Utilizing Concept Correlations for Effective Imbalanced Data Classification

Yilin Yan[1], Yang Liu[1], Mei-Ling Shyu[1], Min Chen[2]
[1]Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33146, USA
[2]University of Washington Bothell, Bothell, WA 98011, USA
{y.yan4, y.liu39}@umiami.edu, shyu@miami.edu, minchen2@uw.edu

***Abstract*—Data imbalance is a challenging and common problem in data mining and machine learning areas, and has attracted significant research efforts. A data set is considered imbalanced when the data instances (samples) are not close to uniformly distributed across different classes/categories, which is very common in real-world data sets. It is likely to result in biased classification results. In this paper, a two-phase classification framework is proposed to make the classification of imbalanced data more accurate and stable. The proposed framework is based on the correlations generated between concepts. The general idea is to identify negative data instances which have certain positive correlations with data instances in the target concept to facilitate the classification task. The experimental results show that our framework is effective in imbalanced data classification and is robust to feature descriptors by comparing it with four existing approaches using four different kinds of feature representations.**

Keywords: Imbalanced data; classification; skewed data; rare class mining; correlation

I. INTRODUCTION

Recently, the development in computer science has enabled explosive growth and availability of data. Noticeably, in many real-world applications, large amounts of data are generated with skewed distributions (or called data imbalance) since the events of interests occur infrequently [1]-[3]. For example, there are often more samples of normal cells (considered negative class) than the abnormal (positive) ones in cancer research. Another example is in natural language processing (NLP) where positive examples/data instances are vastly outnumbered by negative ones when doing information extraction. Similar situations are observed in other areas, such as fraud detection in banking operations, network intrusion detection, risk management, and failure prediction of technical equipment.

Generally speaking, data imbalance problem is ubiquitous and challenging. In an imbalanced data set, the class that has more data instances is defined as a majority class; while the one with fewer data instances is called a minority class. Since most classifiers are modeled by exploring data statistics, as a result, they may be biased towards the majority classes and hence show very poor classification accuracy on the minority classes. It is also possible that a classifier labels all data instances as the majority classes and ignores the minority classes. There are techniques in the literature that attempt to solve the data imbalance problem [1]-[18]. However, imbalanced data classification remains an open challenge, and more work is needed to tackle it.

In this paper, a two-phase classification framework is proposed that utilizes the concept correlations to effectively tackle the data imbalance problem. Here, concepts refer to high-level semantic objects such as car, road, tree, etc. Using the proposed framework, the majority of the negative data instances that do not have positive correlations to the data instances in the target concept will be identified and filtered so that they will not be used to train the final classification model. This improves the quality of the training data set that leads to a better classifier.

This paper is organized as follows. In section 2, existing classification approaches to handle the data imbalance problem are reviewed in details. Section 3 introduces the proposed framework and each important component. In section 4, experimental results and observations are presented. Finally, the last section summarizes this paper and suggests future research directions.

II. RELATED WORK

In general, imbalanced data classification techniques fall into three categories, namely sampling-based, algorithm-based, and feature selection-based approaches.

A. Sampling-based

The most popular classification algorithms for imbalanced data sets are sampling-based approaches. Oversampling and undersampling methodologies have received significant attentions to counter the effect of imbalanced data sets [1]. Studies have tested different variants of oversampling and undersampling techniques, and presented (sometimes conflicting) viewpoints on the usefulness of oversampling versus downsampling [5] for imbalanced data sets.

In general, downsampling is to select a part of negative samples (data instances) to build a model with a similar number of positive samples. It is very efficient as it uses only a subset of the majority class. The main disadvantage is that many data instances in the majority class are ignored and may result in loss of information. Liu et al. proposed two algorithms to overcome this deficiency [6]. "Easy Ensemble" samples several subsets from the majority class, trains a classification model using each of them, and integrates the outputs of those models to produce the final predication results. "Balance Cascade" trains the models sequentially. In each step, the majority class

data instances that are correctly classified by the current trained models are removed from the next round.

In terms of oversampling, duplicate or similar positive data instances are generated by certain algorithms to make the data set balanced. Zhang et al. presented an improved oversampling approach based on the synthetic minority over-sampling technique (SMOTE) [7][8]. First, data distribution of the minority class is used to estimate whether different types of data instances are overlapped. Next, synthetic data instances are generated in different classes when classes overlap significantly with each other. In addition, weights are increased for those positive samples that are far from the borderline. However, oversampling can potentially lead to overfitting.

B. Algorithm-based

The common goal of algorithm-based approaches is to optimize the performance of learning algorithms on unseen data to address the class/data imbalance problem. One-class learning methods recognize the data instances belonging to a specific class and reject the others. Under certain conditions, such as in a multi-dimensional data set, one-class learning achieves better performance than the peers [9]. Cost-sensitive learning methods try to maximize loss functions associated with a data set to improve the classification performance. These learning methods are motivated by the observation that most real-world applications do not have uniform costs for misclassifications. The actual costs associated with each kind of errors are unknown typically, so these methods need to determine the cost matrix based on the data and apply it to the learning stage. A closely related idea to cost-sensitive learners is shifting the bias of a machine to favor the minority class.

