Semantic Concept Detection Using Weighted Discretization Multiple Correspondence Analysis for Disaster Information Management

Samira Pouyanfar and Shu-Ching Chen
School of Computing and Information Sciences
Florida International University
Miami, FL 33199, USA
{*spouy001, chens*}*@cs.fiu.edu*

***Abstract*—Multimedia semantic concept detection is an emerging research area in recent years. One of the prominent challenges in multimedia concept detection is data imbalance. In this study, a multimedia data mining framework for interesting concept detection in videos is presented. First, the Minimum Description Length (MDL) discretization algorithm is extended to handle the imbalanced data. Thereafter, a novel Weighted Discretization Multiple Correspondence Analysis (WD-MCA) algorithm based on the Multiple Correspondence Analysis (MCA) approach is proposed to maximize the correlation between the feature value pairs and concept classes by incorporating the discretization information captured from the MDL module. The proposed framework achieves promising performance to videos containing disaster events. The experimental results demonstrate the effectiveness of the WD-MCA algorithm, specifically for imbalanced datasets, compared to several existing methods.**

***Keywords*-Weighted discretization; Multiple Correspondence Analysis (MCA); imbalanced data; video concept detection; disaster information management**

I. Introduction

Nowadays, multimedia data consisting of audio, text, image, and video has grown tremendously [1][2][3][4][5]. Social networks such as Facebook, Instagram, and Twitter as well as multimedia sharing websites including YouTube, Flickr, SlideShare, etc. are the main sources of multimedia data widely used by ordinary users and even scientists for research purposes. With such an increase in the amount of multimedia data through the Internet, the main question raised is how one can analyze this high volume and variety of data in an efficient and effective way. To answer this question, many research studies have been done recently in multimedia big data analysis [6][7].

Among various multimedia applications, video concept detection has attracted lots of attention in both academia and industry due to the rich content and information in the videos [8][9]. In the literature, various data mining approaches have been proposed to detect concepts and interesting events in videos [10][1][11][12][13]. Example classifiers include neural networks [14], decision trees [15], Multiple Correspondence Analysis [16], etc. However, one main remaining challenge is that of bridging the gap between the low-level visual features and the high level concepts in the videos.

Another critical challenge in multimedia data is how to process data with skewed distributions or in other words, the imbalanced datasets. This can be seen commonly in real world multimedia applications where the classes are not distributed uniformly [17][18][19][20][21][22]. There are usually two classes: the major classes (or called the negative classes) and the minor one (or called the positive class), where we are more interested in detecting the minor class. For instance, in medical lab results, cancer instances are rare but more important than those instances for regular diseases. Other applications of imbalanced data are fraud activities detection, bomb detection, failure predictions of technical equipment, etc. [23][24]. In such conditions, conventional machine learning and data mining algorithms often fail to detect the minor class, and they are biased toward the negative classes, which may have serious effects. Suppose an instance of a medical lab result is predicted as non-cancer (a negative class), while in reality the patient has the cancer. This error is called false negative, which can cause very serious harm.

To overcome the aforementioned challenges, in this paper, a new Weighted Discretization Multiple Correspondence Analysis (WD-MCA) is proposed. It contains a new weighting factor for the discretization algorithm, which is later utilized in the Multiple Correspondence Analysis (MCA) classifier. By assigning reasonable weights to the instances in the minor class, it would be possible to improve the interesting concept detection in multimedia data. This observation motivates us to propose a new data mining framework to handle the imbalanced data problem. For this purpose, the supervised discretization function introduced in [25] is extended to penalize the negative classes and bias the learning model toward the positive class. Moreover, the discretization factor is integrated to the MCA weighting function for effective video concept detection.

The rest of this paper is organized as follows. Section II discusses the existing work in multimedia data analysis.

In section III, the proposed framework is introduced and each component of the WD-MCA is discussed in details. Section IV gives the experimental results and observations. Finally, conclusions and recommendations for future work are presented.

II. Related Work

Regarding the data imbalance issue, conventional approaches can be mainly categorized into the following groups [23][26]: Sampling methods, cost sensitives learning, and hybrid algorithms. Typically, sampling methods modify the data distribution in order to balance the dataset and improve the classification results. There are two main resampling approaches in the literature: over-sampling the minority (positive) class [27] or under-sampling the majority (negative) class [28]. Either way can be used in any machine learning algorithm as a preprocessing phase. Another solution is Cost Sensitive Learning (CSL) which modifies the learning process by incorporating the misclassification costs of the different classes [29]. Currently, CSL has been applied in various learning algorithms such as decision trees [30], AdaBoost [31], and Naive Bayes [32]. Recently, various hybrid methods have been proposed, which combine the traditional solutions for data imbalance subject [33].

