Classification of images with balanced class for faster retrieval of big data and accuracy using CNN

Hassan Mashraa AlBuqami

Department of Information Science, Faculty of Arts and Humanities, King Abdulaziz University, Saudi Arabia
E-mail: Halbugami2020@gmail.com

Abstract

As huge amounts of structured and unstructured data are available in big data, faster retrieval of the needed information in a faster and more accurate way is essential. The classified data will be helpful to retrieve the big data in an effective way. CNN (Convolutional neural networks) is one such technique to deal with huge volumes of images. CNN's are widely used in images and videos that extract the object's features. To overcome the drawbacks related to class imbalance, a new clustered and centroid-based algorithm is suggested to balance the class. This helps in proper classification and to deal with applications where the minimum class data would play a major role e.g., Data of patients suffering from cancer. The proposed cluster-based classification method will help balance the class and retrieve the most accurate data faster.

2010 Mathematics Subject Classification. **62H30**. 68N01, 65D17.
Keywords. big data, class, classification, CNN, cluster.

1 Introduction

To come up with proper and accurate decisions, a huge volume of data is involved. Big data involves large volumes of data. The data are of different forms structured, semi-structured, and unstructured. It is mandatory to retrieve the needed data in a faster and more accurate way so that it can be helpful for decision-making. As zettabytes of data are involved in recent days, traditional retrieval methods cannot be suitable for handling big data. From a vast variety of sources, big data are gathered so that they can be useful for decision-making. Sources include customer feedback, social media, transactional information, etc. When data is in a fixed form i.e. in a structured form, it is easier to use it when compared to unstructured data. Characteristics of big data include huge volume, different varieties like PDFs, and images, speed of data generation, and variability. This sort of big data helps in taking decisions and improving customer satisfaction. Content-based information retrieval works by using metadata, images, etc. [1]. The proposed method uses CNN for feature extraction in images.

To deal with the retrieval of data, a classification-based method is suggested. Based on the features extracted from the image, the images are classified. Features like color, texture, shapes, semantics, structure, etc. are extracted. Hoping that the classification part can be helpful for faster retrieval of data. Image feature extraction using the SVM technique is suggested in [2]. SVM's are primarily not suitable for large datasets and it is time-consuming to train them. As we are dealing with big data, CNN method can be more supportive when compared to SVM. CNN (Convolutional Neural Network) algorithm for image classification is used in the proposed method. Image classifications are majorly used in healthcare. CNN works by using an image as an input and extracts the features of the image to classify it at its best. The convolutional layer is formed from it

Advanced Studies: Euro-Tbilisi Mathematical Journal 16, supplement issue 2 (2023), pp. 57–65.
DOI: 10.32513/asetmj/1932200823205

Tbilisi Centre for Mathematical Sciences.

Received by the editors: 15 November 2022.
Accepted for publication: 15 January 2023.

and to it, the pooling layer is applied to avoid overfitting of data. If the images are slightly tilted etc. CNN technique may face issues to classify the images accurately. To sort this out the data sets are created in an effective way. And in most classification algorithms, the main drawback is because of imbalanced classes.

On classifications of data, if one of the classes is a majority class and the other is a minor class this might lead to imbalanced data, which can affect the accuracy of the model. The minor class might be ignored resulting in the predictions based on the major class alone. Eg. Patients records, when there is a necessity to classify the data, records without cancer will definitely be high when compared to patients records who are suffering from cancer. When our model is based on Cancer patients, its accuracy can be a question mark. To sort this out balancing the class becomes mandatory.

