11]. Since it is usually uncertain how much each type of error would actually cost, these approaches must calculate the cost matrix from the data and apply it to the learning stage. A related concept for learners who are cost-sensitive is to modify a machine's bias to support the minority class. Despite the fact that certain research has indicated their potential to enhance classification performance on unbalanced, they are far from extensive or systematic[17].

There are various kinds of methods for manipulating data. To counteract the effect of imbalanced data sets, sampling-based techniques such as oversampling and under-sampling have garnered considerable interest. Research has examined various forms of oversampling and under-sampling strategies and given (sometimes opposing) opinions regarding the relative merits of oversampling versus down-sampling for unbalanced data sets. In oversampling, specific algorithms produce similar or duplicate positive data examples in order to balance the data collection. A number of studies offered an enhanced method for oversampling that relied on the synthetic minority over-sampling technique (SMOTE)[15]. Overfitting, however, may result from oversampling. Conversely, down-sampling involves choosing a portion of negative samples (data instances) in order to construct a model that has a comparable amount of positive samples. Because only a portion of the majority class is used, it is incredibly efficient. The primary drawback is that a large number of data instances in the majority class are disregarded, potentially leading to knowledge loss. Two algorithms were suggested by Liu et al. to address this shortcoming. In order to provide the final prediction results, "Easy Ensemble" takes multiple subsets from the majority class, trains a classification model using each of them, and then combines the model's outputs. The models are trained successively using "Balance Cascade." The majority classifies data instances in each phase that the already trained models correctly classify.

B. Recent progress in deep learning.

Due to deep learning's capacity to achieve optimal performance for a variety of tasks, research efforts in a wide range of fields, including signal and information processing, machine learning, artificial intelligence, etc., have been drawn to the field in recent years. Deep learning's basic concept originated in artificial neural network research. As a result, a wide range of deep learning techniques have been researched, such as deep neural networks (DNN), Boltzmann machines (BM), restricted Boltzmann machines (RBM), deep belief networks (DBN), and so on. A more thorough overview of recent deep learning research may be found in. Of these, the convolutional neural network (CNN), a discriminative deep architecture that falls under the DNN category, has demonstrated state-of-the-art performance in a number of computer vision and image processing applications and competitions. Each CNN module is made up of a pooling layer and a convolutional layer. To create a deep model, these modules are frequently stacked one on top of the other. Many weights are shared by the convolutional layer, and the pooling layer lowers the data rate from the layer below by subsampling the convolutional layer's output.

Although CNNs have demonstrated promising results for classification tasks in a variety of applications, its performance on a highly imbalanced data set is yet unknown. Thus, we examine in this study the effectiveness of CNNs for classifying imbalanced data and, more importantly, how to expand them for improved performance. In particular, we suggest expanding CNNs by appropriately combining them with a bootstrapping sampling technique that complements CNNs' special features. In order to improve CNN's performance on multimedia data classification with or without imbalanced data distributions, our proposed bootstrapping sampling method incorporates oversampling with decision fusion, in contrast to the negative bootstrap method in, which combines random sampling and adaptive selection to iteratively find relevant negatives.

C. Framework.

Based on two principles, convolutional neural networks (CNNs) are deep learning models that are modifications of multilayer perceptions intended to require as little pre-processing as possible. First, each hidden unit should only connect to a limited portion of the input units (referred to as feature maps in CNNs) by limiting the connections between the hidden units and the input units. The notion of locally connected networks is also influenced by the biological finding that the receptive fields of neurons in the visual cortex are localised. The fact that natural images are stationary is another concept that can be used to lower the computing cost of images. This indicates that every portion of the image has the same statistics as every other portion. In order to acquire a distinct feature activation value at each place in the image, we can convolve the features that have been learned over small patches that have been randomly sampled from a larger image. We can utilise the features directly or use their aggregated statistics for classification after acquiring them through convolution. When opposed to using all of the extracted features, the aggregated statistics generally have a significantly smaller dimension and can also produce better results (less over-fitting).