In recent years, with the advent of new technologies and easy access of multimedia data in the social networks and sharing websites, multimedia concept detection has become a hot topic, both in industry and academia [34][35][36][37]. Current video search engines often use textual descriptions and video tags to retrieve videos. However, due to the limitation and subjectivity of video metadata, such engines may provide a very low performance. Thus, automatic concept detection is crucial in multimedia analysis [38]. Ha et al. [39] proposed a new framework using two different correlation-based approaches integrated with a well-known deep learning method called Convolutional Neural Network (CNN) to automatically detect semantic concepts from NUS-WIDE image dataset [40]. The Positive Enhanced Ensemble Learning (PEEL) framework is presented in [41], which addresses the video concept/event detection, specifically for soccer videos. By integrating the ensemble learning algorithm with a sampling-based mechanism, it outperforms the existing single models and ensemble classifiers. The TRECVID data is a very large real world dataset focusing on information retrieval, specifically on video content based retrieval [42]. Recently, many research studies have been done based on the TRECVID dataset, which made considerable contributions in this area and improved the video semantic concept detection, especially for the imbalanced datasets [43].

Despite the fact that many real world applications deal with continuous features, most of the machine learning algorithms can only be applied to nominal or discrete numerical features [44]. Therefore, discretization continuous-valued features are considered as a significant step in the preprocessing phase. Discretization algorithms can be classified into supervised and unsupervised methods [45]. Fayyad and Irani [25] proposed a supervised discretization algorithm using an information entropy heuristic called the Minimum Description Length (MDL) principle. In this algorithm, first, the continuous features are sorted, then the potential cutting points are calculated from classes' boundaries based on the MDL principal. An unsupervised discretization algorithm based on the Self-Organizing Map (SOM) is presented in [46]. Unlike the K-means clustering which requires the number of clusters beforehand, SOM only requires the maximum number of requested intervals and can effectively partition the feature values into nominal values. In this study, however, we extend the MDL approach to improve the discretization algorithm, specifically for the imbalanced datasets.

In this paper, MCA is used as a classifier due to its powerful nature which is able to measure the correlation between the attributes and classes [47]. In the literature, MCA is widely applied to several multimedia applications including feature selection, discretization, data pruning, and classification [48][16][49]. In the current studies, MCA analyzes each instance by using the equal weight function for all feature sets. However, in this paper, MCA is extended to incorporate the discretization information to enhance the classification efficiency.

This study concentrates on binary classification algorithm. The contributions are as follows. First, the MDL discretization algorithm is extended to handle the imbalanced datasets using a novel costing factor. Then, the discretization factor is combined with the MCA weighting function to improve the classification performance.

III. The Proposed Weighted Descritization Multiple Correspondence Analysis Framework

The proposed WD-MCA framework is depicted in Figure 1. The whole framework can be divided in three main steps: the preprocessing component (the top left panel), the training process (the right panel), and the testing phase (the bottom left panel). The preprocessing phase includes shot boundary detection routine, visual feature extraction, and data splitting. As this step is domain specific, other applications may apply different preprocessing routines. For example, each data type (e.g., audio, speech, image, text, and video) may require a specific feature set and various preprocessing techniques. The next step is the training process where a learning model is trained using the proposed WD-MCA algorithm which contains the weighted MDL discretization algorithm and the MCA based discretization factor. On the other side, testing data instances are discretized using the training discretization information and the WD-MCA model is used as a classifier to detect the semantic video concepts.

A. Preprocessing

The preprocessing phase is domain specific and each application applies different preprocessing routines. Specifically, for video analysis, the preprocessing includes shot boundary detection, key-frame selection, and feature extraction as explained in more details as follows.

In this paper, an automatic and effectual shot boundary detection algorithm described in [50] is applied on the raw video. This algorithm is based on an unsupervised method for image segmentation as well as object tracking techniques. The segmentation algorithm first clusters the feature map of every video frame and groups the frame pixels into several classes. Then, these segmentation maps are compared to see how different they are. In addition, an object tracking algorithm is used to detect moving objects and luminance changes, which improves the final matching results. Moreover, it can be further used for other purposes such as content analysis and video indexing.

After final shots are extracted, a key-frame is selected as a representative of each shot. Key-frames are helpful for video summarization, and therefore, it is important to select the most distinctive one which represents the contents of the whole shot. For this reason, the first frame of each shot is chosen because it is the cut-point separating successive shots in the shot boundary detection algorithm.

In this study, several low-level visual features are extracted from raw videos as described in [51]. Histogram of Oriented Gradient (HOG) [52] has been proven to be an effective and robust visual descriptor in many image processing applications such as object recognition, human detection, and action recognition, to name a few. Color and Edge Directivity Descriptor (CEDD) [53] is another well-known visual descriptor, which incorporates a histogram's texture and color information. Many research studies have leveraged the CEDD features for image indexing and retrieval. These two feature sets, plus other low-level visual attributes such as texture wavelet, color histogram, and color moment are integrated as the final feature set.

Finally, the dataset is split into training set and testing set for further procedures.

B. Training Phase

The proposed training algorithm includes two main components. First, a weighted discretization algorithm is applied to the training set and then to the discretized dataset. Second, the discretization factors are used to train the MCA algorithm. The algorithm and its technical details are described as follows.