Techniques like over-sampling on minor classes, under-sampling on major classes, SMOTE (Synthetic Minority Oversampling Technique), ensembling techniques, etc. can be used to balance the class to support accurate classification. Class imbalances are handled using 3 ways of technique. 1. Data-level Methods like oversampling, and under-sampling 2. Algorithm level 3. Hybrid one with Data and Algorithm level. Synthetic samples are created in the SMOTE method, the advantage is that duplicate samples are not formed. The method works on the principle of moving the data point a little bit, this makes sure that the new sample is not the same point and is also a bit similar to the existing point. As the approach works by considering the minority class neighbors alone, it faces certain problems. SMOM method by assigning weights helps to avoid overgeneralization to some extent [3]. Especially when the positive class occurrence is less, the imbalanced class is having a huge impact. In certain scenarios, because of rare occurrences of events, class imbalances occur [4]. SMOTE Technique only considers minority class dataset, this is overcome by using Borderline SMOTE in the proposed technique.

In the proposed method, the clustering-based method is suggested to handle imbalanced classes in a more effective way. If the samples are not properly classified, in applications like medical diagnosis it will have serious impacts [5].

Big data retrieval using Deep Learning Modified Neural Networks (DLMNN) is suggested in [6]. Classification using the c4.5 algorithm is suggested in [7]. Here the under-sampling technique is used.

Feature extraction on big video datasets using deep learning techniques is suggested in [8],[9]. Features like subtitles, speech, and objects are considered in this method. SMOTE technique with normalization and CNN is suggested in [10]. Context-based personalized video retrieval results are suggested in [11].

Cluster-based and SMOTE-based approaches are considered in past research, but all varieties of data are not included in the classification process and prediction-making. Here in the proposed method class balancing is achieved and also from all the clusters equal amount of dataset is considered for the classification process, hence can be used for applications like medical emergencies for proper predictions.

2 Research methodology

It is mandatory to balance the classes to perform classification in a more accurate way. For example, in the case of medical records, patients suffering from cancer will be fewer when compared to the whole set of medical records. When this is the scenario, there might be a possibility for minor classes to be left out while making decisions resulting in improper predictions. A proper class balancing technique is suggested in the proposed method which can be helpful in classifying the class effectively and can ease the process of data retrieval where large volumes of data are involved.

Here CNN is used for feature extraction and classification of images. When compared to the other classification algorithms, the pre-process required is very less in the case of CNN method. CNN technique assigns weights to various objects of the image, to be able o perform the classification process. The image is decoded to RGB colours to perform the activities. Normalization is carried out after the decoding process.

1. Accuracy Accuracy is calculated using the formula,

   (TRUE POSITIVE + TRUE NEGATIVE) / (TRUE POSITIVE + TRUE NEGATIVE +FALSE POSITIVE + FALSE NEGATIVE) [3]. (1)

   The accuracy value obtained is high in spite of lower classification, if there are imbalances in a class. Eg. Cancer patient records 99% of the Majority class and 1% of the Minority class. The results can be dominated by the majority class samples which results in a high accuracy rate too, which is not true. Without balancing the class if we perform classification even with the CNN-based technique, the results will not be accurate.

2. Error Rate The error rate is calculated using the formula, 1 – Accuracy [3]. (2) The error rate value is also dominated by the majority classes resulting in wrong results because of the class imbalances. To specially deal with class imbalances, Metrics like 1. Precision, 2. Recall, 3. Selectivity is used.

   (a) Precision is calculated using the formula TRUE POSITIVE / (TRUE POSITIVE + FALSE POSITIVE) [3] (3)

   (b) Recall is calculated using the formula, TRUE POSITIVE / (TRUE POSITIVE + FALSE NEGATIVE) [3] (4)

   (c) Selectivity is calculated using the formula, TRUE NEGATIVE / (TRUE NEGATIVE + FALSE POSITIVE) [3] (5)

   As certain metrics give false results because of class imbalances, class balancing techniques play a major role in the classification process.

2.1 SMOTE technique algorithm

1. A point is identified from the minority class

2. All its nearest neighbors are identified. Let's say there are 5 neighbors near the point identified from the minority class.

3. Based on the number of samples needed, the nearest neighbors are selected, if needed in a random way.

The problem with the SMOTE technique is that it only considers the minority class and if there are any outliers in the minority classes, it can create issues.