GASEN (Genetic Algorithm based Selective Ensemble Network) has been proven very effective to select a subset of neural networks to form an ensemble classifier or a regressor of the enhanced generation ability. Che et al. tested GASEN on dozens of data sets and finds that there is some potential for improving GASEN's performance on class-imbalance learning [11]. However, such studies on GASEN are far from extensive or systematic. Machine learning algorithms, such as genetic programming (GP), can also generate biased classifiers when data sets are imbalanced. Bhowan et al. used new fitness functions in the GP learning process and empirically showed better performance by the evolved classifiers on both minority and majority classes [12].

C. Feature Selection-based

The goal of feature selection, in general, is to select *n* features from a feature set that allow a classifier to reach an optimal performance, where *n* is a user-defined parameter. As a key step for many machine learning and data mining algorithms especially for high-dimensional data sets, feature selection has been thoroughly studied, where filters are used to score each feature independently based on a rule [13][14]. However, its importance in resolving the data imbalance problem is a recent direction [15]. This direction is motivated by the fact that in real-life data, the data imbalance problem is commonly accompanied with the issue of high data dimensionality which both sampling techniques and algorithm-based approaches may be insufficient to deal with [9]. Therefore, a number of research work has been conducted to perform feature selection to tackle the data imbalance problem recently. For example, Ertekin [16] studied the performance of feature selection metrics in classifying text data drawn from the Yahoo Web hierarchy. They applied nine different metrics and measured the power of the best features using the naive Bayes classifier.

Wasikowski et al. presented the first systematic comparison of different approaches using seven feature selection metrics. They evaluated the performance of these metrics based on the receiver operating characteristic (AUC) and the precision-recall curve (PRC) [9]. Jamali et al. discussed a prior knowledge for an expert system, which can identify the best performed feature selection metric based on the data characteristics regardless of the classifier used [17]. Zheng et al. investigated the usefulness of explicit control of combination within a proposed feature selection framework using multinomial naive Bayes and regularized logistic regression as classifiers [18].

III. FRAMEWORK

Different from most existing approaches, a two-stage framework is proposed in this paper to train the classifiers in order to solve the data imbalance problem. To demonstrate its effectiveness, in our current study, this framework is applied to detect semantic concepts in a large set of video files using keyframes extracted from them. However, it is worth noting that our framework is generally applicable to a wide range of applications that have imbalanced data distributions.

A. Feature selection

In order to use the proposed framework to detect semantic concepts in videos, the first step is to represent the images (i.e., keyframes) using a set of descriptors or features. Such a representation should be able to cover most of the important information contained in the images. Four kinds of features are utilized in this study, namely HOG (Histogram of oriented gradients), HSV (Hue, Saturation, and Value), Gabor, and CEDD (Color and Edge Directivity Descriptor). The first three are widely used features, focusing on gradient, color, and wavelet, respectively. CEDD is a relatively new type of features which incorporates the color histogram and texture information.

HOG [19] is similar to the edge orientation histograms, scale-invariant feature transform descriptors, and shape contexts, but differs in that it is computed on a dense grid of uniformly spaced cells and uses the overlapping local contrast normalization for an improved accuracy [20]. HOG counts the occurrences of gradient orientation in the localized portions of an image to describe the inner visual characteristics of an object. It combines the angles into eight bins which are uniformly divided over 360 degrees, and each bin accumulates the number of edge points whose angles fall in it [21][22]. Let I_x and I_y denote the central differences at point (x, y), $M(x, y)$ be the gradient magnitude, and $\theta(x, y)$ be its orientation. They are defined as follows.

$$I_x = I(x+1, y) - I(x-1, y);$$

$$I_y = I(x, y+1) - I(x, y-1);$$

$$M(x, y) = \sqrt{I_x^2 + I_y^2};$$

$$\theta(x, y) = \tan^{-1} I_x / I_y .$$

The contribution of an edge point to the HOG is weighted by its gradient magnitude $M(x, y)$. In a real implementation, the analyzed object may further be divided to four grids. Then, an ensemble of HOG descriptors with 32 bins can be formed for action analyses (32 = 4 grids ×8 bins).

HSV is utilized for color histogram since it is perceptually uniform that matches with the human vision system. It is a non-linear color model, by which the color signals can be expressed as three kinds of attributes: Hue, Saturation, and Value [23]. Here, Hue refers to the wavelength of the dominant color that ranges from 0 to 360, Saturation represents the purity of color from 0 to 100% (full saturation), and Value is the color brightness from 0 (black) to 100% (white). This color model is expressed by the Munsell three-dimensional coordinate system [24]. After the HSV values are extracted from the images, they are further processed through an appropriate quantization step before they are used for histogram calculation. This can significantly reduce the required computational load and get a uniformed histogram.

Gabor Feature extracted by the Gabor wavelet is widely used in many research areas and proven to be efficient. A 2D Gabor filter is first utilized which is a band-pass spatial filter with selectivity to both orientation and spatial frequency. With the oscillation orientation and frequency, the Gabor feature vector of each key point is extracted based on the principal component analysis (PCA). To determine whether a candidate key point is selected for feature extraction, the point is considered to be the center of a window whose size is dependent on its type [25]. If there is no selected point within the window, the point under consideration will be selected [26]. Based on the selected key points, the Gabor feature vectors are generated. Changing the values of scale and orientation can identify an optimized parameter setting. In this paper, we use 48-dimension uniformed Gabor feature vectors.