1) The Weighted MDL Discretization Algorithm: In this component, the Minimum Description Length (MDL) approach [25] is extended to improve the discretization step by considering the importance of positive instances in an imbalanced dataset. For this purpose, a weighting factor is

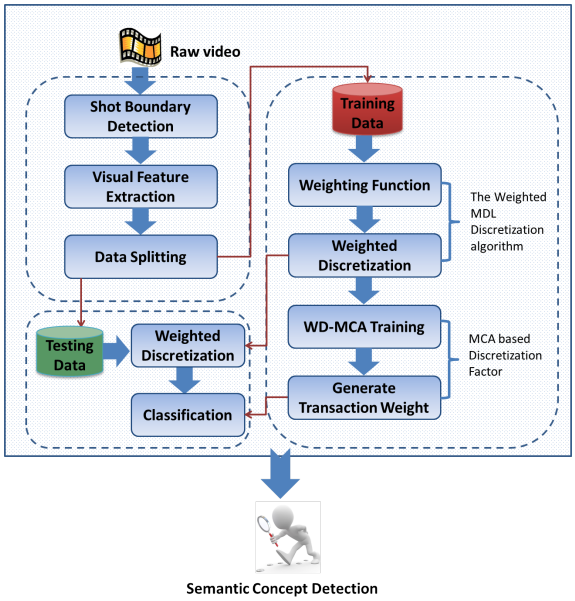

Figure 1: Illustration of the proposed WD-MCA framework

proposed to assign a weight to each positive instance using Equation (1).

$$w_c(i) = \begin{cases} 1 + (ps/L) * \vartheta & \text{if } c = 1 \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

where $w_c(i)$ is the weighting factor for the i^{th} instance, c is the class concept (positive=1 and negative=0), ps is the number of positive instances in the corresponding concept, L is the total number of training instances, and ϑ is a predefined constant. For instance, $w_c(i) = 1.4$ for a positive instance i in a dataset with 200 positive instances out of 10000 training instances, where ϑ=20. The purpose of using the constant factor (ϑ) is to increase the weight of positive instances in an imbalanced dataset. As ps/L is a very small number, especially for a highly imbalanced dataset, it is multiplied by a larger number (ϑ) to increase the weighting factor for positive instances. If ϑ is very small, the weighting factor for positive instances would be very close to the negative ones. On the other hand, if it is very large (>100), the results will be overfitted to the positive class. Therefore, a number between 10 to 50 (depending on the ps/L factor) is reasonable.

In order to find the best cut-point for each feature, the MDL algorithm is applied as follows. First, all the instances are sorted. Next, the class count $Count_c$ in the dataset is calculated as shown in Equation (2).

$$Count_c = \sum_{i=1}^{L} w_c(i). \quad (2)$$

To continue the previous example, $Count_1 = 280$, which increases by 1.4 times for the positive instances. Afterward,

the entropy and information gain [54] of the given dataset is computed using $Count_c$ for both positive and negative classes. Finally, a cut-point of a dataset T, including N instances is evaluated using Equation (3), and $Delta$ is defined in Equation (4).

$$InfoGain > \frac{log_2(N-1)}{N} + \frac{Delta}{N}. \tag{3}$$

$$\begin{aligned} Delta = log(3^{CL} - 2) - ((CL * priorentropy) - \\ (CL_{right} * entropy_{right}) - (CL_{left} * entropy_{left})), \end{aligned} \tag{4}$$

where CL is the total number of classes ($CL = 2$ for binary classification), $priorentropy$ is the entropy value before the split, $entropy_{right}$ and $entropy_{left}$ are the entropy values of the right and left subsets, respectively, and CL_{right}, CL_{left} are the total number of classes of right and left subsets, respectively.

The cut-points will be iteratively generated for both left and right sides of the given dataset until the condition in Equation (3) is true. As a result, all features are discretized into several feature items. Finally, the total number of discretized subsets for each feature is stored in the $DisCount_j$, where $j = 1, 2, \cdots, M$ and M is the total number of features.

2) MCA based Discretization Factor: Multiple correspondence Analysis (MCA) is a modified version of original correspondence analysis which captures the correlation between features and classes. In this paper, the MCA algorithm is enhanced using the discretization information captured from the previous component. In multimedia databases, rows represent data instances and columns represent features as well as the corresponding concept labels. MCA captures the correspondences between rows and columns which will be later leveraged in the classification step to bridge the gap between the low-level visual features and high-level concepts.

Algorithm 1 illustrates the whole procedure of the Weighted Discretization Multiple Correspondence Analysis (WD-MCA) approach. The WD-MCA input includes a matrix containing all training instances T and feature values F; its output is the Weight Matrix ($WM^c_{j,\varphi}$) calculated using the correlation information. First, as described in the previous section, each feature is discretized using the weighted MDL discretization algorithm called $WDISC$ in line 1 in Algorithm 1, which generates the discretized data as depicted in Table I. Let the total number of feature items for all features be $DisCount$, new training instances being discretized into nominal intervals be T', feature items be $F'_{j,\varphi}$, and c_1 and c_2 be the positive and negative classes. Afterwards, an indicator matrix (Ind) is constructed whose dimension is $(DisCount + CL) * (DisCount + CL)$ as shown in Table II. This table is a binary representation of the discretized features, where the rows indicate the training instances and the columns indicate the feature items (feature-value pairs). Therefore, each instance can only belong to one