2.2 Borderline SMOTE

The outliers are identified and removed as they add to the noise. Outliers are found by finding the neighbors to the point. If the neighbors belong to the majority class, the point can be removed. The noisy instances are removed from the samples [4]. And border points are added. Border points are one which has neighbors of both majority classes and minority classes. As Borderline considers majority classes too, the drawback of SMOTE technique is overcome. As a variant, Border points that have the shortest path from the minority class to the majority class are the ones that are first considered for sampling the minority class.

2.3 Clustering-based class balancing technique

Based on the accuracy percentage, it's not possible to believe that it is a good classifier. When there are more instances of a particular class the model can predict the majority class for all scenarios and can have a greater accuracy percentage. This accuracy percentage cannot guarantee the correctness of the model. So, it is mandatory to perform class balancing to predict the results in a better way. Random sampling or synthetic samples may not come up with proper results. Here a clustering-based class balancing technique is suggested. This method is not based on duplicate sampling or replica sampling; hence this method performs well when compared to the existing class balancing technique. In the proposed method, the majority class is divided into n number of clusters. Based on the features extracted, similar ones form a group and become a cluster. Let us consider that the majority class is of 75% and the minor class is 25%. An equal number of clusters are formed in the major class and minor class, eg 5 clusters of each class. Till there are instances in the minor class clusters, a record is considered from each cluster to be a sample. As for the majority class, random n samples are considered and as for the minority class, all samples (n) are considered. This method makes sure that all sorts of records are made use of for the classification process as each cluster is unique in its own way. Thereby classification results can be more accurate and an equal number of instances for each class can be used resulting in a balanced class. Here there are no duplicate records, all sorts of the original record are used to their full extent, as no cluster is left out.

Algorithm

- Step 1: Equal number of clusters are formed for both the class
- Step 2: A loop is created, which works till there are instances in the minor clusters
- Step 3: In each iteration, a random majority class cluster object and a minority class An object is added to the sample list.
- Step 4: Equal number of instances are used as samples in both the class resulting in Balanced class.
- Step 5: If there are very few records in a cluster of minority classes i.e., ¡ 5%, the Borderline SMOTE technique is used to oversample the cluster.

  Eg. Majority class with clusters 1,2,3,4,5 Minority class with clusters 1,2,3,4,5

As for the example, there are 25% of minority class samples. Resulting in roughly 5% of samples from each cluster. This would result in 25% of samples from the majority clusters and 25% of samples from the minority clusters. A total of 50% of the total samples is used for the classification purpose, which is a very balanced class scenario resulting in giving equal importance to minority classes so