CEDD is a new low-level feature that can be used for image retrieval based on multiple feature extraction algorithms. It incorporates both color and texture information to form a 144-demension vector [27]. One of the most important attributes of CEDD is its low computational complexity for feature extraction, in comparison with the needs of the most MPEG-7 descriptors [28]. For the color part, a set of fuzzy rules [29] undertakes the extraction of color information in the HSV color space. The fuzzy system forms a 24-bins histogram with 3 channels of HSV as inputs, and each bin represents a preset color. In order to extract texture features, the MPEG-7 Edge Histogram Descriptor [30] is utilized which can detect edges in vertical, horizontal, 45-degree, 135-degree, and non-directional edges. Adding the 5-edge descriptor with the original information, each region contains 6 fixed texture regions and totally a 144-bin histogram.

B. Classification

SVM is one of the state-of-the-art algorithms in the data mining area [31] including multimedia classification [32][33][34]. The general idea is to build a separating hyperplane to classify the data instances so that the geometric margin is maximized. In order to handle the case that the classes are linearly inseparable, the kernel trick is utilized. In this paper, LibSVM, one of the most popular off-the-shelf software implementations, is used [35]. There are only four common kernels, and γ, r, and d are kernel parameters:

$$\text{linear: } K(x_i, x_j) = x_i^T x_j ;$$

$$\text{polynomial: } K(x_i, x_j) = (\gamma x_i^T x_j + r)^d , \ \gamma > 0 ;$$

$$\text{RBF: } K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} , \gamma > 0 ;$$

$$\text{sigmoid: } K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) .$$

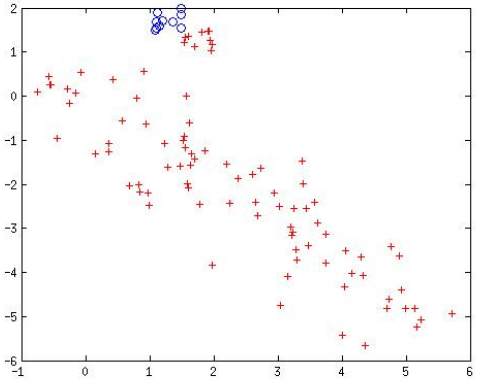

Figure 1. An example of two-dimensional imbalanced data.

C. Data imbalance problem

An example imbalanced data set is given in Figure 1, where the circles represent the positive data instances and the crosses represent the negative ones. Due to the data imbalance problem, the same SVM (with the same kernel and parameters) can produce vastly different borders on different training sets that are randomly picked from the same data set (in Figure 1). We further investigate such an observation by repeatedly picking training sets from the same data set and building SVM models, and notice that the borders can be categorized into three different types: horizontal, left-leaning, and right-leaning (an example for each type is shown in Figures 2(a), 2(b), and 2(c), respectively). This shows that the classification results keep changing for the imbalanced data.

Motivated by the fact that in most data sets, many negative data instances are far from the positive ones, we propose to consider them isolated and build two different models separately. We believe these models, when trained properly, can collectively provide better classification results. The main idea of our two-phase framework is to combine positive instances with their correlated negative instances in the first phase and to identify the real positive ones in the second phase.

D. Positive correlation

In our study, positive correlations are utilized to find the so-called related concepts, namely the negative data instances that are considered to be similar to the positive data instances (i.e., the ones belong to the target concept). Formally, a positive

correlation is defined as a relationship between two concepts that the increase of the occurrence in one concept would increase the other. If two concepts have a positive correlation, it means they are likely to appear together. For example, the concept "Plant" and "Vegetation" are likely to appear together, which means there is a strong positive correlation between them. Therefore, they are both first considered positive.

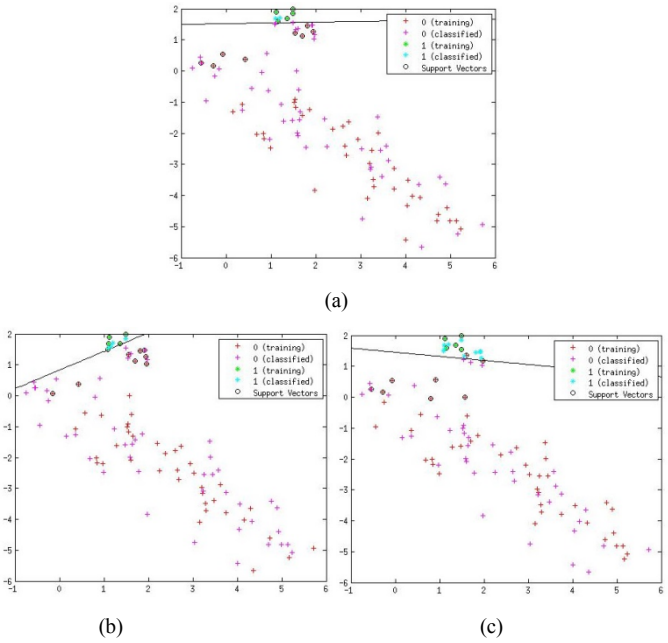

Figure 2. Different types of decision boundaries generated by the same classifier on different training sets which are randomly picked from the same imbalanced data set: (a) horizontal, (b) left-leaning, (c) right leaning.