Table I: Discretized data

	F_1	F_2	$\cdots$	F_M	Class
t'_1	$F'_{1,1}$	$F'_{2,1}$	$\cdots$	$F'_{M,1}$	c_1
t'_2	$F'_{1,2}$	$F'_{2,1}$	$\cdots$	$F'_{M,1}$	c_2
t'_3	$F'_{1,2}$	$F'_{2,3}$	$\cdots$	$F'_{M,2}$	c_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
t'_N	$F'_{1,1}$	$F'_{2,2}$	$\cdots$	$F'_{M,2}$	c_1

Table II: Indicator matrix

	$F'_{1,1}$	$F'_{1,2}$	$F'_{2,1}$	$\cdots$	$F'_{M,1}$	$F'_{M,2}$	c_1	c_2
t'_1	1	0	0	$\cdots$	1	0	1	0
t'_2	0	1	1	$\cdots$	1	0	0	1
t'_3	0	1	0	$\cdots$	0	1	0	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
t'_N	1	0	0	$\cdots$	1	0	1	0

of the feature items. In this case, the corresponding indicator value equals 1.

After that, Burt matrix ($Burt$) is calculated by the inner product of the indicator matrix ($Burt = Ind^T Ind$) as shown in Table III. Each number in the $Burt$ matrix represents the number of occurrences of a specific feature item. For instance, $Burt(F'_{1,2}, c_2) = 2$ if there are two instances with feature item $F'_{1,2}$ that belong to class c_2. Then, $Burt$ is normalized by the grand total (G) of Ind ($Z = Burt/G$). Thereafter, Z is transformed to a new projected space using the eigenvectors V and eigenvalues E extracted from the Singular Value Decomposition (SVD) and the diagonal matrix D is derived from the singular vector V. Next, the correlation (weight $W^c_{j,\varphi}$, as shown in line 10 in Algorithm 1) between classes and feature-value pairs is calculated using the cosine value of the angle between them. The smaller the angle value is, the higher correlated the feature-value pairs and classes are. The final MCA weight value for each feature-value pair is then calculated using the corresponding weight value. For more details regarding the MCA process, please refer to [16].

In this paper, a penalized factor obtained from the discretization algorithm is utilized to reduce the weight of the features with higher feature items. In other words, the smaller the discretization count is, the more valuable information it has. For this purpose, in each iteration, the final weight value is calculated using the $WM^c_{j,\varphi}$ combined with the penalized factor $dw * DisCount_j$, as shown in line 14 in Algorithm 1, where dw is the discretization weight obtained from the weighted MDL algorithm in Section III-B1, and $DisCount_j$ is the number of feature items for each feature j. Eventually, the final weights are stored in the weight matrix $WM^c_{j,\varphi}$.

C. Testing Phase

The testing module includes the weighted discretization and classification steps. First, the testing dataset is discretized into nominal features using the weighted discretization algorithm described in Section III-B1. Then, the final

Table III: Burt matrix

	$F'_{1,1}$	$F'_{1,2}$	$F'_{2,1}$	$\cdots$	c_1	c_2
$F'_{1,1}$	2	0	1	$\cdots$	2	0
$F'_{1,2}$	0	2	1	$\cdots$	0	2
$F'_{2,1}$	1	1	1	$\cdots$	1	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
c_1	2	0	0	$\cdots$	2	0
c_2	0	2	1	$\cdots$	0	2

Algorithm 1 WD-MCA

Input: Training instances $T\{t_i,\ i = 1, 2, \cdots, N\}$, feature set $F = \{f_j,\ j = 1, 2, \cdots, M\}$.
Output: Weight matrix $WM^c_{j,\varphi}$.

$\{T', F'_{j,\varphi}, DisCount\} \leftarrow$ WDISC(T, F);
for all $f_j \in F, (j = 1, \cdots, M)$ **do**
  Create Indicator matrix Ind;
  Create Burt matrix $Burt$;
  $\{Z, V, E\} \leftarrow MCA(Burt)$;
  Create correspondence matrix CM;
  Derive diagonal matrix D from V;
  for all $F'_{j,\varphi}\ (\varphi = 1, \cdots, DisCount_j)$ **do**
    for all $C_{j,c}\ (c = 1, \cdots, CL)$ **do**
      Calculate $W^c_{j,\varphi}$;
    end for
  end for
  for all $F'_{j,\varphi}\ (\varphi = 1, \cdots, DisCount_j)$ **do**
    for all $C_{j,c}\ (c = 1, \cdots, CL)$ **do**
      $WM^c_{j,\varphi} \leftarrow WM^c_{j,\varphi} + W^c_{j,\varphi}$;
    end for
    $WM^c_{j,\varphi} \leftarrow WM^c_{j,\varphi}/(dw * DisCount_j)$;
  end for
end for
return $WM^c_{j,\varphi}$

features are fed to the classification component to predict the concept class of each testing instance. The $WM^c_{j,\varphi}$ matrix created in the training phase (depicted in Algorithm 1) is used in the testing phase to generate the ranking score for each instance. This score is calculated by accumulating all the weights within one instance i and then is normalized by the total number of features (M) as shown in Equation (5).

$$Score_i = \frac{\sum_{j=1}^{M}(1 - mw_j(i))^2}{M} \tag{5}$$

After the score matrix SM is calculated for all training instances, it can be directly used to rank the testing data. For classifying instances, a threshold needs to be generated based on the training performance as described in [55]. In the first step, training scores are sorted by the descending order, and then the candidate thresholds are selected based on the indexes of the scores with target class label. Finally, the best threshold is generated by iteratively evaluating the performance of the candidate thresholds.