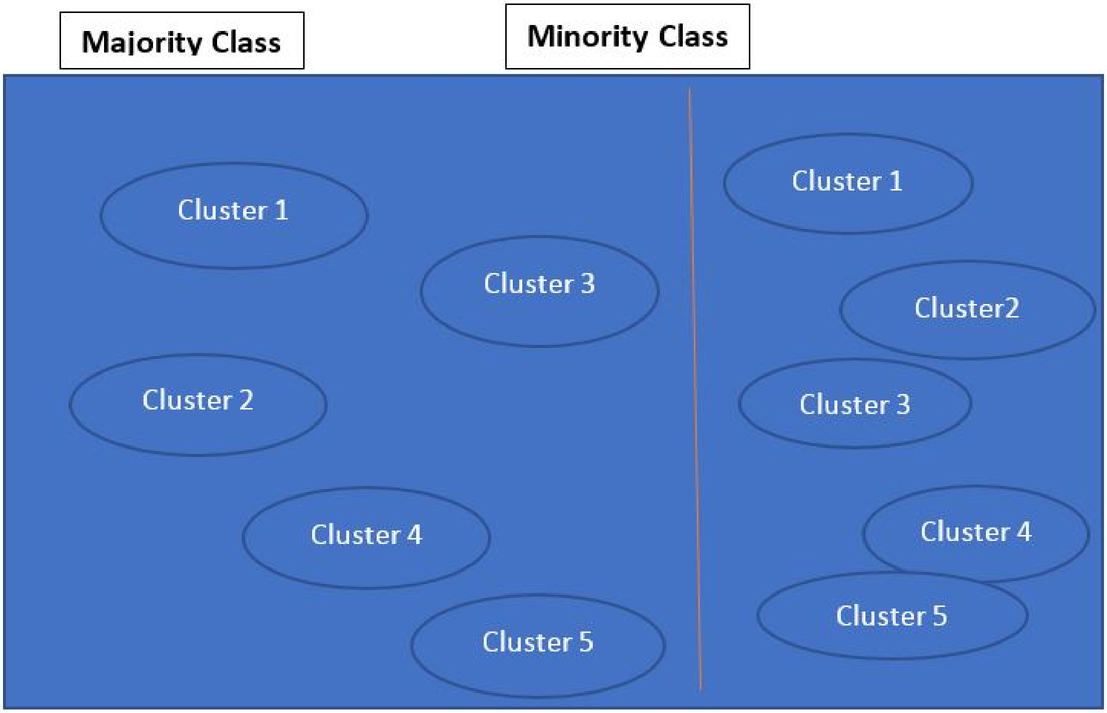

Figure 1. An equal number of clusters are formed for both the majority class and the minority class.

that the minority class is not left out for the prediction process resulting in the proper prediction of cancer patients. The Borderline smote technique also considers majority-class neighbors, thereby overcoming the drawback of SMOTE technique.

A properly classified record set can be useful to fasten the retrieval process. As it's tedious to deal with big data, this balanced classification method can be supportive in the retrieval process, as in Figure 1.

3 Results

This clustering with a Borderline smote-based technique is useful for classifying the class in a more accurate way as the classes are totally balanced, as shown in Figure 2.

Equal samples are used from both classes after using the proposed clustering technique with Borderline SMOTE, as shown in Figure 3 and Table 1.

4 Discussions

As data is growing at a faster rate, big data retrieval is an important aspect. When the data is properly classified, it can be helpful for faster retrieval of Data. Here clustering-based technique helps in the proper classification of imbalanced classes too. As imbalanced classes lead to false