Many algorithms have been developed to select the most significant positive correlations from the variables. In [36][37][38], the Apriori algorithm and its variants were introduced for the association rule mining technique. Positive correlations have also played an important role in multimedia annotation problems [39][40]. However, though the approaches in mining positive correlations have performed well in previous cases, when coming to large imbalanced data sets, there are some limitations. In order to address this issue, we extend the work in [39] to true positive correlation mining. The idea is that if two concepts (e.g., "Plant" and "Vegetation"), called target concept and reference concept, are truly correlated, the correlation between them would remain with or without the existence of any other concepts (e.g., "Field," "Sky," etc.). In other words, if "Plant" and "Vegetation" have positive correlation only when "Field" is presented but not in any other cases, their correlation may be falsely identified and is actually caused by "Plant"-"Field" and "Vegetation"-"Field" correlations. We therefore define these exogenous concepts as control concepts for further filtering. Formally, assume Ω is the set of interested concepts in the study. For a target concept $T \in \Omega$ (e.g., "Plant"), we want to check whether it has a true positive correlation with a reference concept $R \in \Omega$ (e.g., "Vegetation") by computing the Integrated Correlation Factor (ICF) between them given a control concept C (e.g., "Field"), where $C \in \Omega, C \neq T, C \neq R$.

$$ICF(T,R|C) = \frac{1}{|\Omega| - 2}\rho(C_T, C_R|C_C^+)$$

Here, $|\Omega|$ is the number of concepts in Ω, C_c indicates a control concept, C_c^+ is the set of data instances where concept C exists, C_T and C_R are vectors indicating the existence (1 being yes and 0 otherwise) of target concept and reference concept in the data instances, and $\rho(C_T, C_R|C_c^+)$ indicates the Pearson product-moment correlation coefficient between the target concept and reference concept under the condition of C_c^+. It should be noted that Pearson correlation is undefined under some circumstances, and hence a default value was used in the system in our study [41]. The following reasons further justify the need of introducing the control concept in our system. First, there are many concepts that are normally considered mutually exclusive, such as "indoor" and "outdoor," but are in fact correlate to each other along their borderlines (e.g., front door area is close to indoor and outdoor). Such boundary correlated concepts are hard to detect under normal two-variable systems. However, in our proposed framework, these concepts can be effectively mined because any control concepts that appear normally in one concept but rarely in the other concept would filter the data set and make the correlation more explicit. Second, based on the TRECVID data set [42], all of the newly reconstructed data instances are under the condition that the control concept is labeled as positive, which means our data instances are all viewed by human annotators. This increases the credibility of our data set.

E. Model built based on positive correlations

Based on the trained positive correlations, we can build a two-phase framework for the imbalanced data set. First, those data instances that have a positive correlation with the data instances of the target concept are used as the positive training set and the rest of the negative ones are the negative training set. These two training sets are used to build the SVM model at the first phase. Please note that those data instances having a positive correlation with the target concept may be true positive or false positive. Therefore, another SVM model is built in the second phase to recognize the true positive data instances. If a testing data instance is labeled as negative in phase 1, it does not need to be tested by phase 2. A positive decision is made only when both two phases make the positive decision. Figure 3 and Figure 4 present the training and testing phases of our proposed framework.

To demonstrate the effectiveness of the proposed 2-phase framework, the same example two-dimensional imbalanced data instances (in Figure 1) are classified by the same SVM model, and are used to show how the decision boundaries behave in phase 1, phase 2, and the proposed 2-phase framework. In Figure 5, it can be easily seen that the decision boundaries generated from different tests behave similarly at Phase 1. The same happens at Phase 2 as shown in Figure 6. In addition, when both phases are integrated, the decision boundaries can

better separate the truly positive data instances from the negative ones, which can be clearly observed in Figure 7. The results from these figures show that by using the concept correlations in two phases, it is much more effective to find the stable decision boundaries to separate the positive data instances (the minority class) from a large number of negative data instances in an imbalanced data set, resulting in a high and stable classification accuracy.

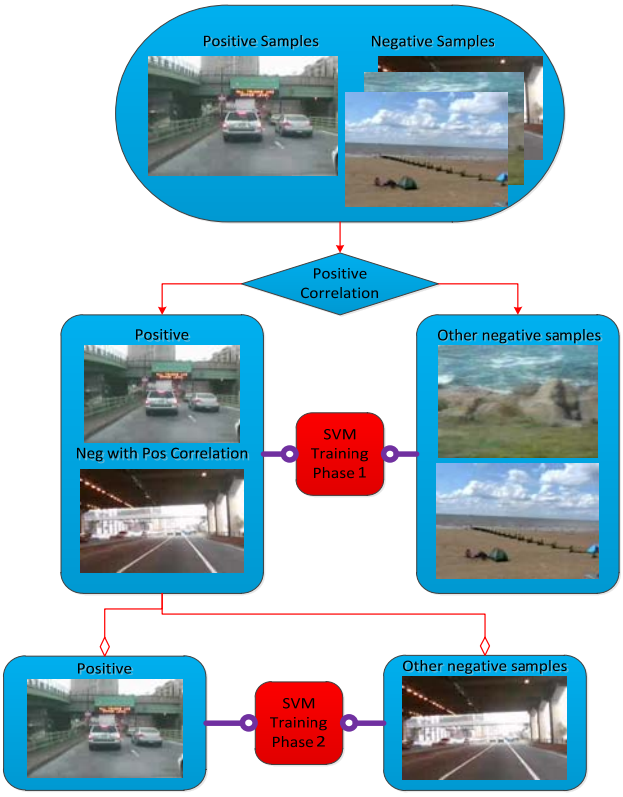

Figure 3. Training stage of the proposed framework.