Table IV: Disaster dataset statistics

No.	Concept	#Positive Instances	P/N ratio
1	damage	410	0.060
2	fire	309	0.045
3	mud-rock	143	0.021
4	lightening	674	0.098
5	snow	221	0.032

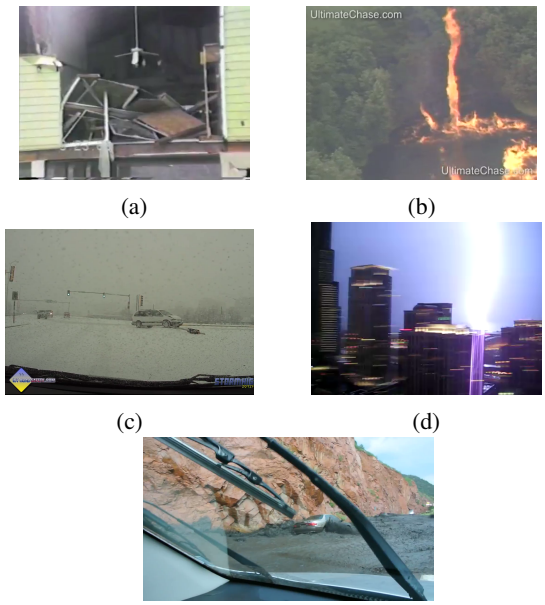

Figure 2: Different sample concepts in the disaster dataset: (a) damage, (b) fire, (c) snow (d) lightening, (e) mud-rock

IV. Experimental Analysis

A. Dataset Description

Although the proposed WD-MCA can be used as a general framework for various multimedia applications (with data including video, image, audio, and/or text), in this paper, a specific task is selected called semantic concept detection from videos containing disaster events. Automatic disaster detection from videos, a new and demanding topic, can be beneficial for classifying videos including disaster events from non-disaster ones. For this purpose, the proposed framework is tested using a new disaster dataset. Specifically, it contains about 80 different YouTupe videos with 5 disaster concepts. Figure 2 depicts a key-frame sample extracted from the videos for each disaster concept. The detailed statistics of the dataset is summarized in Table IV. In total, the dataset includes 6884 video shots and the average ratio of the positive instances to the negative ones (P/N) is 0.051, which shows the non-uniform distribution of the dataset.

B. Evaluation Criteria

In the imbalanced datasets, accuracy may not be the best metric to show the effectiveness of the classification algorithm because most conventional classifiers are biased toward the major (negative) class and may have very high performance on negative classes. However, as the minor (positive) class is more important and critical to be detected, the proposed WD-MCA framework is evaluated using a common measurement metric for imbalanced data called F1 score as defined in Equation (6), where precision and recall are defined as shown in Equations (7) and (8), respectively. Here, TP, FP, and FN refer to the numbers of true positive, false positive, and false negative data instances, respectively.

$$F1 = 2 * \frac{Precision * Recall}{(Precision + Recall)}; \tag{6}$$

$$Precision = \frac{TP}{TP + FP}; \tag{7}$$

$$Recall = \frac{TP}{TP + FN}. \tag{8}$$

C. Evaluation Results

As mentioned earlier, the first step in the proposed framework is preprocessing the data which includes shot boundary detection, key-frame selection, visual feature extraction, and finally data splitting. After applying an automatic shot boundary detection approach [50], the first frame of each shot is selected as a representative of that shot. Then, several visual features as described earlier are extracted from each key-frame. In total, there are 6884 instances and 707 features for each instance. Then, the dataset is divided into three training and testing sets through a 3-fold validation which contains approximately equal numbers of positive and negative instances (P/N ratio is almost equal).

In the training phase (see Section III-B), the training set is discretized using the proposed weighted MDL discretization algorithm and then the testing set is discretized using the same discretization scheme. Afterward, the discretized training instances are passed to the WD-MCA module to train the model.

For evaluating the proposed WD-MCA model, an experiment is conducted using the testing instances to see how accurate they are classified. The performance results are compared to two well-known existing methods: standard MCA [16] and Decision Tree (DT), which achieved very high performance for other imbalanced datasets [15]. The detailed comparison results for each concept and each framework are presented in Table V. As can be seen from this table, the proposed WD-MCA outperforms other methods in terms of F1 score for all disaster concepts. For the fire concept, for instance, it has a promising performance (F1=91%) and improves the classification result by about 8% and 2% compared to DT and MCA, respectively. For the snow concept, the average F1 score is a little bit low