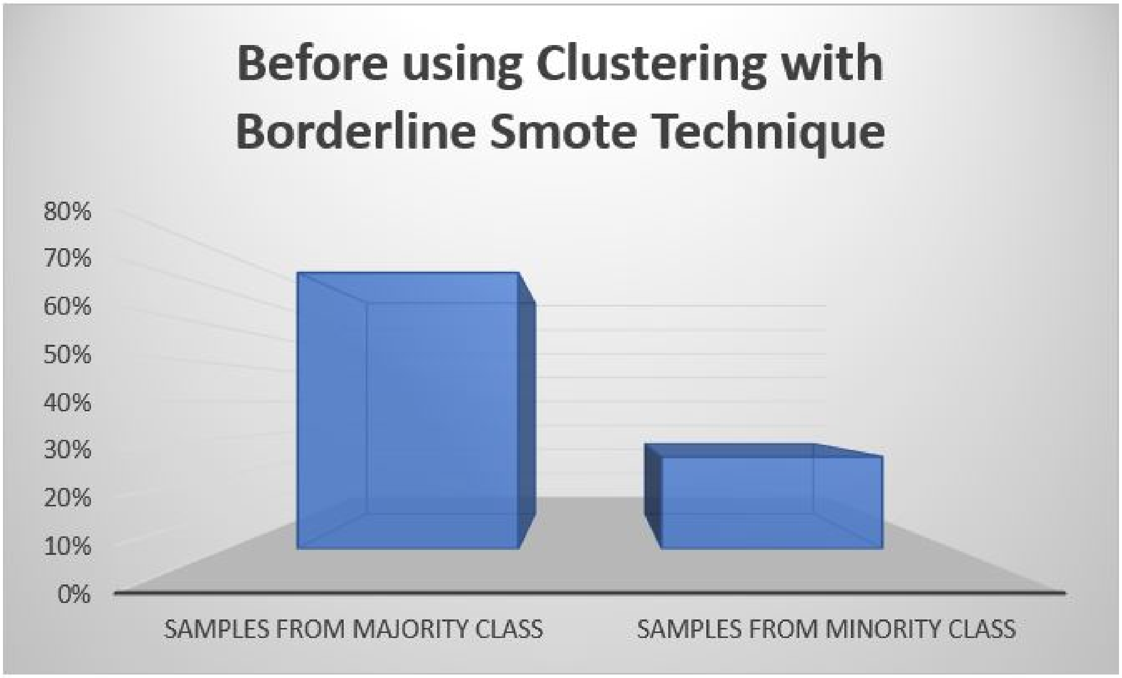

Figure 2. Before using Clustering with Borderline Smote Technique

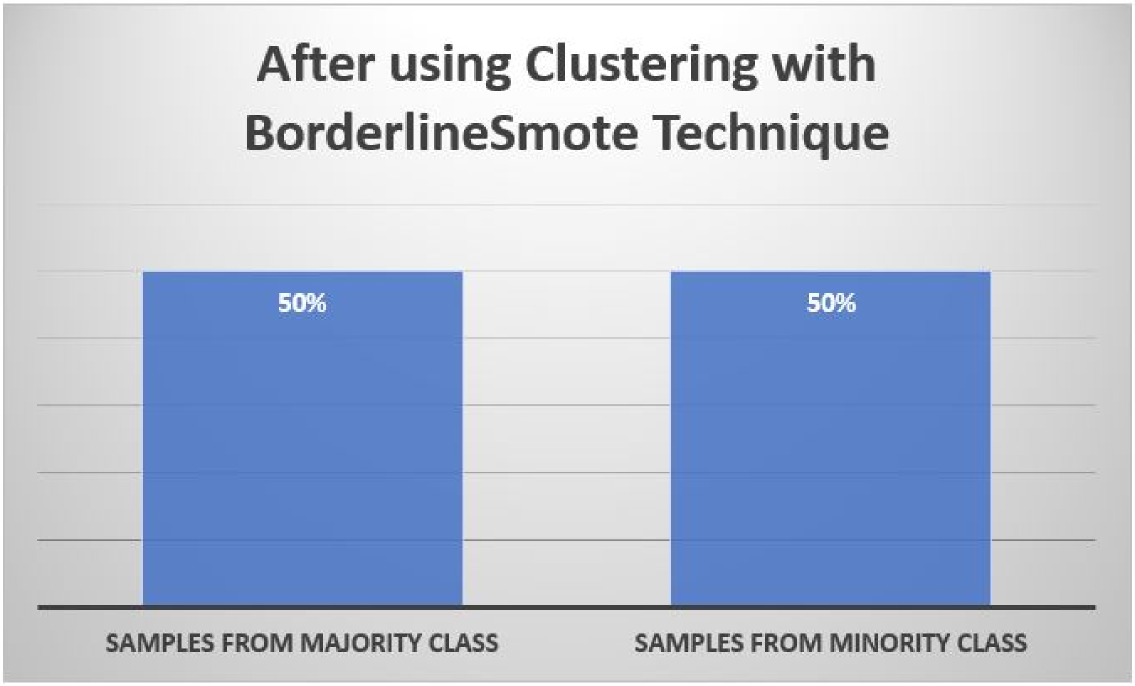

Figure 3. After using Clustering with BorderlineSmote Technique

Table 1. Existing and proposed methods

Existing Method	Proposed Method
Imbalanced Classes	Balanced Classes
Outliers are possible	As the Borderline SMOTE technique is used, outliers are removed.
Possibility for false Accuracy and Error rate value	Classes are balanced using the Clustering based technique, and positive accuracy, and Error rate results can be obtained.
Only the minority class might be considered while oversampling the minority class.	Both the majority class and minority class are considered while oversampling the minority class.
Improper Classifications, thereby unable to support proper retrieval of Big data.	Proper classification can be helpful in the faster retrieval of Big data.

accuracy values, the class is balanced using the Clustering based technique with borderline smote.