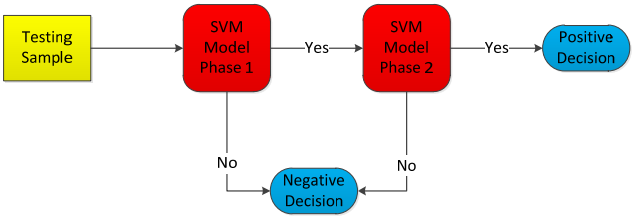

Figure 4. Testing stage of the proposed framework.

F. *Performance Measure*

In general, a classifier is evaluated by a confusion matrix as illustrated in Table I. The columns are the predicted class and the rows are the state of nature (actual class). In the confusion matrix, TN is the number of negative examples correctly classified (True Negatives), FP is the number of negative examples incorrectly classified as positive (False Positives), FN is the number of positive examples incorrectly classified as negative (False Negatives), and TP is the number of positive examples correctly classified (True Positives). For performance comparison, the precision and recall metrics [43] are commonly used and are derived from the confusion matrix as follows.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

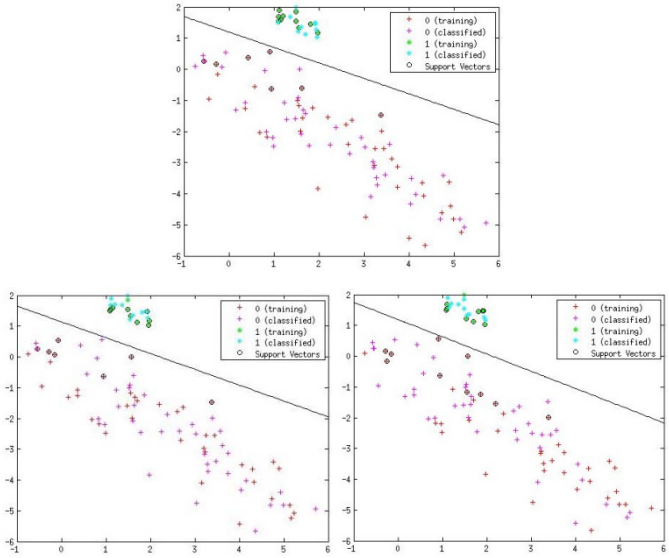

Figure 5. Decision boundaries generated in phase 1.

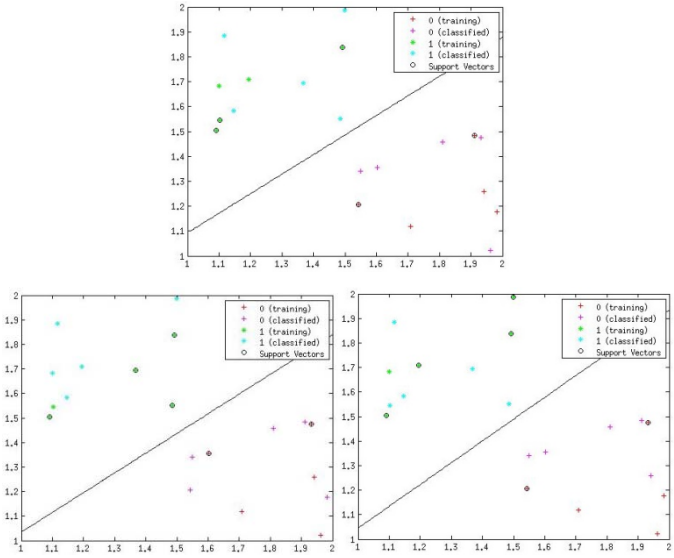

Figure 6. Decision boundaries generated in phase 2.

TABLE I. CONFUSION MATRIX

	Predicted Positive	**Predicted Negative**
State of nature Positive	True Positives (TP)	False Negatives (FN)
State of nature Negative	False Positives (FP)	True Negatives (TN)

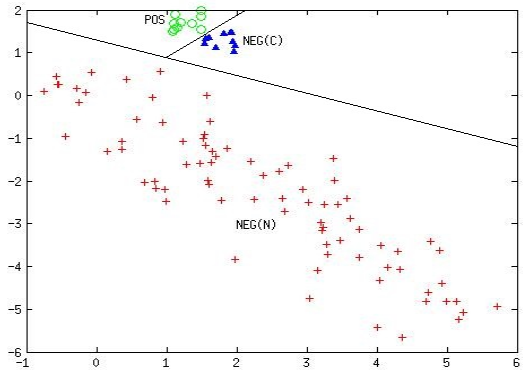

Figure 7. Decision boundaries generated in our proposed 2-phase framework (integrating phase 1 and phase 2), where NEG(C) represents the negative data instances having a false positive correlations with the data instances of the target concept.