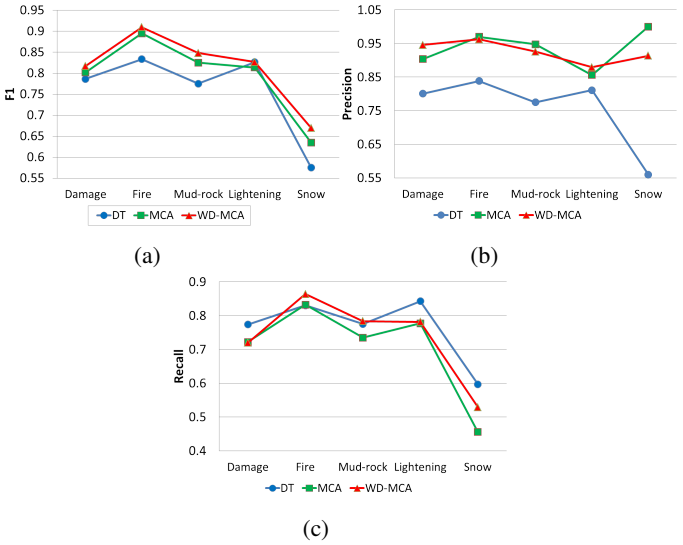

Figure 3: Average comparison results on disaster dataset: (a) F1 score, (b) Precision, (c) Recall

(F1=67%). However, it is still about 10% and 4% higher than that of DT and MCA, respectively. In overall, the average F1 score of all three folds for all 5 disaster concepts is 85%. In an imbalanced dataset where detecting positive instances such as disaster, cancer, fraud, and bomb is very vital, such an improvement (even small for some concepts) is very significant. In addition, only positive instances are used for evaluating the results and the performance of negative instances is not considered. The average comparison results including F1, precision, and recall for each concept are shown in Figure 3. As can be inferred from Figure 3b, MCA has higher precision values in three concepts, while its low recall values, as shown in Figure 3c, decrease its overall performance. The DT algorithm has the lowest precision values in all the concepts, which reduces its overall F1 scores. All in all, the proposed WD-MCA achieves the highest average results, demonstrating the effectiveness of integrating the weighting discretization function with the standard MCA algorithm.

V. Conclusion

Multimedia analysis has attracted lots of attention in recent years. One of the significant applications in multimedia is video semantic concept/event detection. In particular, the data imbalance problem, an open issue in multimedia analysis systems, is selected because conventional data mining algorithms are often unable to detect the minor (positive) class in such non-uniform data distribution. To overcome this challenge, a Weighted Discretization algorithm based on the MCA classifier (WD-MCA) is proposed to improve the correlation between classes and feature-value pairs. Specifically, the MDL discretization approach is extended to tackle the imbalanced data issue by applying a weighting factor to

Table V: Detailed comparison results on disaster dataset

disaster concept	fold #	DT			MCA			WD-MCA		
		precision	recall	F1	precision	recall	F1	precision	recall	F1
damage	fold 1	0.826	0.875	0.85	0.874	0.765	0.816	0.945	0.757	0.841
	fold 2	0.796	0.796	0.796	0.952	0.73	0.826	0.971	0.723	0.828
	fold 3	0.805	0.752	0.777	0.885	0.672	0.763	0.921	0.679	0.782
	average	0.801	0.774	0.787	0.904	0.722	0.802	0.946	0.720	**0.817**
fire	fold 1	0.865	0.800	0.831	0.971	0.825	0.892	0.972	0.875	0.921
	fold 2	0.825	0.846	0.835	1.000	0.878	0.932	0.947	0.923	0.935
	fold 3	0.825	0.846	0.835	0.939	0.795	0.861	0.969	0.795	0.873
	average	0.838	0.831	0.834	0.970	0.833	0.895	0.963	0.864	**0.910**
mud-rock	fold 1	0.723	0.723	0.723	0.929	0.830	0.876	0.929	0.830	0.876
	fold 2	0.857	0.875	0.866	0.970	0.667	0.79	0.921	0.729	0.814
	fold 3	0.745	0.729	0.737	0.944	0.708	0.81	0.927	0.792	0.854
	average	0.775	0.776	0.775	0.948	0.735	0.825	0.926	0.784	**0.848**
lightening	fold 1	0.775	0.844	0.808	0.902	0.741	0.814	0.913	0.746	0.821
	fold 2	0.816	0.827	0.821	0.814	0.800	0.807	0.839	0.809	0.824
	fold 3	0.843	0.858	0.85	0.852	0.791	0.82	0.886	0.791	0.835
	averagee	0.811	0.843	**0.827**	0.856	0.777	0.814	0.879	0.782	**0.827**
snow	fold 1	0.580	0.548	0.563	1.000	0.448	0.648	0.900	0.493	0.637
	fold 2	0.533	0.662	0.590	1.000	0.500	0.667	0.935	0.581	0.717
	fold 3	0.566	0.581	0.573	1.000	0.419	0.590	0.905	0.514	0.655
	average	0.560	0.597	0.575	1.000	0.456	0.635	0.913	0.529	**0.670**

the minor (positive) class. Moreover, the discretization factor is integrated with the MCA algorithm to enhance the multimedia semantic concept detection. The whole WD-MCA framework is successfully evaluated on videos containing the disaster events. This dataset includes few positive instances and has a highly imbalanced P/N ratio. The experimental results show the effectiveness and high performance of the proposed algorithm compared to several existing data mining algorithms in terms of the F1 score.