Convolutional layer: several feature maps make up a convolutional layer. The process of computing the feature map at the lth layer involves convolving the feature maps from its previous layer using an activation function f that incorporates learnable kernels and additive bias, as stated in Equation (1). The input data is represented by the first layer in this case, which also shows a selection of input maps and the activation function (f), which is typically selected to be the logistic (sigmoid) function.

$$X_j^l = f\left(\sum_{i\in M_j} X_i^{l-1} * K_{ij}^l + b_j^l\right), l \geq 2; \quad (1)$$

$$X_j^l = f\left(\beta_j^l pool(X_j^{l-1}) + b_j^l\right), l \geq 2. \quad (2)$$

represents a pooling process on Pooling Layer (2) that typically computes the aggregated statistics of the input maps, including their mean and maximum values. The layer could be referred to as mean pooling, max pooling, etc., depending on the pooling technique used. Usually, this layer is placed following each convolutional layer. Fully-connected layer: fully-connected layers, which come after a number of convolutional and pooling layers, are responsible for the high-level reasoning in neural networks. All of the neurons in the preceding layer—which could be convolutional, pooling, or fully connected—are connected to every single neuron via a fully connected layer.

RESULT

As previously indicated, deep learning has achieved great success in numerous study areas; yet, very little work has been done to apply it to the categorization of unbalanced (multimedia) data. As demonstrated in Figure 1, where the y-axis displays prediction error rates during the convergence process and the x-axis indicates the number of

iterations, deploying deep learning models directly on a skewed data set typically results in poor classification performance. The prediction error rate decreases gradually using standard CNN processing, as seen in Figure 1. Nevertheless, as Figure 1(b) illustrates, the error rate may significantly fluctuate or even rise when we use CNN on an unbalanced data set. This is due to the fact that most deep learning techniques—including CNNs—split the training set into multiple batches during training. However, because of the skewed distribution of the training set, when splitting an imbalanced data set, some of these batches might only contain negative instances rather than any positive ones. Consequently, these trained models exhibit subpar performance on the testing set.

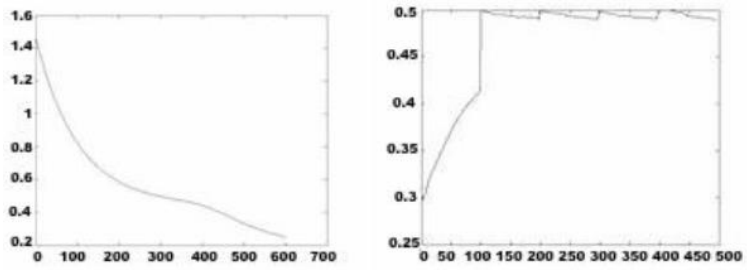


Figure 1, Different types of total error rate convergence generated.

We suggest adding a bootstrapping technique to the CNN algorithm in order to address this problem. In formal terms, bootstrapped sampling has the following definition. Let n and m represent the number of positive and negative examples in the imbalanced training set with n>>m, respectively. According to our suggested framework, Sn and Sp represent the numbers of negative and positive instances, and each batch is formed with the same size S=(Sn+Sp) and negative to positive ratio S=(Sn/Sp). Totally, N batches will be generated, where

$$N = \lfloor n/s_n \rfloor. \quad (3)$$

Put differently, any remaining negative occurrences will not be taken into account during the training process when sn is not precisely divisible by N. The amount of negative examples that are ignored in comparison to those used in training is little because the total number of negative instances (n) in the training set is large and the batch size (sn) is often small. After that, we combine the m positive examples with sn negative instances for each batch, picking one positive instance at random from the m positive instances for sp times. Each training iteration will generate batches via the completion of this process a single time. In order to prevent overfitting, a random approach like this guarantees that every positive instance has an equal chance of being chosen and trained with various negative instances. Table I shows how the procedure works. From the initial imbalanced data, the bootstrapping procedure creates a pseudo-balanced training set for each iteration. After that, we may train the CNN model using it.