The main goal for learning from the imbalanced data sets is to improve the recall without sacrificing the precision. However, recall and precision goals can often be conflicting, since the increase of true positive data instances for the minority class may also increase the number of false positives, which will reduce the precision. F-measure, also known as F_1 score or F-value, can combine the trade-offs between precision and recall, and is therefore considered an objective and ultimate quality metric of a classifier. It is defined as follows.

$$F-measure = 2 \times \frac{precision \times recall}{precision + recall}$$

Figure 8. Sample keyframes with annotated concepts in the TRECVID data set, the concepts are bicycling, tree, politics, and face, respectively.

IV. EXPERIMENTAL RESULTS

In TRECVID 2012 project, the semantic indexing (SIN) task aims to recognize the semantic concept contained within a video shot, which can be an essential technology for retrieval, categorization, and other video exploitations. It has several challenges such as data imbalance, scalability, and semantic gap [44][45]. The research directions to address these challenges may include developing robust learning approaches that adapt to the increasing size and the diversity of the videos, fusing information from other sources such as audio and text, and detecting the low-level and mid-level features that have high discrimination abilities.

A. Experimental Setup

In the experiment, the IACC.1.A data set is chosen from the TRECVID 2012 benchmark [42]. It contains approximately 8000 Internet Archive videos (50GB, 200 hours) with creative commons licenses in MPEG-4/H.264 with the duration between 10 seconds and 3.5 minutes. Most videos have some metadata provided by the donor available, e.g., title, keywords, and descriptions. These videos are collected from the Internet and are diversified in terms of the creator, content, style, production qualities, and original collection devices. The videos are segmented into a number of shots and each shot is represented by a keyframe. The shot boundaries and keyframes are also given in the data set. The labels are provided by collaborative annotation effort organized by NIST. In this study, each keyframe is treated as a data instance. Figure 8 shows four sample keyframes with the labeled concepts. SVM is used as the classifier in this study. We adopt the empirical setting in LibSVM, and for comparison purposes, the cross validation scheme is employed to compare with some existing approaches. In each iteration, two third of the dataset is used to train a model, the rest is for testing.

B. Experimental Results and Analyses

This data set is chosen because it has severe data imbalance problem as some basic statistics shown in Table II [42]. As can be seen, the positive data instances account for less than 1% in the data set, which poses great challenges in data retrieval in practice.

TABLE II. INFORMATION OF IACC.1.A DATA SET

Data Set	IACC.1.A
TRECVID Year	2012
No. of Concepts	130
No. of Instances	144774
Average Positive No.	865.42
Average P/N Ratio	0.0062

TABLE III. TOP TEN POSITIVE RELATED CONCEPT PAIRS

Target Concept	**Related Concept**	**ICF**
Plant	Vegetation	0.69306
Car	Ground_Vehicles	0.56938
Government-Leader	Politicians	0.54638
Daytime_Outdoor	Outdoor	0.41878
Road	Streets	0.37638
Anchorperson	News_Studio	0.34946
Beach	Waterscape_Waterfront	0.34690
Trees	Vegetation	0.34369
Anchorperson	Reporters	0.33728
Building	Suburban	0.31346

In this experiment, the classification of each class is considered as a two-class problem, namely positive and negative. The

target concept is considered as positive; while the rest 129 concepts are considered as negative. As shown in Table III, we first compute the ICFs among the concepts and identify the top ten positive related concept pairs. The column "Target Concept" contains ten concepts that we aim to classify one by one in the experiment while the "Related Concept" helps in Phase 1 of the modeling. The performance of the proposed framework is compared to four existing approaches (down-sampling, over-sampling, adapting SVM [46] and feature selection) by reporting the average performance across these ten concepts. To eliminate the possible influence of features on the classification performance, four different types of features (HOG, HSV, Gabor, and CEDD) are used in the comparisons. Their results are presented in Tables IV, V, VI, and VII, respectively. In the tables, the top performance in each column is highlighted. As we can see, the performance of the classification approaches is largely influenced by the choice of feature set. For instance, most of the approaches including the proposed framework perform better by using HOG or CEDD features (Tables IV and VII) than using HSV or Gabor (Tables V and VI). Nevertheless, though other classification approaches may achieve higher precision or recall values using certain features, , our proposed framework consistently outperforms all of them using the F-measure metric, which as discussed earlier better reflects the effectiveness of a classifier than other metrics.

TABLE IV. EXPERIMENTAL RESULTS BY HOG

	Precision	Recall	F-measure
Down-sampling	5.90%	82.99%	0.1103
Over-sampling	**100%**	36.93%	0.5394
Adapting SVM	52.77%	90.87%	0.6677
Feature Selection	40.67%	95.02%	0.5697
Proposed Framework	58.35%	**98.34%**	**0.7324**

TABLE V. EXPERIMENTAL RESULTS BY HSV

	Precision	Recall	F-measure
Down-sampling	4.42%	**85.06%**	0.0841
Over-sampling	**100%**	0.41%	0.0083
Adapting SVM	22.14%	70.54%	0.3370
Feature Selection	23.28%	68.88%	0.3480
Proposed Framework	28.86%	81.13%	**0.4257**

TABLE VI. EXPERIMENTAL RESULTS BY GABOR

	Precision	Recall	F-measure
Down-sampling	5.72%	**87.55%**	0.1073
Over-sampling	**100%**	1.24%	0.0246
Adapting SVM	33.46%	74.69%	0.4621
Feature Selection	32.10%	68.46%	0.4371
Proposed Framework	39.42%	81.46%	**0.5313**