However, there are still some limitations that need to be overcome. The proposed framework is tested on a new dataset collected by our team, which is not publicly available. In the future, this framework will be extended to detect more concepts from various datasets and applications. Furthermore, in the current framework, only low-level visual features are used for video analysis. Some mid-level and high-level features including spatio-temporal and textual information (e.g., object motion features, and video metadata) will be investigated and utilized to improve the concept detection performance.

ACKNOWLEDGMENT

For Shu-Ching Chen, this research is partially supported by DHS's VACCINE Center under Award Number 2009-ST-061-CI0001 and NSF HRD-0833093, HRD-1547798, CNS-1126619, and CNS-1461926.

REFERENCES

[1] S.-C. Chen, A. Ghafoor, and R. L. Kashyap, *Semantic models for multimedia database searching and browsing*. Springer Science & Business Media, 2000.

[2] S.-C. Chen and R. Kashyap, "Temporal and spatial semantic models for multimedia presentations," in *Proceedings of the 1997 International Symposium on Multimedia Information Processing*, 1997, pp. 441–446.

[3] S.-C. Chen, M.-L. Shyu, and R. Kashyap, "Augmented transition network as a semantic model for video data," *International Journal of Networking and Information Systems, Special Issue on Video Data*, vol. 3, no. 1, pp. 9–25, 2000.

[4] M.-L. Shyu, C. Haruechaiyasak, S.-C. Chen, and N. Zhao, "Collaborative filtering by mining association rules from user access sequences," in *Proceedings of the International Workshop on Challenges inWeb Information Retrieval and Integration (WIRI)*. IEEE, 2005, pp. 128–135.

[5] M.-L. Shyu, S.-C. Chen, and C. Haruechaiyasak, "Mining user access behavior on the www," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, vol. 3. IEEE, 2001, pp. 1717–1722.

[6] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86 –91, july-sept. 2006.

[7] M. L. Shyu, Z. Xie, M. Chen, and S. C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 252–259, Feb 2008.

[8] S.-C. Chen, M.-L. Shyu, and C. Zhang, "An intelligent framework for spatio-temporal vehicle tracking," in *Proceedings of the 4th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2001, pp. 213–218.

[9] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying overlapped objects for video indexing and modeling in multimedia database systems," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 4, pp. 715–734, 2001.

[10] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna, "Semantic event detection via temporal analysis and multimodal data mining," *IEEE Signal Processing Magazine*, vol. 23, pp. 38–46, March 2006.

[11] M.-L. Shyu, S.-C. Chen, M. Chen, and C. Zhang, "A unified framework for image database clustering and content-based retrieval," in *Proceedings of the 2nd ACM International Workshop on Multimedia Databases*, ser. MMDB '04. New York, NY, USA: ACM, 2004, pp. 19–27. [Online]. Available: http://doi.acm.org/10.1145/1032604.1032609

[12] X. Chen, C. Zhang, S.-C. Chen, and M. Chen, "A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval," in *Proceedings of the IEEE International Symposium on Multimedia (ISM)*. IEEE, 2005, pp. 37–44.

[13] X. Chen, C. Zhang, S.-C. Chen, and S. Rubin, "A human-centered multiple instance learning framework for semantic video retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 39, no. 2, pp. 228–233, 2009.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1725–1732. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2014.223

[15] S.-C. Chen, M.-L. Shyu, C. Zhang, and M. Chen, "A multimodal data mining framework for soccer goal detection based on decision tree logic," *International Journal of Computer Applications in Technology*, vol. 27, no. 4, pp. 312–323, 2006.

[16] L. Lin, M.-L. Shyu, G. Ravitz, and S.-C. Chen, "Video semantic concept detection via associative classification," in *The 10th IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2009, pp. 418–421.

[17] S.-C. Chen, S. Rubin, M.-L. Shyu, and C. Zhang, "A dynamic user concept pattern learning framework for content-based image retrieval," *IEEE Transactions on Systems, Man, and Cybernetics: Part C*, vol. 36, pp. 489–495, November 2006.

[18] X. Huang, S.-C. Chen, M.-L. Shyu, and C. Zhang, "User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval," in *Proceedings of the Third International Workshop on Multimedia Data Mining, in conjunction with the 8th ACM International Conference on Knowledge Discovery & Data Mining*, July 2002, pp. 100–108.

[19] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht, "An effective content-based visual image retrieval system," in *Proceedings of the IEEE International Computer Software and Applications Conference*, August 2002, pp. 914–919.

[20] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Video semantic concept discovery using multimodal-based association classification," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, July 2007, pp. 859–862.

[21] M.-L. Shyu, T. Quirino, Z. Xie, S.-C. Chen, and L. Chang, "Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems," *ACM Transactions on Autonomous and Adaptive Systems*, vol. 2, no. 3, Sep. 2007. [Online]. Available: http://doi.acm.org/10.1145/1278460.1278463

[22] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Effective feature space reduction with imbalanced data for semantic concept detection," in *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing (SUTC2008)*. IEEE, 2008, pp. 262–269.