Table 1, Setup CNN Model

Layer	Layer size	Output size
Input (m*m)		
Convolution 1	k_1*n_1*n_1	k_1*(m-n_1+1)*(m-n_1+1)
Pooling 1	p_1*p_1	k_1*(m-n_1+1)/p_1*(m-n_1+1)/p_1 [let m_2=(m-n_1+1)/p_1]
Convolution 2	k_2*n_2*n_2	k_2*(m_2-n_2+1)*(m_2-n_2+1)
Pooling 2	p_2*p_2	k_2*(m_2-n_2+1)/p_2*(m_2-n_2+1)/p_2
Output		2

Our suggested multimedia data categorization framework is evaluated on the TREC Video Retrieval Evaluation (TRECVID) 2023 data set, a sizable benchmark set with a significantly unbalanced data distribution, to show how effective it is. A fixed batch size might not be appropriate for every testing idea because the TRECVID 2023 data set has varying degrees of data imbalance for different concepts. As a result, the number of positive training instances in the training set is used to dynamically determine the batch size. We doubled the batch size in our experiment compared to the total number of successful training cases.

Figure 2, Sample keyframes with annotated concepts.

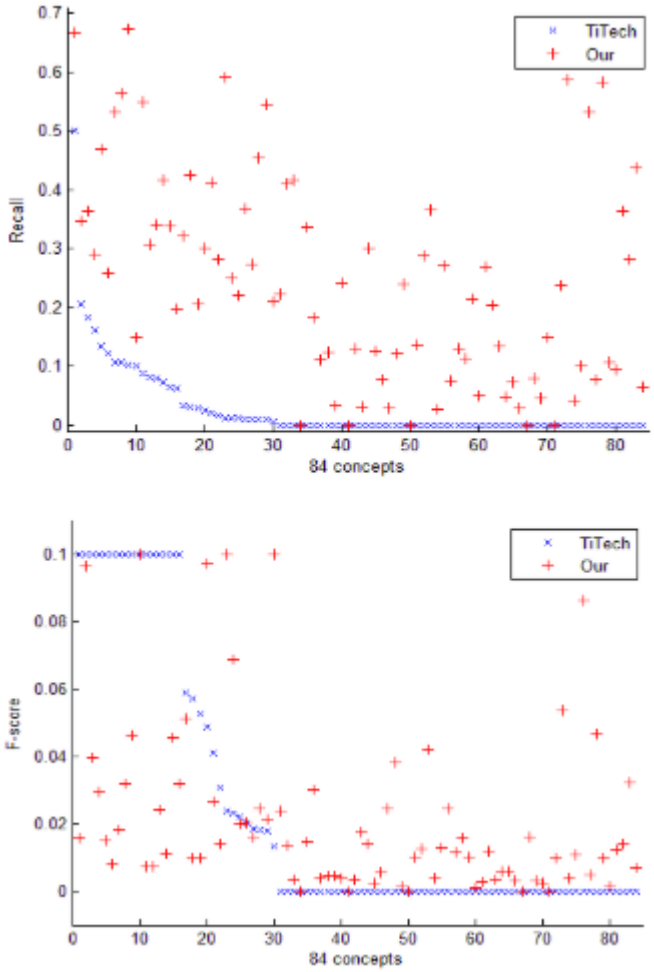


Figure 3, Recall and F-Score Comparsion from 84 Concept

We evaluate our system on 84 highly unbalanced ideas with P/N ratios ranging from 0.0001 to 0.0005. There are 113,046 instances used for testing and 243,355 instances utilized for training. According to [39], the F-score illustrates the trade-off between recall and precision in imbalanced data categorization, when recall is valued more than precision. Therefore, as illustrated in Figure 3, our framework's recall and F-score values are contrasted with those of TiTech (Tokyo Institute of Technology), which placed first in the TRECVID 2023 semantic indexing challenge.

As can be shown, for two thirds of the 84 ideas, our F-scores surpass those of the TiTech group, and for nearly all of the concepts, our recall values are significantly higher. The exception to this is that for four concepts, both frameworks are unable to identify any real positive instance because of incomplete and noisy data annotations. It is also noteworthy that our system achieves an average recall value of roughly 0.3, whereas the TiTech group finds only zero or one genuine positive out of 50 concepts. This amply illustrates how well CNNs operate with the bootstrapping approach in our framework for classifying imbalanced multimedia data, particularly in light of the study's findings that CNN performance is significantly poorer than that of every other classifier employed in the experiment.