TABLE VII. EXPERIMENTAL RESULTS BY CEDD

	Precision	Recall	F-measure
Down-sampling	5.29%	84.65%	0.0997
Over-sampling	**100%**	18.26%	0.3088
Adapting SVM	67.14%	97.51%	0.7953
Feature Selection	55.42%	97.51%	0.7068
Proposed Framework	70.85%	**99.01%**	**0.8260**

V. CONCLUSIONS AND FUTURE WORK

It is challenging to obtain reasonable classification accuracies when the data set is imbalanced, since the data instances in the majority class usually overshadows those in the minority class (the target concept). In the paper, we propose a novel 2-phase classification framework that utilizes concept correlations to tackle the data imbalance problem. Experimental results that compare four existing approaches with four types of features demonstrate that the proposed framework is capable of utilizing the concept correlations to separate the positive data instances of the minority class from a large number of negative data instances in an imbalanced data set. It is important to note that the proposed framework can also be applied in many other research areas and applications that are challenged by the data imbalance problem like homeland security, network security, disaster information management, to name a few.

In this paper, we used the correlation between concepts regardless of how strong the correlations are. In the future, the correlation values can also be considered in the framework since the larger a correlation value between two concepts is, the stronger the concepts may be correlated.

REFERENCES

[1] C. Chen and M.-L. Shyu, "Integration of Semantics Information and Clustering in Binary-class Classification for Handling Imbalanced Multimedia Data," Edited by Tansel Ozyer, Keivan Kianmehr, Mehmet Tan, and Jia Zeng, *Information Reuse and Integration in Academia and Industry*, Chapter 14, Springer Verlag, 2013.

[2] C. Chen and M.-L. Shyu, "Clustering-based Binary-class Classification for Imbalanced Data Sets," *The 12th IEEE International Conference on Information Reuse and Integration (IRI 2011)*, pp. 384-389, Las Vegas, Nevada, USA, August 2011.

[3] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Effective Feature Space Reduction with Imbalanced Data for Semantic Concept Detection," *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC2008)*, pp. 262-269, Taichung, Taiwan, June 2008.

[4] T. Jo, and N. Japkowicz, "Class Imbalances versus Small Disjuncts," *SIGKDD Explorations*, 6(1), pp. 40-49, June 2004.

[5] G. E. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explorations*, 6(1), pp. 20-29, June 2004.

[6] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(2), pp.539-550, April 2009.

[7] L. Zhang and W. Wang, "A Re-sampling Method for Class Imbalance Learning with Credit Data," *2011 International Conference on Information Technology, Computer Engineering and Management Sciences (ICM)*, pp. 393-397, September 2011.

[8] N. V. Chawla and K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Oversampling Technique," *Journal of Artijcial Intelligence Research*, 16, pp. 321-357, 2002.

[9] M. Wasikowski and X-W. Chen, "Combating the Small Sample Class Imbalance Problem Using Feature Selection," *IEEE Transactions on Knowledge and Data Engineering*, 22(10), pp. 1388-1400, October 2010.

[10] G. Deng and Y. Jiang, "I-fuzzy equivalence relation and I-transitive approximations," *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pp. 80-84, May 2012.

[11] J. Che, Q. Wu, and H. Dong, "An Empirical Study on Ensemble Selection for Class-imbalance Data Sets," *2010 5th International Conference on Computer Science and Education*, pp. 477-480, August 2010.

[12] U. Bhowan, M. Johnston, and M. Zhang, "Developing New Fitness Functions in Genetic Programming for Classification With Unbalanced Data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(2), pp. 406-421, April 2012.

[13] R. Longadge, S. S. Dongre, and L. Malik, "Class Imbalance Problem in Data Mining: Review," *International Journal of Computer Science and Network*, 2(1), February 2013.

[14] Q. Zhu, L. Lin, M.-L. Shyu, S.-C. Chen, "Feature Selection Using Correlation and Reliability Based Scoring Metric for Video Semantic Detection," *2010 IEEE Fourth International Conference on Semantic Computing*, pp. 462-469, September 2010.

[15] S. Hukerikar, A. Tumma, A. Nikam, and V. Attar, "SkewBoost: An Algorithm for Classifying Imbalanced Datasets," *2011 2nd International Conference on Computer & Communication Technology*, pp. 46-52, 2011.

[16] S. Ertekin1, J. Huang, L. Bottou, and C. L. Giles, "Active Learning in Imbalanced Data Classification," *The Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pp. 127-136, 2007.

[17] I. Jamali, M. Bazmara, and S. Jafari, "Feature Selection in Imbalance Data Sets," *International Journal of Computer Science Issues*, 9(3), May 2012.

[18] Z. Zheng, X. Wu, and R. Srihari, "Feature Selection for Text Categorization on Imbalanced Data," *SIGKDD Explorations*, 6(1), pp. 80-89, 2004.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Conference on Computer Vision and Pattern Recognition*, 1, pp. 886-893, June 2005.

[20] L.-C. Chen, J.-W. Hsieh, Y. Yan, and D.-Y. Chen, "Vehicle Make and Model Recognition using Sparse Representation and Symmetrical SURFs," *2013 16th International IEEE Conference on Intelligent Transportation Systems*, pp. 1143-1148, Hague, Netherlands, October 2013.