[23] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42–47, 2012.

[24] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," *CoRR*, vol. abs/1305.1707, 2013.

[25] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1993, pp. 1022–1029.

[26] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[27] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *Advances in intelligent computing*. Springer, 2005, pp. 878–887.

[28] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5718–5727, 2009.

[29] N. Thai-Nghe, Z. Gantner, and L. Schmidt-Thieme, "Cost-sensitive learning methods for imbalanced data," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2010, pp. 1–8.

[30] S. Lomax and S. Vadera, "A survey of cost-sensitive decision tree induction algorithms," *ACM Computing Surveys (CSUR)*, vol. 45, no. 2, p. 16, 2013.

[31] J. Zheng, "Cost-sensitive boosting neural networks for software defect prediction," *Expert Systems with Applications*, vol. 37, no. 6, pp. 4537–4543, 2010.

[32] X. Chai, L. Deng, Q. Yang, and C. X. Ling, "Test-cost sensitive naive bayes classification," in *Fourth IEEE International Conference on Data Mining (ICDM)*. IEEE, 2004, pp. 51–58.

[33] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.

[34] S.-C. Chen, S. Sista, M.-L. Shyu, and R. Kashyap, "Augmented transition networks as video browsing models for multimedia databases and multimedia information systems," in *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*, 1999, pp. 175–182.

[35] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht, "Image retrieval by color, texture, and spatial information," in *Proceedings of the 8th International Conference on Distributed Multimedia Systems*, September 2002, pp. 152–159.

[36] M.-L. Shyu, C. Haruechaiyasak, and S.-C. Chen, "Category cluster discovery from distributed www directories," *Information Sciences*, vol. 155, no. 3, pp. 181–197, 2003.

[37] S. Pouyanfar and H. Sameti, "Music emotion recognition using two level classification," in *Iranian Conference on Intelligent Systems (ICIS)*. IEEE, 2014, pp. 1–6.

[38] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "Generalized affinity-based association rule mining for multimedia database queries," *Knowledge and Information Systems (KAIS): An International Journal*, vol. 3, no. 3, pp. 319–337, 2001.

[39] H.-Y. Ha, Y. Yang, S. Pouyanfar, H. Tian, and S.-C. Chen, "Correlation-based deep learning for multimedia semantic concept detection," in *Web Information Systems Engineering (WISE)*. Springer, 2015, pp. 473–487.

[40] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM international conference on image and video retrieval*. ACM, 2009, p. 48.

[41] Y. Yang and S.-C. Chen, "Ensemble learning from imbalanced data set for video event detection," in *2015 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2015, pp. 82–89.

[42] "Trec video retrieval evaluation: Trecvid," retrieved at: 2016-04-04. [Online]. Available: http://trecvid.nist.gov/

[43] Y. Sun, K. Sudo, Y. Taniguchi, H. Li, Y. Guan, and L. Liu, "Trecvid 2013 semantic video concept detection by ntt-md-dut," in *TRECVID 2013*, 2013.

[44] M.-L. Shyu, K. Sarinnapakorn, I. Kuruppu-Appuhamilage, S.-C. Chen, L. Chang, and T. Goldring, "Handling nominal features in anomaly intrusion detection problems," in *Proceedings of the 15th International Workshop on Research Issues on Data Engineering (RIDE Workshop of Stream Data Mining and Applications RIDE-SDMA '2005), in conjunction with The 21st International Conference on Data Engineering (ICDE 2005)*. IEEE, 2005, pp. 55–62.

[45] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Machine learning: proceedings of the twelfth international conference*, vol. 12, 1995, pp. 194–202.

[46] M. Vannucci and V. Colla, "Meaningful discretization of continuous features for association rules mining by means of a som." in *ESANN*, 2004, pp. 489–494.

[47] H. Abdi and D. Valentin, "Multiple correspondence analysis," *Encyclopedia of measurement and statistics*, pp. 651–657, 2007.

[48] L. Lin, M.-L. Shyu, and S.-C. Chen, "Enhancing concept detection by pruning data with mca-based transaction weights," in *11th IEEE International Symposium on Multimedia (ISM)*. IEEE, 2009, pp. 304–311.

[49] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Effective supervised discretization for classification based on correlation maximization," in *2011 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2011, pp. 390–395.

[50] S.-C. Chen, M.-L. Shyu, and C. Zhang, "Innovative shot boundary detection for video indexing," *Video data management and information retrieval*, pp. 217–236, 2005.

[51] Y. Yang and S.-C. Chen, "Disaster image filtering and summarization based on multi-layered affinity propagation," in *2012 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2012, pp. 100–103.

[52] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE, 2005, pp. 886–893.

[53] S. A. Chatzichristofis and Y. S. Boutalis, "Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval," in *Computer vision systems*. Springer, 2008, pp. 312–322.

[54] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[55] Y. Yang, H.-Y. Ha, F. Fleites, S.-C. Chen, and S. Luis, "Hierarchical disaster image classification for situation report enhancement," in *2011 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2011, pp. 181–186.