CONCLUTIONS

In the paper, we propose to combine a bootstrapping strategy with CNNs, a deep learning approach, to increase its capabilities. A set of pseudo-balanced training batches are created during the bootstrapping phase using the data set's attributes, and they are then fed into the CNN for classification. The experimental results show that our suggested approach is effective in classifying multimedia data with a highly skewed data distribution, using the TRECVID data set. Furthermore, it has been demonstrated that our extended CNN framework can operate well on low-level features, which significantly cuts down on the amount of time needed for deep learning training—a departure from many previous studies in the field that use raw media data as their input.

Acknowledgement

As writers, we would want to express our gratitude to the school administration for their help in providing study materials and for considering our research as our final assignment for our studies at Universitas Perwira Purbalingga. This study is being conducted as a means of contributing knowledge and experience to every creative community wherever they may be.

REFERENCE

[1] L. Wang, T. Liu, G. Wang, K. L. Chan, and Q. Yang, "Video tracking using learned hierarchical features," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1424–1435, 2015.

[2] H. Xue, Y. Liu, D. Cai, and X. He, "Tracking people in RGBD videos using deep learning and motion clues," *Neurocomputing*, vol. 204, pp. 70–76, 2016.

[3] D. Zhang, H. Maei, X. Wang, and Y.-F. Wang, "Deep reinforcement learning for visual object tracking in videos," *arXiv Prepr. arXiv1701.08936*, 2017.

[4] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol.

381, pp. 61–88, 2020.

[5] G. Zheng and Y. Xu, "Efficient face detection and tracking in video sequences based on deep learning," *Inf. Sci. (Ny).*, vol. 568, pp. 265–285, 2021.

[6] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 3943–3968, 2021.

[7] S. Pang, J. J. del Coz, Z. Yu, O. Luaces, and J. D\'\iez, "Deep learning to frame objects for visual target tracking," *Eng. Appl. Artif. Intell.*, vol. 65, pp. 406–420, 2017.

[8] G. Chandan, A. Jain, H. Jain, and others, "Real time object detection and tracking using Deep Learning and OpenCV," in *2018 International Conference on inventive research in computing applications (ICIRCA)*, 2018, pp. 1305–1308.

[9] Y. Yoon *et al.*, "Analyzing basketball movements and pass relationships using realtime object tracking techniques based on deep learning," *IEEE Access*, vol. 7, pp. 56564–56576, 2019.

[10] S. Pang, J. J. Del Coz, Z. Yu, O. Luaces, and J. D\'\iez, "Deep learning and preference learning for object tracking: a combined approach," *Neural Process. Lett.*, vol. 47, pp. 859–876, 2018.

[11] Y. Li *et al.*, "Deep learning-based object tracking in satellite videos: A comprehensive survey with a new dataset," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 181–212, 2022.

[12] H. V. R. Aradhya and others, "Object detection and tracking using deep learning and artificial intelligence for video surveillance applications," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 12, 2019.

[13] P. R. Kamble, A. G. Keskar, and K. M. Bhurchandi, "A deep learning ball tracking system in soccer videos," *Opto-Electronics Rev.*, vol. 27, no. 1, pp. 58–69, 2019.

[14] X.-Q. Zhang, R.-H. Jiang, C.-X. Fan, T.-Y. Tong, T. Wang, and P.-C. Huang, "Advances in deep learning methods for visual tracking: Literature review and fundamentals," *Int. J. Autom. Comput.*, vol. 18, no. 3, pp. 311–333, 2021.

[15] L. Jiao, D. Wang, Y. Bai, P. Chen, and F. Liu, "Deep learning in visual tracking: A review," *IEEE Trans. neural networks Learn. Syst.*, vol. 34, no. 9, pp. 5497–5516, 2021.

[16] S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra, "Deep learning in multi-object detection and tracking: state of the art," *Appl. Intell.*, vol. 51, pp. 6400–6429, 2021.

[17] D. Meimetis, I. Daramouskas, I. Perikos, and I. Hatzilygeroudis, "Real-time multiple object tracking using deep learning methods," *Neural Comput. Appl.*, vol. 35, no. 1, pp. 89–118, 2023.