[21] L.-C. Chen, J.-W. Hsieh, Y. Yan, and B.-Y. Wong, "Real-time Vehicle Make and Model Recognition from Roads," *2013 12th Conference on Information Technology and Applications in Outlying Islands*, pp. 1033-1040, Kinmen, Taiwan, May 2013.

[22] K.-T. Chuang, J.-W. Hsieh, and Y. Yan, "Modeling and Recognizing Action Contexts in Persons Using Sparse Representation," *2012 International Computer Symposium*, 21, pp. 531-541, Hualien, Taiwan, December 2012.

[23] H.-Y. Yang, J.-F. Wu, Y.-J. Yu, and X.-Y. Wang, "Content Based Image Retrieval Using Color Edge Histogram in HSV Color Space," *Journal of Image and Graphics*, 13(10), pp. 2035-2038, December, 2008.

[24] L. Lei, X. Wang, B. Yang, and J. Peng, "Image dimensionality reduction based on the HSV feature," *2010 9th IEEE International Conference on Cognitive Informatics (ICCI)*, pp. 127-131, July 2010.

[25] D.-H. Liu, K.-M. Lam, and L.-S. Shen, "Optimal sampling of Gabor features for face recognition," *Pattern Recognition Letters*, 25(2), pp. 267-276, January 2003.

[26] G. Du, L. Gong, and F. Su, "An effective Gabor-feature selection method for face recognition," *IEEE International Conference on Network Infrastructure and Digital Content*, pp. 722-725, November 2009.

[27] S. A. Chatzichristofis and Y. S. Boutalis, "CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval," *In Proceedings of the 6th International Conference on Computer Vision Systems*, pp. 312-322, May 2008.

[28] C. S. Won, D. K. Park, and S.-J. Park, "Efficient Use of MPEG-7 Edge Histogram Descriptor," *ETRI Journal*, 24(1), Febuary 2002.

[29] S. A. Chatzichristofis and Y. S. Boutalis, "FCTH: Fuzzy color and texture histogram — A low level feature for accurate image retrieval," *9th International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 191-196, Klagenfurt, Austria, May 2008.

[30] Y. Zhou, L. Li, T. Zhao, and H. Zhang, "Region-based high-level semantics extraction with CEDD," *2010 2nd IEEE International Conference on Network Infrastructure and Digital Content*, pp. 404-408, September 2010.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, 20(3), pp. 273–297, September 1995.

[32] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," *In Proceedings of the 17th IEEE International Conference on Pattern Recognition*, 3, pp. 32–36, August 2004.

[33] D. Liu, M.-L. Shyu, and G. Zhao, "Spatial-temporal motion information integration for action detection and recognition in non-static background," *2013 IEEE 14th International Conference on Information Reuse and Integration*, pp. 626-633, August 2013.

[34] D. Liu, M.-L. Shyu, Q. Zhu, and S.-C. Chen, "Moving Object Detection under Object Occlusion Situations in Video Sequences," *2011 IEEE International Symposium on Multimedia*, pp. 271-278, December 2011.

[35] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2(3), pp. 1-27, May 2011.

[36] R. Agrawal and R. Srikant , "Fast algorithms for mining association rules," *In Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-499, San Francisco, USA, 1994.

[37] M. J. Zaki, "Scalable algorithms for association mining," *IEEE Transactions on Knowledge and Data Engineering*, 12(3), pp. 372–390, May/June 2000.

[38] Z. H. Deng, Z. Wang, and J. Jiang, "A New Algorithm for Fast Mining Frequent Itemsets Using N-Lists," *Science China Information Sciences*, 55(9), pp. 2008 - 2030, September 2012.

[39] T. Meng and M.-L. Shyu, "Concept-concept association information integration and multi-model collaboration for multimedia semantic concept detection," *Information Systems Frontiers*, pp. 1-13, April, 2013.

[40] T. Meng and M.-L. Shyu, "Leveraging Concept Association Network for Multimedia Rare Concept Mining and Retrieval," *2012 IEEE International Conference on Multimedia and Expo*, pp. 860-865, July 2012.

[41] T. Meng, "Association Affinity Network Based Multi-Model Collaboration for Multimedia Big Data Management and Retrieval," *Open Access Dissertations*, pp. 1-1115, December 2013.

[42] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," *In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 321-330, 2006.

[43] M. Buckland and F. Gey, "The Relationship Between Recall and Precision," *Journal of the American Society for Information Science*, 45(1), pp. 12-19, January 1999.

[44] T. Meng and M.-L. Shyu, "Leveraging Concept Association Network for Multimedia Rare Concept Mining and Retrieval," *In Proceedings of the 2012 IEEE International Confer-ence on Multimedia and Expo (ICME '12)*, pp. 860-865, Washington DC, USA, July 2012.

[45] T. Meng and M.-L. Shyu, "Automatic annotation of drosophila developmental stages using association classification and information integration," *2011 IEEE International Conference on Information Resue and Integration (IRI 2011)*, pp. 142–147, Las Vegas, Nevada, August 2011.

[46] N. V. ChawlaData, "Data Mining for Imbalanced Datasets: An Overview," Edited by Oded Maimon and Lior Rokach, *Data Mining and Knowledge Discovery Handbook*, Chapter 40, pp. 853-867, 2005.