Here similar samples are grouped together, and in a random way, the samples are taken from the clusters in an equal proportion. This results in proper classification as no part is left out and every class is treated in equal proportion too. The proposed method can be especially helpful when there are fewer chances for occurrences when compared to the majority classes. E.g. Cancer patients among the medical samples. This might be of the ratio 99% and 1% respectively. If the classification technique is applied as such, the minority class can be left out and the majority class might dominate the entire classification resulting in false accuracy ad error values.

Here drawbacks of SMOTE technique are also addressed and the Borderline SMOTE technique is added where the shortest distance between majority class and minority class border points are also considered as a variant which is used along with the cluster-based algorithm to form a hybrid algorithm that can be useful when Minority class samples are extremely meagre.

This hybrid algorithm makes sure that all types of data are considered for classification and prediction making. And noise cancellation is also enabled making it usable for most important applications too to predict accurately.

5 Conclusions and future works

Balanced classes are essential for the proper classification of data. The proposed method makes sure that the classes are properly balanced resulting in true accuracy value. The cluster-based Borderline SMOTE technique can be applied in areas where minority classes play an important role in predictions. CNN techniques are useful to deal with images. It helps in feature extraction and classification but imbalanced class is the major issue. The proposed method works on balancing the classes so that the minority class is not left out in the prediction process. As huge volumes of data are involved and as the Big data grows each day, proper classification algorithms can be supportive for faster retrieval of big data. To the classified data, keywords-based search methodology can be added to further retrieve the big data at a faster rate.

References

[1] Y. Deng, C. Xing, and L. Cai, *Building Image Feature Extraction Using Data Mining Technology.*, Cognitive-Inspired Semantic Representation and Analytics for Multimedia Data **22** (2022), 8006437.

[2] Z. Tuanfei, L. Yaping, and L. Yonghe, *Synthetic minority oversampling technique for multiclass imbalance problems*, Pattern Recognition **72** (2017), 327-340.

[3] M. Justin, T. Johnson, and M. Khoshgoftaar, *Survey on deep learning with class imbalance*, Journal of Big Data (2019).

[4] L. Kai, R. Bingyu, G. Tao, W. Jiajun, Y. Jiazuo, W. Kexiang, and H. Jicun, *A hybrid cluster-borderline SMOTE method for imbalanced data of rock groutability classification.*, Bulletin of Engineering Geology and the Environment **81** (2022), 1-15.

[5] W. Shujuan, D. Yuntao, S. Jihong, and X. ingxue, *Research on expansion and classification of imbalanced data based on SMOTE algorithm.*, Nature **11(1)** (2021), 1-11.

[6] N. Wahyu, S. Muhammad, and S. Agung, *Clustering Based Undersampling for Handling Class Imbalance in C4.5 Classification Algorithm.*, Journal of Physics: Conference Series **1** (2020), 1-10.

[7] T. Prasanth and M. Gunasekaran, *Effective Big Data Retrieval Using Deep Learning Modified Neural Networks.*, Mobile Networks and Applications **24** (2019), 282-294.

[8] H. Israr, M. Ehsan Ullah, and Q. Ghazia, *Big Data Retrieval: Taxonomy, Techniques and Feature Analysis*, IJCSNS International Journal of Computer Science and Network Security **18(11)** (2018), 5528.

[9] P. Thuong-Cang, P. Anh-Cang, and C. Hung-Phi, *Content-Based Video Big Data Retrieval with Extensive Features and Deep Learning*, Appl. Sci. **12** (2022), 6753.

[10] H. Sadiq, S. Solomon, and H. Javad, *Effective Class-Imbalance learning based on SMOTE and Convolutional Neural Networks*, Machine Learning **85** (2022), 325-336.

[11] F. Yinan, Z. Pan, X. Jie, J. Shouling, and W. Dapeng, *Video Big Data Retrieval Over Media Cloud: A Context-Aware Online Learning Approach.*, IEEE Transactions on Multimedia **21(7)** (2019).